Li *Language Testing in Asia* (2018) 8:3
DOI 10.1186/s40468-018-0054-7

Language Testing in Asia

RESEARCH Open Access

Developing a test of L2 Chinese pragmatic comprehension ability

Shuai Li

Correspondence: sli12@gsu.edu
Department of World Languages and Cultures, Georgia State University, 25 Park Place, Suite 1933, Atlanta, GA 30303, USA

Abstract

Background: This study aims to develop a test for assessing pragmatic comprehension ability in Chinese as a second language (L2). Following the framework of an argument-based approach to test validation, this study attempts to obtain backing for the Evaluation and Explanation inferences.

Methods: Test items were developed based on two sources of authentic language use (i.e., field notes and a corpus of natural language use). Following a series of piloting, 107 examinees of L2 Chinese completed the test (k = 39) in the main study. Among them, nine examinees had retrospective interviews that probed the knowledge, strategies and processes involved in completing the test.

Results: The assumption underlying the Evaluation inference was supported by satisfactory statistical characteristics of the test (e.g., item/test difficulty, item discrimination, distractor functioning, and item/person fit); moreover, the two assumptions associated with the Explanation inference were backed by quantitative and qualitative evidence demonstrating that variations in test performance were attributable to the targeted construct of pragmatic comprehension ability.

Conclusion: The test appears to be a reliable instrument for assessing pragmatic comprehension ability in L2 Chinese. The test results can be used to inform decision-making on curriculum development for interested Chinese programs.

Keywords: L2 Chinese, Pragmatic comprehension, Implicature, Implied meaning, Argument-based approach to validation

Background

Over the last three decades, pragmatic competence, the ability to conduct socially appropriate communication by linking linguistic forms to communicative functions in social contexts (Leech 2014; Taguchi and Roever 2017; Timpe-Laughlin et al. 2015), has become one of the central components in major models of communicative language ability (e.g., Bachman and Palmer 2010; Purpura 2004). Correspondingly, research on assessing second language (L2) pragmatic competence has witnessed development since the early 1990s. This development is mainly characterized by an expanded repertoire of pragmatic constructs to be assessed (e.g., Roever et al. 2014; Timpe 2013, Youn 2015; for a review, see Taguchi and Roever 2017). Although speech acts are the most thoroughly investigated pragmatic construct in the literature, research on assessing other constructs (e.g., implicature, routine, and interactional competence) remains limited. In fact, construct coverage is a major challenge that

Springer Open

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

researchers in L2 pragmatics assessment face, along with issues regarding test practicality, coping with examinees with a wide range of proficiency, and linking test scores to real-world implications (Taguchi and Roever 2017). Moreover, with very few exceptions (e.g., Yamashita 1996), the bulk of published studies in the field have focused on English. There is thus a need to expand the target language repertoire so as to examine the generalizability of existing findings and to serve a wider examinee population.

In response to the calls for research on assessing under-represented pragmatic constructs and for targeting languages other than English, this study aimed at developing a test of L2 Chinese pragmatic comprehension ability (i.e., the ability to understand non-literal, implied meaning). Only a few studies have focused on assessing pragmatic comprehension ability (Roever 2005, 2006; Taguchi 2009; Walters 2009). These studies have shown a lack of consensus on how pragmatic comprehension ability is defined, operationalized, and empirically linked to test performance. Moreover, existing studies have all employed multiple-choice questions as the format of assessment, but there is a need to further investigate this assessment format by exploring optimal ways of delivering prompts and alternative strategies for developing distractors. Finally, previous research exclusively focused on L2 English. Aside from Taguchi et al.'s (2013) study, which focused on acquisition rather than assessment, little is known about what can be done to assess L2 Chinese pragmatic comprehension ability. To fill these gaps in the literature, this study followed an argument-based approach to test validation (Kane 2013) and examined the soundness of the evaluation and explanation inferences in a validity argument associated with developing a L2 Chinese pragmatic comprehension test.

Pragmatic comprehension ability: definition and operationalization

Pragmatic comprehension ability refers to the ability to understand non-literal, implied meaning during verbal communication. Such implied meaning is called *conversational implicature* (Grice 1975), as shown in the following example:

Man: Has Jim come back from school?

Woman: It is not 3:00 pm yet.

In this example, for the man to correctly interpret the woman's implied meaning (i.e., whether Jim has come home or not), he must assume that the woman's utterance is relevant to their communication; he also needs to process certain contextual information (e.g., Jim's daily routine, the current time of the day, etc.).

As the relevance theory (Sperber and Wilson 1995) posits, implicature comprehension is heavily influenced by implicature strength, which is related to the amount of contextual cues that one must process for meaning interpretation. Comprehension of a weakly conveyed implicature entails processing a larger amount of contextual information before one can reach an interpretation, and this leads to greater processing effort; in contrast, understanding a strongly conveyed implicature requires processing a smaller amount of contextual information, therefore reducing the processing effort necessary for meaning interpretation.

Pragmatics research has shown that conventionality can enhance implicature strength (Taguchi 2008, 2009). Conventionality refers to "common knowledge of the way things are done" and comprises of *conventions of language* and *conventions of usage* (Morgan 1978, p. 279). Concerning conventions of language, certain linguistic forms are

conventionalized to convey pragmatic functions (hereafter Lin). For example, people say *there you go* to express agreement. For conventions of usage, certain discourse patterns can reliably convey communicative functions (hereafter Dis). For example, people often provide a reason (e.g., *I have already had another plan.*) without explicitly saying *no* to reject an invitation (e.g., *Would you like to join us for dinner?*). When implicatures are conveyed through conventionalized linguistic forms (Lin) or predicable discourse patterns (Dis), listeners do not have to extensively process contextual cues for meaning interpretation, because meaning can be derived directly from those conventionalized features. Consequently, the processing effort involved in implicature comprehension is reduced.

On the other hand, implicatures can be expressed without conventionalized features (hereafter Non), such as in the example shown at the beginning of this section. Different from implicatures encoded through conventionalized means (i.e., Lin and Dis), the linguistic forms and discourse patterns people use to convey a non-conventionalized implicature are numerous and unpredictable (Taguchi 2008). Moreover, due to lack of conventionality, comprehension of non-conventionalized implicatures is typically effortful because listeners must draw on various sources of contextual information.

In summary, drawing on the insights from the relevance theory, this study understands pragmatic comprehension ability from the perspective of implicature strength and the associated processing load. Due to the critical role of conventionality in influencing implicature strength and because of the two types of theorized conventionality (i.e., conventions of language and conventions of usage), pragmatic comprehension ability is operationalized as the ability to correctly interpret the meaning of three types of implicatures, i.e., implicatures conveyed through conventionalized linguistic forms (Lin), fixed discourse patterns (Dis), and non-conventionalized utterances (Non).

Research on assessing L2 pragmatic comprehension

In L2 pragmatics assessment, only a few studies (Roever 2005, 2006; Taguchi 2009; Walters 2009) pioneered to develop instruments for testing pragmatic comprehension ability. These studies operationalized pragmatic comprehension ability in different ways; moreover, they reported mixed findings regarding the success of their instruments. These observations entail further empirical research in this area.

Informed by conversation analysis (CA), Walters (2009) operationalized pragmatic comprehension as the ability to understand three types of communicative actions: assessment responses, compliment responses, and pre-sequence responses. He developed a 10-item test in L2 English. Each item included an aural prompt (i.e., a dialog) and four written multiple-choice options. While Walter's study did not adopt an argument-based approach to validation, the results raised issues relevant to the evaluation and explanation inferences of a validity argument for assessing pragmatic comprehension. To illustrate, Walter's test items exhibited various degrees of *sequential indeterminacy*, i.e., the response turn in a conversation can be highly variable among individuals. Therefore, even the native speakers of his study often failed to choose the correct answers, leading to a wide range of item means (from 37 to 93%). The noticeable sequential indeterminacy, in addition to the low reliability of the test (Spearman-Brown split-half coefficient $r = -.137$), indicates a need to investigate the

appropriateness of the prompts (i.e., a dialog), the functioning of multiple-choice options, and the scoring criteria. Walter's study would thus benefit from examining the types of empirical evidence in support of a (hypothetical) evaluation inference. Moreover, Walters did not provide theoretical and empirical evidence to demonstrate the link between expected test performance and the focal construct measured by his test. Hence, it was not clear to what extent test performance reflected pragmatic comprehension ability as operationalized in his study. In other words, the backing for a (hypothetical) explanation inference was missing.

In another study, Roever (2005, 2006) developed a test battery for assessing pragmatic knowledge involved in speech act production, routine judgment, and implicature comprehension. Again, the target language was English. Pragmatic comprehension ability, as assessed by the implicature section of the test battery, was operationalized as the ability to understand idiosyncratic implicatures (equivalent to the Non type as discussed above) and formulaic implicatures (incorporating the Dis and Lin types). The implicature section contained 12 multiple-choice items. Each item included a brief scenario description, a short written dialog, and four options. Among the four options, the correct one was derived from native English speakers' interpretation of the targeted implicature, whereas the three distractors were developed based on non-native speakers' misconceptions. The mean scores of the 12 items ranged from 31.44 to 78.95% with an average of 60.41% (Cronbach's $\alpha = .80$). Test scores increased with proficiency level (operationalized as class levels) but were not affected by L2 exposure (i.e., comparing exposure and non-exposure learner scores). Formulaic implicatures were found to be more difficult than idiosyncratic implicatures.

Roever's instrument can be further improved when viewed from the perspective of an argument-based approach to validation. In terms of the evaluation inference, the written mode of prompt delivery does not resemble how people typically produce and comprehend conversational implicatures. Presenting prompts aurally can better simulate the way implicatures are conveyed in real life. In addition, because only four of Roever's 12 items targeted implicatures conveyed through fixed discourse patterns (Dis) or conventionalized linguistic forms (Lin), more items are needed to represent these two implicature types. Moreover, additional item analyses (e.g., item discrimination, distractor analysis, and examination of fit) other than item means and test reliability would be necessary to better evaluate item/test functioning. Viewed against the explanation inference, additional evidence is desirable to better evaluate the link between pragmatic comprehension ability and test performance. For example, Roever conceptualized the focal construct of his instrument as comprising of two different types of implicatures. To support this theorized structure of pragmatic comprehension ability, he reported a significant difference in test scores based on the two implicature types. Additional backing, such as differences in the knowledge, strategies, and processes involved in comprehending different types of implicatures (Taguchi 2008); a strong correlation between the implicature types; and appropriate Rasch-calibrated item fit statistics, would be needed to further support the argument for the explanation inference (construct validity) involved in his test.

The literature review suggests a need for researching what constitutes appropriate backing to support the evaluation and explanation inferences in a validity argument

associated with developing instruments for assessing L2 pragmatic comprehension ability. To this end, scholars can be informed by related interlanguage pragmatics research, notably Taguchi's studies on L2 English (2005, 2009), L2 Japanese (2008), and L2 Chinese (Taguchi et al. 2013). Of particular relevance is Taguchi et al.'s (2013) study, which is the only one addressing pragmatic comprehension in L2 Chinese. The researchers developed a computerized instrument for assessing the comprehension of implicatures encoded through indirect refusals, conventionalized linguistic forms, and non-conventionalized utterances. Each item included an aurally delivered dialog followed by four written multiple-choice options. Taguchi et al. reported, among other things, a positive proficiency effect on comprehension accuracy, an effect of implicature type on comprehension accuracy (e.g., implicatures encoded through indirect refusals were the easiest to comprehend compared with implicatures conveyed through conventionalized linguistic forms and non-conventionalized utterances). These findings constitute a good reference for evaluating the soundness of the assumptions underlying the explanation inference for future research. However, Taguchi et al. did not report reliability coefficients for their instrument, nor did they provide individual item statistics (e.g., difficulty, discrimination, and fit statistics) and results of the distractor analysis. Therefore, it was not possible to examine their instrument against the assumptions of the evaluation inference. Finally, because the multiple-choice options of Taguchi et al.'s instrument were written in English, the instrument has a restricted application to only those who know English. A test that can be administered to learners with various L1 backgrounds would be desirable.

This study: rationale and purpose

The literature review reveals several issues surrounding research on developing instruments for assessing L2 pragmatic comprehension ability: different construct definition and operationalization with insufficient evidence to support the theorized structure of the targeted construct, non-optimal item delivery format, mixed findings of test reliability, and insufficient analyses for evaluating item/test functioning (e.g., distractor analysis). From the perspective of an argument-based approach to test validation (Kane 2013), those issues appear to be related to the evaluation and explanation inferences. For example, appropriate item delivery format and distractor analyses are within the scope of the evaluation inference, which concerns the conditions that allow test performance to be appropriately evaluated to reflect the targeted construct (i.e., pragmatic comprehension ability). On the other hand, operationalization of the construct of pragmatic comprehension ability, and collecting sufficient evidence to establish the link between test performance and the targeted construct being assessed, falls under the explanation inference.

This study thus focused on obtaining backing in support of the evaluation and explanation inferences in a validity argument associated with developing a L2 Chinese pragmatic comprehension test. Table 1 outlines the inferences, warrants, assumptions, and backing of this study. The evaluation inference in this study is based on the assumption that test performance is appropriately evaluated to reflect examinees' pragmatic comprehension ability (Assumption 1). To this end, an investigation into the various statistical characteristics of the test (e.g., item difficulty, item discrimination,

Table 1 An overview of the evaluation and explanation inferences

Inferences	Warrants	Assumptions	Backing
Evaluation	1. Observations of performance on the test are properly evaluated to provide observed scores reflective of pragmatic comprehension ability.	1. Statistical characteristics of test items are appropriate for the intended purpose of the test.	1. Item/test statistics (e.g., item/test difficulty, item discrimination, item fit, distractor analysis) are within an expected range.
Explanation	2. Expected scores are attributable to a construct of pragmatic comprehension ability.	2. Test performance varies according to well-documented examinee factor(s) affecting pragmatic comprehension.	2. Positive effects of proficiency on test performance.
		3. The internal structure of the test conforms to a theoretical view of the construct of pragmatic comprehension ability.	3a. Strong correlations between different types of implicatures. 3b. Different implicature types lead to different test performance. 3c. Different linguistic knowledge, processes, and strategies are involved in interpreting different types of implicatures.

item/person fit, item/person separation, and distractor functioning) can help evaluate the soundness of this assumption.

The explanation inference in this study rests upon two assumptions showing that test performance is attributable to pragmatic comprehension ability. First, the assumption that test performance varies according to known factors affecting pragmatic comprehension (Assumption 2) can be supported by demonstrating the well-documented effects of proficiency on pragmatic comprehension (e.g., Roever 2005, 2006; Taguchi 2008, 2009; Taguchi et al. 2013). Second, the assumption that the internal structure of the test corresponds to a theoretical view of the construct of pragmatic comprehension ability (Assumption 3) can be backed by three lines of evidence based on theoretical predictions and existing empirical findings: (a) because the three types of implicature are conceptualized to represent different facets of a shared construct, one can expect strong relationships in test performance between them; (b) according to the pragmatics theories reviewed earlier, the varying degrees of conventionality (e.g., with or without conventionalized features) and the different kinds of conventionality (i.e., conventions of language and conventions of use) require different levels of processing effort, leading to variations in test performance; and (c) comprehension of the three types of implicature likely draws on different knowledge, processes, and strategies because of the differences in conventionality (e.g., Taguchi 2008). This study was guided by three research questions (RQs):

RQ1: To what extent are the statistical characteristics of the test appropriate for assessing pragmatic comprehension ability among learners with a relatively wide proficiency range? (Assumption 1)

RQ2: To what extent does test performance vary according to proficiency of L2 Chinese? (Assumption 2)

RQ3: To what extent does the internal structure of the test correspond to a theoretical view of pragmatic comprehension ability? (Assumption 3)

Methods

Item development and piloting

Test items were developed based on a corpus of natural spoken Chinese and field notes taken in China. The corpus, *Chinese Annotated Dialogue and Conversation Corpus*[1], was the only available natural Chinese speech corpus found for the purpose of this study. The corpus consists of 12 extended free-topic conversations (approximately 16.2 h in total) between native Chinese speakers (e.g., classmates, colleagues) recorded in daily settings (e.g., dorms, offices). The researcher read through the transcripts and identified 29 utterances expressing implied meaning. The other resource was the researcher's field notes taken in China in 2013 for this project. The field notes contained utterances encoding Chinese indirect communication that the researcher observed in daily settings (e.g., university campus, restaurants, offices, and home) between friends, relatives, colleagues, and strangers. The researcher read through the field notes and identified 19 utterances expressing implied meaning. For example, one field note entry documented the use of 再说吧 (literal meaning: "Say it again"; implied meaning: "Let's discuss it later") to turn down an invitation between two friends, and this utterance was later included in the instrument.

Based on the identified utterances, 48 dialogs (i.e., test prompts) were developed with the last turn of each dialog encoding implicature. To ensure examinees' familiarity with the topics and scenarios presented in the test prompts, the researcher analyzed the elementary-level textbooks used by the targeted examinees and identified two generic categories of daily interaction scenarios: campus life and personal life. The former category contains topics such as discussing classes, talking about textbooks/books/assignments, making appointments, scheduling activities/events, and borrowing/lending stuff, and the latter category includes topics such as talking about traveling experiences; shopping; discussing food, movies, and scenic spots; making personal plans; and discussing personal items. These scenarios and topics were represented in the 48 dialogs. In this way, the target language domain was operationalized as daily interactions represented in the textbooks used by the targeted examinee population.

The researcher then developed a meta-pragmatic judgment questionnaire including the 48 dialogs, with the last turn of each dialog underlined. Fifteen native Chinese speakers completed the questionnaire for which they (a) wrote out their interpretations of the underlined turn (which contained a targeted implicature) and (b) rated the degree of real-life authenticity for each dialog on a 5-point scale (i.e., 1 point being least plausible in real life and 5 points being truly authentic). Subsequent interviews were held with each native speaker to collect alternative interpretations and comments on the authenticity of the dialogs. Seven dialogs showed varied interpretations, and another two were found with low authenticity scores (i.e., with a mean below 3.5 out of 5). These nine dialogs were removed and 39 dialogs remained.

Four multiple-choice options were created for each dialog: one correct answer, one incorrect answer containing words/phrases from the last turn of the dialog (type A), one incorrect answer based on the meaning of the entire dialog (type B), and one incorrect answer with the opposite interpretation (type C). The three distractor types represented common error types reported in previous research (e.g., Taguchi 2008, 2009; Taguchi et al. 2013).

According to Taguchi (2008), type A distractor is based on the recency effect (Deese and Kaufman, 1957), that is, listeners tend to best recall the last words/phrases that they hear. Type B distractor is based on the keyword processing strategy reported in research on L2 listening comprehension (Ross 1997), that is, listeners tend to rely on initial keyword-referent association for comprehension. Finally, type C distractor is based on the finding that, when listeners are not able to correctly derive the implied meaning, they sometimes reach a completely opposite interpretation.

In order to minimize the confounding influence of examinees' varied grammatical/lexical knowledge on pragmatic comprehension, all prompts and multiple-choice options were written with the 600 words and 16 categories of grammatical structures outlined in the level-3 syllabus of the new HSK test (Hanban 2010).

Two domain experts reviewed the 39 items. The experts were both native Chinese speakers with Ph.D. degrees in applied linguistics. They both have published on L2 Chinese pragmatics and have been involved in L2 Chinese education for more than 10 years. For each test item, the experts (a) decided whether the item assessed pragmatic comprehension ability and if yes, whether the intended option was the only correct answer; (b) reviewed whether the multiple-choice options conformed to the development guidelines described above; and (c) categorized the items into one of the three implicature types (i.e., Lin, Dis, and Non). The researcher reviewed the experts' written responses, conducted follow-up interviews for clarification, and revised the items accordingly. As a result, there were 13 Lin, 13 Dis, and 13 Non items. Two native Chinese speakers (one male, one female) with experience of teaching college Chinese language courses read the 39 dialogs with a speed adjusted to those who have learnt three semesters of college Chinese. Table 2 lists the technical features of the test. The Appendix lists three sample items.

The 39 items were computerized with the software *LiveCode* (http://livecode.com). The order of the items and the multiple-choice options were randomized, and a common random order was used across examinees. For each item, examinees first listened to a dialog. Immediately afterwards, four multiple-choice options appeared on the computer screen. Examinees pressed 1, 2, 3, or 4 on the keyboard to indicate their choices. The computer program then automatically started the next item till the end of the test.

The 39 items were first piloted with 30 native Chinese speakers. The p values (proportion-correct) ranged from 93.33 to 100.00% across the items, with a mean of 99.11%. The eight items that did not reach 100% accuracy rate were reviewed and revised. The researcher then recruited eight examinees of L2 Chinese with various L1 backgrounds and proficiency levels to complete the revised version of the test. They reported no unknown words/grammatical structures, nor did they

Table 2 Item features

	Dialog length (in syllables)	Audio length (in seconds)	Speech rate (syllables per 60 s)	Number of turns	Total option length (in characters)	Real-life authenticity rating
Mean	67.02	25.42	158.19	5.13	40.92	4.42
SD	0.81	0.32	2.11	0.86	0.84	0.35
Min.	66.00	25.03	153.37	4.00	40.00	3.50
Max.	68.00	25.98	163.00	6.00	42.00	4.93

have a problem with the testing procedures. Examinees from first and second year classes found the speed of speech appropriate, while those from 3rd and 4th year classes found the speed somewhat slower than what they were able to handle. Because these comments on the speed of speech were expected, no change was made to the test.

Examinees

A total of 107 learners of L2 Chinese were recruited from a university in China for the main study. Among these examinees, 98 completed the computerized test and nine completed one-on-one retrospective interviews (see the "Procedures" section below). The 98 examinees came from first year ($n = 36$), second year ($n = 24$), third year ($n = 26$), and fourth year ($n = 12$) Chinese classes, and there were 59 females and 39 males (mean age = 23.35 years, $SD = 3.84$); moreover, they represented various L1 backgrounds including Bahasa Indonesia ($n = 12$), Bulgarian ($n = 1$), English ($n = 3$), French ($n = 3$), Hindi ($n = 1$), Italian ($n = 5$), Japanese ($n = 1$), Kazakhstan ($n = 3$), Korean ($n = 25$), Lao ($n = 1$), Mongolian ($n = 4$), Nepali ($n = 1$), Persian ($n = 1$), Russian ($n = 2$), Romanian ($n = 1$), Spanish ($n = 9$), Swedish ($n = 1$), Thai ($n = 10$), Turkish ($n = 1$), Urdu ($n = 1$), Uzbekistan ($n = 11$), and Visayan ($n = 1$).

Procedures

The 98 examinees completed the test individually in a quiet room on campus. After familiarizing themselves with the testing procedure with two practice items, they completed the test, followed by a demographic background survey. The test sessions were monitored by research assistant(s). As for the interviews, each of the nine examinees completed the test item by item in a quiet room on campus with the researcher. Upon completing each item, the researcher asked the same question 你为什么选这个? (*Why did you select this option?*) to prompt the examinees to report the rationale underlying their choices as well as their thinking processes involved in completing the test item. All verbal reports were audio recorded and later transcribed for analysis.

Data analyses

Comprehension accuracy was operationalized as test scores based on correct answers. Each correct response received one point, and each incorrect response received no point. The score range of the test was thus 0 to 39. The researcher used WINSTEPS version 3.80.1 to conduct data analysis with the Rasch dichotomous model. Both raw scores and logits were reported as descriptive statistics. Inferential statistics were performed with SPSS 16.0. Unless stated otherwise, Rasch-calibrated measures were used for statistical analyses. In cases where the assumptions underlying certain parametric procedures were not met, alternative non-parametric procedures were used. Finally, following Taguchi (2008), the interview data were qualitatively analyzed with a focus on the strategies and processes involved in interpreting the implied meaning encoded through the three types of implicature.

Table 3 Item statistics

Item	Mean	*SD*	Corrected item-total correlation	Rasch analysis			
				Measure	Model S.E.	Infit MNSQ	Infit ZSTD
L1	0.66	0.48	0.34	0.55	0.25	1.17	1.4
L2	0.83	0.38	0.44	− 0.63	0.30	0.96	− 0.1
L3	0.62	0.49	0.49	0.79	0.25	0.95	− 0.4
L4	0.76	0.43	0.42	− 0.06	0.27	1.04	0.3
L5	0.63	0.48	0.62	0.73	0.25	0.78	− 2.0
L6	0.51	0.50	0.32	1.43	0.24	1.17	1.6
L7	0.59	0.49	0.49	0.97	0.24	0.94	− 0.6
L8	0.43	0.50	0.27	1.89	0.24	1.21	2.0
L9	0.76	0.43	0.40	− 0.06	0.27	1.07	0.6
L10	0.66	0.48	0.53	0.55	0.25	0.92	− 0.6
L11	0.16	0.37	0.29	3.72	0.32	0.88	− 0.6
L12	0.77	0.43	0.30	− 0.14	0.28	1.19	1.3
L13	0.68	0.47	0.33	0.42	0.25	1.20	1.6
D1	0.79	0.41	0.43	− 0.29	0.28	1.01	0.1
D2	0.88	0.33	0.33	− 1.15	0.34	1.12	0.6
D3	0.78	0.42	0.57	− 0.21	0.28	0.84	− 1.1
D4	0.71	0.45	0.52	0.22	0.26	0.92	− 0.6
D5	0.88	0.33	0.42	− 1.15	0.34	0.91	− 0.3
D6	0.96	0.20	0.34	− 2.56	0.54	0.93	0.0
D7	0.83	0.38	0.52	− 0.63	0.30	0.88	− 0.7
D8	0.56	0.50	0.52	1.14	0.24	0.94	− 0.5
D9	0.91	0.29	0.50	− 1.55	0.38	0.77	− 0.8
D10	0.90	0.30	0.37	− 1.40	0.37	1.04	0.3
D11	0.91	0.29	0.52	− 1.55	0.38	0.79	− 0.8
D12	0.68	0.47	0.48	0.42	0.25	0.99	0.0
D13	0.81	0.40	0.50	− 0.46	0.29	0.92	− 0.5
N1	0.46	0.50	0.49	1.71	0.24	0.91	− 0.9
N2	0.77	0.43	0.28	− 0.14	0.28	1.22	1.5
N3	0.79	0.41	0.49	− 0.29	0.28	0.95	− 0.3
N4	0.70	0.46	0.30	0.29	0.26	1.24	1.8
N5	0.76	0.43	0.58	− 0.06	0.27	0.84	− 1.1
N6	0.67	0.47	0.56	0.48	0.25	0.89	− 0.9
N7	0.86	0.35	0.59	− 0.93	0.32	0.77	− 1.2
N8	0.75	0.44	0.36	0.01	0.27	1.14	1.0
N9	0.77	0.43	0.33	− 0.14	0.28	1.16	1.1
N10	0.83	0.38	0.57	− 0.63	0.30	0.79	− 1.2
N11	0.79	0.41	0.46	− 0.29	0.28	0.99	0.0
N12	0.60	0.49	0.55	0.91	0.24	0.88	− 1.1
N13	0.93	0.26	0.31	− 1.87	0.43	1.02	0.2
Mean	0.73	–	0.44	0.00	.29	0.98	0.0
SD	–	–	–	1.15	.06	0.14	1.0

Results

RQ1 focused on the statistical characteristics of the test. The Cronbach's α of the test was .91 (based on raw scores). Table 3 presents the item measures. The individual item means (p values based on raw scores) ranged from 0.16 (16%) to 0.96 (96%) with an average of 0.73 (73%). Individual item discrimination statistics (item-total correlation coefficients based on raw scores) ranged between .26 and .62 with an average of .46. Rasch analyses showed that item difficulty spanned over 6.28 logits from − 2.56 to 3.72 logits (a positive logit value denotes a higher level of item difficulty than a negative logit value does). The item separation index was 3.58 (5.11 strata) with a reliability coefficient of .93. The item infit mean square (MNSQ) statistics were all within the acceptable range of 0.5 to 1.5 (Boone et al. 2014, p.166; Wright and Linacre 1994) and that the item infit z standardized (ZSTD) statistics were not lower than − 2.0 or higher than 2.0. Examinee ability showed a wide range of 8.00 logits spanning from − 2.31 to 5.69 logits with a mean of 1.49 logits (SD = 1.48 logits) (a positive logit value denotes a higher level of examinee ability than a negative logit value does). The person separation index was 2.32 (3.43 strata) with a reliability coefficient of .84. The person infit MNSQ statistics showed that all but three (3.06%) examinees were within the acceptable range

Table 4 Distractor analyses

	Correct response	Distractor type A	Distractor type B	Distractor type C
Linguistic (Lin, k = 13)				
CSL-1 (n = 36)	48.50%	23.93%	5.56%	22.01%
CSL-2 (n = 24)	65.38%	16.98%	6.09%	11.54%
CSL-3&4 (n = 38)	72.67%	12.75%	5.06%	9.51%
Total (N = 98)	63.34%	18.21%	5.65%	14.84%
Average logit (N = 98)	1.99	0.70	0.62	0.31
Discourse (Dis, k = 13)				
CSL-1 (n = 36)	72.22%	17.09%	4.27%	6.41%
CSL-2 (n = 24)	82.05%	9.29%	5.13%	3.53%
CSL-3&4 (n = 38)	89.68%	6.28%	2.63%	1.42%
Total (N = 98)	81.40%	10.99%	3.84%	3.77%
Average logit (N = 98)	1.82	0.58	− 0.24	0.13
Non-conventional (Non, k = 13)				
CSL-1 (n = 36)	63.03%	17.74%	8.97%	10.26%
CSL-2 (n = 24)	74.68%	15.06%	5.45%	4.81%
CSL-3&4 (n = 38)	84.61%	8.50%	3.44%	3.44%
Total (N = 98)	74.25%	13.50%	5.96%	6.28%
Average logit (N = 98)	1.91	0.38	0.23	0.02
All items (k = 39)				
CSL-1 (n = 36)	61.25%	19.59%	6.27%	12.89%
CSL-2 (n = 24)	74.04%	13.78%	5.56%	6.62%
CSL-3&4 (n = 38)	82.32%	9.18%	3.71%	4.79%
Total (N = 98)	72.55%	14.13%	5.10%	8.21%
Average logit (N = 98)	1.93	0.53	0.14	0.05

Type A an incorrect answer containing words/phrases from the last turn, *Type B* an incorrect answer based on the meaning of the entire dialog, *Type C* an incorrect answer with opposite interpretation

of 0.5 to 1.5. Similarly, the person infit ZSTD statistics revealed that three out of the 98 examinees (3.06%) had a value greater than 2.0 (but less than 3.0).

Table 4 displays the results of the distractor analysis. Third year and fourth year examinees were grouped together (i.e., Chinese as a second language (CSL)-3&4) due to the small sample size of the latter group. All three distractor types led to incorrect responses regardless of item type and proficiency level. Type A distractors (i.e., containing words/phrases from the last turn) showed the strongest effect of distraction. On the other hand, the distracting effects were more or less comparable between type B (i.e., based on the meaning of the entire dialog) and type C (i.e., an incorrect answer with opposite interpretation) for Dis and Non items; for Lin items, type C distractors resulted in considerably higher percentages of incorrect responses than type B distracters did. In fact, for Lin items, the distracting effect of type C approached that of type A. Table 4 also shows the average logit values of the examinees who made incorrect selections. Type A distractors demonstrated the highest mean logit values across item types. On the other hand, the mean logit values for type B distractors were higher than those for type C distractors in terms of Lin and Non items; however, the pattern was reversed for Dis items.

RQ2 asked whether proficiency affected test performance. Table 5 lists the descriptive statistics based on the entire test and based on the three item types. Table 6 shows the results of statistical comparisons. Proficiency was operationalized as three class levels (i.e., CSL-1, CSL-2, and CSL-3&4). One-way between-subject ANOVA revealed that proficiency significantly affected overall test performance, $F(2, 95) = 14.62$, $p < .001$, $\eta_p^2 = .24$. Follow-up pairwise comparisons with Bonferroni corrections showed significant differences between the CSL-1 and CSL-2 groups ($p = .049$), between the CSL-1 and CSL-3&4 groups ($p < .001$), but not between the CSL-2 and CSL-3&4 groups ($p = .062$). Turning to between-group differences within each of the three implicature types, one-way ANOVA revealed a significant effect of proficiency on performance on the Lin items, $F(2, 95) = 13.46$, $p < .001$, $\eta_p^2 = .22$. Subsequent pairwise comparisons with Bonferroni corrections showed significant differences between the CSL-1 and CSL-2 groups ($p = .010$), between the CSL-1 and CSL-3&4 groups ($p < .001$), but not between the CSL-2 and CSL-3&4 groups ($p = .388$) (Table 6). Regarding the other two item types (i.e., Dis and Non), because the data did not meet the normality assumption, *Kruskal–*

Table 5 Descriptive statistics of comprehension accuracy

	Total ($k = 39$)		Lin ($k = 13$)		Dis ($k = 13$)		Non ($k = 13$)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Based on raw scores								
CSL-1 ($n = 36$)	0.61	0.21	0.49	0.21	0.72	0.24	0.63	0.26
CSL-2 ($n = 24$)	0.74	0.17	0.65	0.21	0.82	0.19	0.74	0.21
CSL-3&4 ($n = 38$)	0.82	0.14	0.72	0.19	0.89	0.15	0.84	0.16
Total ($N = 98$)	0.73	0.20	0.62	0.22	0.81	0.20	0.74	0.22
Based on logits								
CSL-1 ($n = 36$)	0.64	1.26	0.16	1.16	1.54	1.71	0.84	1.77
CSL-2 ($n = 24$)	1.49	1.18	0.94	1.40	2.23	1.45	1.57	1.36
CSL-3&4 ($n = 38$)	2.28	1.41	1.49	1.56	3.05	1.41	2.38	1.34
Total ($N = 98$)	1.48	1.47	0.74	1.55	2.29	1.66	1.61	1.64

Table 6 Effects of proficiency on comprehension accuracy

	Total ($k = 39$)	Lin ($k = 13$)	Dis ($k = 13$)	Non ($k = 13$)
CSL-1 vs. CSL-2	$p = .049$*	$p = .010$*	$p = .110$	$p = .087$
CSL-1 vs. CSL-3&4	$p < .001$*	$p < .001$*	$p < .001$*	$p < .001$*
CSL-2 vs. CSL-3&4	$p = .062$	$p = .388$	$p = .020$	$p = .034$

*denotes statistically significant contrasts at 0.5 and .01 levels

Wallis tests were performed, with *Mann–Whitney U* tests as follow-up procedures whenever applicable (the alpha level was adjusted to .017 with Bonferroni corrections). The results showed that proficiency significantly affected performance on the Dis items, χ^2 (2, $N = 98$) = 15.74, $p < .001$. Follow-up comparisons showed no significant difference between CSL-1 and CSL-2 groups ($Z = -1.61$, $p = .110$), a significant difference between CSL-1 and CSL-3&4 groups ($Z = -3.79$, $p < .001$), and no significant difference between CSL-2 and CSL-3&4 groups ($Z = -2.32$, $p = .020$). Finally, in terms of the Non items, proficiency again significantly affected test performance, χ^2 (2, $N = 98$) = 15.83, $p < .001$. Subsequent comparisons showed no significant difference between CSL-1 and CSL-2 groups ($Z = -1.72$, $p = .087$), a significant difference between CSL-1 and CSL-3&4 groups ($Z = -3.85$, $p < .001$), and no significant difference between CSL-2 and CSL-3&4 groups ($Z = -2.11$, $p = .034$). In summary, the CSL-3&4 group consistently outperformed the CSL-1 group in test performance across item types and based on the entire test. The CSL-2 group outperformed the CSL-1 group on the Lin items and on the entire test. There was no other significant contrast.

RQ3 investigated the internal structure of the test. In examining the strength of correlation between performances on the three types of items, Spearman's ρ statistics were calculated due to the violation of the normality assumption. There were strong correlations between the Lin and Dis items, $\rho = .70$ ($p < .001$); between the Lin and Non items, $\rho = .72$ ($p < .001$); and between the Dis and Non items, $\rho = .61$ ($p < .001$).

Regarding the effects of implicature type on test performance, because the data did not meet the normality assumption, *Friedman* tests were performed, with *Wilcoxon* tests as follow-up procedures whenever applicable (the alpha level was adjusted to .017 with Bonferroni corrections). Table 7 summarizes the results of the statistical comparisons. Implicature type led to significant difference in test performance for all examinees as a group, χ^2 (2, $n = 98$) = 78.25, $p < .001$. Follow-up comparisons revealed significant differences for all three contrasts (Table 7). Moreover, the significant effects of implicature type on comprehension were found within each of the three examinee groups, including the CSL-1 group, χ^2 (2, $n = 36$) = 36.78, $p < .001$; the CSL-2 group, χ^2 (2, $n = 24$) = 13.00, $p = .001$; and the CSL-3&4 group, χ^2 (2, $n = 38$) = 31.63, $p < .001$. Follow-up analyses showed that test performance was significantly different between the three

Table 7 Effects of implicature type on comprehension accuracy

	Total ($N = 98$)	CSL-1 ($n = 36$)	CSL-2 ($n = 24$)	CSL-3&4 ($n = 38$)
Lin vs. Dis	$p < .001$*	$p < .001$*	$p < .001$*	$p < .001$*
Lin vs. Non	$p < .001$*	$p < .001$*	$p = .017$*	$p < .001$*
Dis vs. Non	$p < .001$*	$p = .002$*	$p = .026$	$p = .002$*

*denotes statistically significant contrasts at 0.5 and .01 levels

implicature types within each examinee group (see Table 7), with the only exception being the contrast between Dis and Non items for the CSL-2 group.

In additional to the above analyses, examinees' retrospective protocols were analyzed to examine whether different knowledge, inferential processes, and strategies were involved in comprehending the three types of implicature. As it turned out, comprehending implicature encoded through fixed discourse patterns appeared to be the most straightforward process. Because the majority of the Dis items involved indirect refusals, examinees justified their meaning interpretations by directly citing the reason(s) that the interlocutor provided in the dialog, as shown in the following excerpt:

Excerpt 1: Dis item (item #2 in the Appendix)

Examinee #2 (CSL-1, correct interpretation): "女的说7点去的时候,路上车很多,开不快。所以女的意思是早去比较好。".

"The woman says when leaving at 7 o'clock, there will be many cars in the street, (so they) won't be able to drive fast. Therefore the woman means that it is better to leave earlier."

In contrast, in comprehending implicatures conveyed through non-conventionalized utterances, examinees typically explained their inference chain based on the content of the most relevant adjacency pair, and, in doing so, they often drew on background knowledge and personal experience to justify their interpretations. For example, in the following excerpt, the examinee first repeated the last adjacency pair containing the implicature; she then explained her (incorrect) interpretation by referring to a courteous behavior that routinely occurs, that is, closing the door before playing loud music so as to not disturb other people.

Excerpt 2: Non item (item #3 in the Appendix)

Examinee #5 (CSL-2, incorrect interpretation): "这个男的说你想不想听我唱歌,女的说她关门,意思是说她想听,不想打扰别的人。".

"The man says do you want to hear me singing the song, and the woman says she will close the door, meaning that she wants to listen, (but) does not want to disturb other people."

Finally, the processes involved in comprehending implicatures encoded through conventionalized linguistic forms differed according to whether examinees had acquired the relevant pragmalinguistic knowledge. When examinees already knew the pragmatic function of a conventionalized linguistic form, they often directly referred to the linguistic form and paraphrased the implied meaning. For example, in the following excerpt, the examinee correctly explained the conventionalized meaning of the phrase *kě bú shì ma* ("tell me about it" or "you are right").

Excerpt 3: Lin item (item #1 in the Appendix)

Examinee #7 (CSL-3, correct interpretation): "…女的说'可不是嘛'。 '可不是嘛'的意思是'对啊'。".

"…the woman says 'kě bú shì ma'. 'kě bú shì ma' means 'yes'."

On the other hand, when examinees did not have the necessary pragmalinguistic knowledge, they tended to be attracted to the literal meaning of the conventionalized phrase and reach incorrect interpretations, as shown in excerpt 4 (based on the same item as excerpt three). Interestingly, the examinee actually repeated the conventionalized linguistic in its entirety (*kě bú shì ma*); however, in explaining his interpretation, the examinee left out the first character of the phrase and therefore

completely altered the intended meaning. This is because without the first character *kě*, the remaining *bú shì ma* ("not right") no longer encodes the conventionalized pragmatic meaning of agreement but instead expresses a disagreement. Clearly, although the examinee was able to repeat the targeted linguistic form, he did not have the required pragmalinguistic knowledge and failed to reach correct pragmatic comprehension.

Excerpt 4: Lin item (item #1 in the Appendix).

Examinee #1 (CSL-1, incorrect interpretation): "男人说长林是冷还没有意思。女人说'可不是嘛'。这样说的,不是嘛'的意思是有意思。对不对?"

"The man says Changlin is cold and not interesting. The woman says *kě bú shì ma*. In saying so, *bú shì ma* (in this context) means (Changlin is) interesting. Am I right?"

Still another finding emerged from the protocol analyses was that examinees often appeared to be tentative in their meaning interpretation when they were drawn to the literal meaning of the conventionalized linguistic forms. Their uncertainty was demonstrated in their attempts to seek verification from the interviewer, as shown in the end of excerpt 4. This strategy of seeking external verification was not observed in the verbal reports for Dis and Non items.

Discussion

This section first discusses the research findings in light of the three assumptions outlined in Table 1 and then provides a validity argument for the two targeted inferences.

Assumption 1: Appropriateness of the statistical characteristics of the test for its intended purpose

The first backing for Assumption 1 comes from the finding that individual item discrimination statistics were all above the commonly accepted threshold of .25 (Fulcher and Davidson 2007, p.104). Furthermore, the test was found to be able to reliably differentiate more than three levels of pragmatic comprehension ability (3.34 strata, with a reliability coefficient of .84). Because the test was intended for learners enrolled in a 4-year Chinese program, one would expect three to four levels of pragmatic comprehension ability among them. Although three examinees' performance did not fit the Rasch model, the percentage was below the 5% misfit ratio by chance.

Turning to the difficulty measure of the test, the 39 items showed a wide range of difficulty (Table 3), and they represented nearly six levels of difficulty (i.e., 5.73 strata, with a reliability coefficient of .94). The descriptive statistics in Table 5 indicate that the test was somewhat easy, particularly for the CSL-3&4 group. There are two possible explanations. First, the test items were developed by using a restricted set of vocabulary and grammar structures (see the "Methods" section), which limited the sentence complexity and vocabulary width of the test prompts (i.e., dialogs). Consequently, the CSL-3&4 group was probably in a more advantageous position compared with the CSL-1 and CSL-2 groups because the former group was likely to be more familiar with the words and sentence structures. However, the use of a restricted pool of vocabulary and grammar structures was justified in this study because the test was not intended to assess

vocabulary/grammar knowledge but rather the ability to infer the implied meaning of utterances, for which understanding the literal meanings of utterances is a prerequisite. A second and more plausible explanation is that proficiency has well-documented strong positive effects on pragmatic comprehension (e.g., Roever 2005, 2006, 2013; Taguchi 2005, 2008, 2009; Taguchi et al. 2013). A study worth discussing here is Roever (2013), which reported a ceiling effect among a group of advanced ESL learners who took the implicature test developed by Roever (2005, 2006). Because Roever did not use a restricted set of vocabulary and grammar structures in developing his items, the ceiling effect he reported suggests that examinees with advanced proficiency can be expected to do well on a pragmatic comprehension test, and the performance of the CSL-3&4 group in this study confirmed this expectation. Nevertheless, it would be desirable to explore ways to create more difficult items in the future. Given the higher level of difficulty of the Lin items compared with Dis and Non items (Table 5), one possibility is to increase the percentage of Lin items in a future pragmatic comprehension test.

The results of distractor analyses also lent support to Assumption 1 in that all three types of distractors functioned as expected. Different from existing pragmatics assessment studies that either did not specify the rationale for distractor development (Walters, 2009) or used L2 learners' incorrect interpretations as distractors but did not explain why certain incorrect interpretations were selected to be included in tests (e.g., Roever 2005; Liu 2006), this study followed Taguchi et al. (2013) to develop distractors based on relevant psychological theories and SLA research findings. The results showed that all three types of distractors led to incorrect responses. Type A distractors (based on the recency effect) demonstrated the strongest distracting effect across item types and proficiency levels, resulting in the highest percentages of incorrect responses as well as the highest average examinee ability who selected this distractor. This result corroborated previous research findings (Taguchi 2008) and suggests that, when the implied meaning is not readily accessible, examinees tend to rely on memory of the most recent information for meaning inference.

The distracting effects of type B (based on the keyword processing strategy) and type C (representing an opposite interpretation) distractors differed according to item types. In terms of percentage of incorrect responses (see Table 4), type B and type C distractors were more or less comparable for Dis and Non items; yet for Lin items, the examinees were more likely attracted to type C distractors than to type B distractors. Meanwhile, in terms of the average ability of the examinees who chose these two distractors (see Table 4), the ability level associated with type B distractors was higher than that associated with type C distractors for Lin and Non items; in contrast, when it comes to Dis items, the ability level associated with type B distractors was lower than that associated with type C distractors. These findings led to two observations. First, for items requiring the knowledge of conventionalized form-function mappings (i.e., Lin items), a lack of such knowledge tends to lead to completely opposite meaning interpretations, and this is particularly the case for lower ability examinees. Excerpt 4 of the protocol analysis is a good example here: the examinee (from the CSL-1 group) was attracted to the literal meaning of the conventionalized utterances, which led to the opposite meaning interpretation. Second, for items assessing knowledge of fixed discourse

patterns (i.e., Dis items), lower ability examinees are most likely to rely on the content of an entire dialog for meaning interpretation. This makes sense because, as illustrated in excerpt 1, comprehension of the implied meanings embedded in Dis items requires examinees to successfully identify the key turn in an adjacency pair and understand its meaning. Lower ability examinees may not be able to do so and instead may search an entire dialog for cues that can help with meaning interpretation.

Assumption 2: Effects of proficiency on test performance

Appropriate backing for Assumption 2 needs to demonstrate positive effects of proficiency on test performance, because proficiency is the only examinee factor that has consistently been reported to affect pragmatic comprehension (e.g., Roever 2005, 2006; Taguchi 2005, 2008, 2011; Taguchi et al. 2013).This expectation was confirmed in this study, particularly by the comparison between the CSL-1 and CSL-3&4 groups: the latter group consistently outperformed the former group on the entire test and across item types. In comparing adjacent proficiency groups, however, the effects of proficiency were less prominent across item types. For example, between the CSL-1 and CSL-2 groups, significant difference in test performance was observed only for Lin items. Comparing the CSL-2 and CSL-3&4 groups, there was no significant contrast at all. In this study, the examinees were enrolled in an intensive language program with approximately 20 h of formal instruction each week, and they were presumably able to access abundant learning opportunities afforded by the study abroad environment to improve pragmatic comprehension ability. Yet, without focused pragmatics instruction, the pace of acquiring the ability to comprehend implicatures appeared to be quite slow: after all, the examinee went through the first 2 to 3 years of the curriculum before realizing statistically significant, all-round development in this ability (i.e., comparing the CSL-1 and CSL-3&4 groups). Focused instruction on implicature may be needed to facilitate the development of pragmatic comprehension ability.

Assumption 3: The internal structure of the test corresponds to a theoretical view of pragmatic comprehension ability

The first backing to support Assumption 3 entails evidence demonstrating that the test is measuring a unidimensional latent construct. This was confirmed by the results of the Rasch analyses showing that the item infit MNSQ statistics and the item infit ZSTD values were all within acceptable ranges. In addition, the strong correlations between performances on the three types of implicature further supported the claim that the test items measured a shared construct.

Due to the theorized differences in the degree of conventionality encoded in the three types of implicature, additional backing to support Assumption 3 requires evidence demonstrating differential test performance according to the three item types. Indeed, implicatures conveyed through fixed discourse structures (Dis) were the easiest to comprehend. The retrospective protocols (e.g., excerpt 1) complemented this quantitative finding in showing that examinees were able to utilize the conventionalized discourse structure to reduce the processing effort involved in comprehension. This result was consistent with previous research findings (Taguchi 2008; Taguchi et al. 2013) and

demonstrated the facilitative effects of conventions of usage (Morgan 1978) on interpreting implied meaning.

Interestingly, implicatures encoded through conventionalized linguistic forms (Lin) were found to be more difficult to comprehend than non-conventionalized implicature (Non) regardless of examinee proficiency level. This result contrasted with Taguchi et al.'s (2013) findings, which showed that Non items were more difficult than Lin items. The discrepancy is likely due to the specific linguistic conventions included in the two studies. Among Taguchi et al.'s 12 Lin items, five items included rhetorical questions and five ellipsis sentences (i.e., utterances that are intentionally left incomplete in order to convey implied meaning), which they argued to be relatively easy to interpret due to cross-cultural similarity in pragmatic function. This study, however, only included three items with rhetorical questions and one item with ellipsis sentences, and these items (i.e., L2, L9, L12, and L13) received the highest mean scores (p values) among the Lin items (Table 3). The remaining nine items were similar to the sample Lin item (Appendix) in the sense that examinees without the required pragmalinguistic knowledge would likely be drawn to the literal meaning of the utterances and reach incorrect interpretations (see excerpt 4). The lower percentage of conventionalized linguistic features whose pragmatic functions are shared cross-culturally might have made the Lin items in this study difficult. Moreover, analyses of the retrospective verbal reports revealed that the examinees were more tentative in responding to the Lin items (e.g., explicitly requesting external verification) than to the Dis and Non items. This showed that processing implicature encoded through conventionalized linguistic forms was more effortful than processing the other two types of implicature, leading to less accurate meaning interpretations for Lin items than for Dis and Non items. These results suggest that the facilitative effect of conventions of language on implicature interpretation cannot be taken for granted among L2 examinees. Such conventions can reduce processing effort in implicature interpretation only after examinees have acquired the necessary pragmalinguistic knowledge.

Summary: a validity argument for the evaluation and explanation inferences

The evaluation inference focused on the conditions for ensuring appropriate evaluation of test performance so that test scores reflect examinees' pragmatic comprehension ability. Developed based on authentic materials (corpus and field notes), the test items were piloted with native speakers and comparable examinees and were standardized to reduce the influences of potentially confounding variables (e.g., unfamiliar grammar and vocabulary, length of prompts and options, real-life authenticity). Although these measures do not directly support the evaluation inference as specified in Table 1, they constitute the prerequisite conditions for ensuring the soundness of this inference.

Turning to the assumption (Assumption 1) underlying the evaluation inference, it received backing from the satisfactory item discrimination statistics and distractor functioning. Admittedly, the test turned out to be relatively easy for examinees with advanced proficiency in Chinese (i.e., the CSL-3&4 group). Yet, given the well-documented positive effects of proficiency on L2 pragmatic comprehension (Taguchi 2008, 2011; Taguchi et al. 2013; Roever 2005, 2006), particularly the ceiling effect

among advanced ESL examinees as reported in Roever (2013), the CSL-3&4 group's test performance likely reflected their real pragmatic comprehension ability. In supporting the evaluation inference, a unique contribution of this study to the pragmatics assessment literature was to show the feasibility of an alternative approach to developing distractors that functioned satisfactorily. Interestingly, the different types of distractors showed differential distracting effects across implicature types. Such interaction between distractor type and implicature type has not been reported in the literature and warrants future research.

As mentioned earlier, the explanation inference concerns the extent to which test performance can be explained by the targeted construct of pragmatic comprehension ability. The two assumptions (Assumptions 2 and 3) underlying this inference were supported by empirical evidence showing that test performance conformed to a theoretical view of the construct of pragmatic comprehension ability. Rasch analyses, correlation analyses, and retrospective protocol analyses together demonstrated that the test measured three different but highly related facets of a shared latent construct (Assumption 3). The pragmatics theories employed for test development, in addition to the experts' review of the content of the test items, lent further backing to the claim that the test measured pragmatic comprehension ability conceptualized as comprising of three components (Assumption 3). Finally, the confirmation of positive effects of proficiency on pragmatic comprehension was consistent with existing research findings on L2 pragmatic comprehension, thereby supporting Assumption 2. Because previous research on assessing L2 pragmatic comprehension ability has not fully specified the types of empirical evidence that can help verify the internal structure of pragmatic comprehension ability, the results of this study (particularly those for RQ3) can serve as a reference for future research.

Conclusions

One practical use of the results of this study would be to inform decisions on curriculum development for the Chinese program where the examinees came from, because implicature was not a unit of focused instruction in the program. First, regarding whether focused instruction on implicature comprehension is needed, the results showed that the advanced learners (i.e., CSL-3&4) may not necessarily need focused instruction due to their good performance. However, the relatively large standard deviation of the CSL3&4 group showed a lack of homogeneity in pragmatic comprehension ability among the learners in this group. Considering the examinees' test performance across proficiency levels, it seems appropriate to include implicature instruction into the lower division of the Chinese curriculum. Second, regarding the sequence of instruction, implicatures conveyed through fixed discourse patterns (Dis) can be introduced first due to relative easiness. Moreover, because implicatures encoded in pragmalinguistic forms (i.e., those assessed by the Lin items) and implicatures expressed without linguistic conventions (i.e., those assessed by the Non items) were found to be generally difficult, these two implicature types can be introduced later in the curriculum and with more instructional attention.

In terms of how to incorporate implicature comprehension into L2 classroom instruction, interested readers can refer to Taguchi (2007) for detailed suggestions that can be applied to the context of L2 Chinese instruction. Briefly, instructors can

highlight typical discourse patterns (e.g., providing excuses to turn down a request/invitation) and note their role in assisting understanding implied meaning. As for implicatures conveyed through conventionalized linguistic forms, instructors should explicitly teach such pragmalinguistic forms and discuss their pragmatic functions. To this end, it would be efficient to focus on certain categories of pragmalinguistic structures (rather than individual forms), such as ellipsis sentences and rhetoric questions as included in this test. Finally, regarding non-conventionalized implicatures, instructors can encourage learners to utilize multiple contextual cues for meaning interpretation, such as certain paralinguistic cues (e.g., intonation, tone of speech), personal experience, and background knowledge.

This study has several limitations. First, although the results showed that the distractors functioned properly, in hindsight, there is one major limitation in distractor development: type C distractor and the correct answer (the key) contrast directly with each other. As a reviewer pointed out, because type C distractor and the key account for all the possibilities, examinees can quickly develop test-taking strategy, particularly in high-stake test environments, and choose between these two options as the key. This would lead to improper item functioning. Further development of this test should involve revising the rule for writing type C distractor. Second, proficiency was operationalized in terms of learners' class levels. Although this is a widely adopted approach in L2 pragmatic comprehension research (e.g., Roever 2005, 2006; Taguchi 2008; Taguchi et al. 2013), it may not allow a precise understanding of the relationship between proficiency and pragmatic comprehension ability. Future studies should include a standardized proficiency test to better understand how proficiency affects pragmatic comprehension. Finally, given the limited scope of this study, it was not possible to address fundamental issues involved in the test validation process such as fairness of the test. This topic is to be researched during the next phase of this project.

Endnotes

[1]For more information, please visit: http://paslab.phonetics.org.cn/index.php/achievements/resources/cadcc-_han_yu_pu_tong_hua_zi_ran_kou_yu_dui_hua_yu_liao_ku

Appendix

Sample items (English translation not available in the actual test)

Item one: conventionalized linguistic form (Lin)

男:下个星期你要去长林,是吗?

Next week you are going to Changlin, right?

女:对,我要去长林看我一个朋友。

Yes, I am going to Changlin to visit a friend of mine.

男:现在是冬天,长林非常冷啊。

It is winter now, Changlin is very cold.

女:是啊。

Definitely.

男:而且长林那个地方没什么好玩儿的,城市也很小,真的没什么意思。

And Changlin has no fun places to visit, the city is very small, (it is) really not interesting.

女:可不是嘛!

Tell me about it!

1. 女的常常去长林看她朋友
The woman often goes to Changlin to visit her friend.

2. 女的觉得长林没有意思
The woman feels that Changlin is boring. (Correct)

3. 女的喜欢长林这个地方
The woman likes Changlin.

4. 女的可不是没去过长林
The woman has not been to Changlin before.

Item two: fixed discourse pattern (Dis)
男:明天晚上的比赛你去看吗?
Will you go to watch the match tomorrow evening?
女:去啊,你不去吗?
Sure, aren't you going?
男:我也去。对了,比赛几点开始?
I will go too. Oh, when does it start?
女:七点半开始。
At seven thirty.
男:那我吃完饭再去。我们七点开车一起去,怎么样?
Then I will go after dinner. Let's drive there together at Seven, OK?
女:那个时候路上的车会很多,开不快。
There will be lots of cars on the road at that time. (We) won't be able to drive fast enough

1. 女的开车开得不快
The woman cannot drive fast.

2. 女的想七点开车去看比赛
The woman wants to drive to the match at 7 o'clock.

3. 女的觉得七点走不好
The woman doesn't consider it a good idea to leave (for the match) at 7 o'clock. (Correct)

4. 女的常常和男的一起看比赛
The woman and the man often watch matches together.

Item three: non-conventionalized utterance (Non)
女:李明,你在听什么歌呢?
Li Ming, what music are you listening to?
男:我在听"说唱音乐"。
I am listening to rap.
女:说唱音乐?我没听过。什么是说唱音乐?
Rap music? I haven't heard about it. What is rap music?
男:就是一边唱一边说。现在很多人都喜欢听呢。要不要我给你唱一下?

That is singing while speaking. Now many people like listening to it. Do you want me to sing it for you?

女:你一唱歌呀,我就得关门。.

Whenever you sing, I have to close the door.

1. 女的很喜欢听男的唱歌

The woman likes to hear the man singing very much.

2. 女的得关着门听男的唱歌

The woman has to close the door in order to listen to the man singing.

3. 女的和男的都喜欢唱歌

The woman and the man both like singing.

4. 女的不想听男的唱歌

The woman does not like to listen to the man singing. (Correct)

Acknowledgements

I would like to thank Dr. Naoko Taguchi, Dr. Sara C. Weigle, Dr. Dongbo Zhang, Dr. Feng Xiao, and colleagues at Carnegie Mellon University and Beijing Language and Culture University for their generous support at various stages of this project.

Funding

This project was funded by The Center for Urban Language Teaching and Research (CULTR) at Georgia State University (GSU), by GSU's Research Initiation Grant, and by the Humanity and Social Science Youth Foundation of The Ministry of Education of China (Project No. 16YJC740074).

Authors' contributions

Shuai Li is the sole author of this manuscript.

Competing interests

The author declares that he has no competing interests.

Received: 2 November 2017 Accepted: 12 January 2018
Published online: 06 February 2018

References

Bachman, LF, & Palmer, AS (2010). *Language assessment in practice: developing language tests and justifying their use in the real world*. Oxford: Oxford University Press.

Boone, WJ, Staver, JR, Yale, MS (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Deese, J, & Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*(3), 180–187.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment: an advanced resource book. New York, NY: Routledge.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax and Semantics *Vol.* 3 (pp.41–58). Cambridge, MA: Academic Press.

Hanban (2010). *Chinese proficiency test syllabus—level 3*. Beijing: The Commercial Press.

Kane, MT. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Leech, G (2014). *The pragmatics of politeness*. Oxford: Oxford University Press.

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

Morgan, J. L. (1978). Two types of convention in indirect speech acts. In P. Cole (Ed.), Syntax and semantics, *Vol.* 9 (pp. 261-281). New York, NY: Academic Press.

Purpura, J (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Roever, C (2005). *Testing EFL pragmatics*. Frankfurt: Peter Lang.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing*, *23*, 229–256.

Roever, C. (2013). Testing implicature under operational conditions. In S. Ross & G. Kasper (Eds.), Assessing second language pragmatics (pp. 43–64). New York, NY: Palgrave Macmillan.

Roever, C, Fraser, C, Elder, C (2014). *Testing ESL sociopragmatics: development and validation of a web-based test battery*. Frankfurt: Peter Lang.

Ross, S (1997). An introspective analysis of listener inferencing on a second language listening test. In G Kasper, E Kellerman (Eds.), *Communication strategies: psycholinguistic and sociolinguistic perspectives*, (pp. 216–237). New York: Longman.

Sperber, D, & Wilson, D (1995). *Relevance: communication and cognition*, (2nd ed.,). Cambridge: Cambridge University Press.

Taguchi, N. (2005). Comprehension of implied meaning in English as a second language. *Modern Language Journal*, *89*, 543–562.

Taguchi, N. (2007). Development of speed and accuracy in pragmatic comprehension in English as a foreign language. *TESOL Quarterly*, *42*, 313–338.
Taguchi, N. (2008). Pragmatic comprehension in Japanese as a foreign language. *Modern Language Journal*, *92*(4), 558–576.
Taguchi, N. (2009). Corpus-informed assessment of comprehension of conversational implicatures in L2 English. *TESOL Quarterly*, *43*(4), 738–749.
Taguchi, N. (2011). The effect of L2 proficiency and study-abroad experience in pragmatic comprehension. *Language Learning*, *61*, 904–939.
Taguchi, N, Li, S, Liu, Y. (2013). Comprehension of conversational implicature in L2 Chinese. *Pragmatics and Cognition*, *21*(1), 139–157.
Taguchi, N, & Roever, C (2017). *Second language pragmatics*. Oxford: Oxford University Press.
Timpe, V (2013). *Assessing intercultural language learning*. Frankfurt: Peter Lang.
Timpe-Laughlin, V., Wain, J., & Schmidgall, J. (2015). Defining and operationalizing the construct of pragmatic competence: review and recommendations. ETS Research Report Series 2015 (1), 1–43. Princeton, NJ: ETS.
Walters, S. (2009). A conversation analysis–informed test of L2 aural pragmatic comprehension. *TESOL Quarterly*, *43*(1), 29–54.
Wright, BD, & Linacre, JM. (1994). Reasonable mean-square fit values. *Rasch measurement transactions*, *8*(3), 370.
Yamashita, S. O. (1996). Six measures of JSL pragmatics. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.
Youn, S. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(3), 199–225.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com