Non-Expert Raters' Scoring Behavior and Cognition in Assessing Pragmatic Production in L2 Chinese

Shuai Li, Xian Li, Yali Feng, and Ting Wen

Abstract This chapter reports on a study investigating non-expert raters' scoring behavior and cognitive processes involved in evaluating speech acts and pragmatic routines in L2 Chinese. Pragmatic production data were collected from 51 American learners of Chinese, who completed a 12-item oral Discourse Completion Test (DCT). The learners were divided into 15 groups, each including the same six learners and three different learners. A total of 101 non-expert, native Chinese raters evaluated the oral productions of one learner group and were encouraged to verbalize their scoring rationale. Results showed that, although the raters varied significantly in scoring severity, their scoring behaviors were consistent, with very limited instances of scoring bias. Qualitative analysis based on 2753 verbal protocols revealed that the raters predominantly oriented towards criteria related to holistic meaning expression in assessing speech acts and routines. They prioritized criteria related to linguistic expressions (notably those concerning vocabulary knowledge) in evaluating pragmatic routines, and they paid more attention to criteria related to interactional skills in assessing speech acts. Boundary crossing implications are discussed in relation to pragmatics assessment and L2 Chinese teaching.

Keywords Pragmatics assessment · L2 Chinese · Rater cognition · Scoring behavior

S. Li (✉)
Department of World Languages and Cultures, Georgia State University, Atlanta, GA, USA
e-mail: sli12@gsu.edu

X. Li · Y. Feng
Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, USA
e-mail: xli66@gsu.edu; yfeng9@gsu.edu

T. Wen
College of Chinese Studies, Beijing Language and Culture University, Beijing, China
e-mail: wenting@blcu.edu.cn

D. Zhang, R. T. Miller (eds.), *Crossing Boundaries in Researching, Understanding, and Improving Language Education*, Educational Linguistics 58, https://doi.org/10.1007/978-3-031-24078-2_4

1 Introduction

With the publication of the seminal work by Hudson et al. (1995), second language (L2) pragmatics assessment evolved to cross the boundaries of research on interlanguage pragmatics and on L2 performance assessment. Informed by pragmatics theories and research on interlanguage pragmatics, the field has mainly focused on developing instruments for assessing pragmatic competence, leading to expanded construct coverage in pragmatics assessment (Taguchi & Roever, 2017). More recently, thanks to the influence from L2 performance assessment, the field's research agenda has pluralized. Researchers have increasingly paid attention to contingent factors that may affect assessment outcomes, including rater behavior and cognition (e.g., Liu & Xie, 2014; Youn, 2007), rating scale functioning (e.g., Chen & Liu, 2016; Li et al., 2019), and differential item functioning (e.g., Roever, 2007), to name just a few. Among such factors, rater behavior and cognition have attracted much research attention, which reflects sustained interest in this topic in the larger field of L2 performance assessment (for a review, see Han, 2016). Meanwhile, methodological boundary crossing has also characterized the field's development. Whereas early-stage studies adopted a strong psychometric paradigm, more recent research has incorporated additional methodological paradigms such as Conversation Analysis and discursive pragmatics (e.g., Youn, 2015; Walters, 2007).

In L2 pragmatics assessment, researchers typically develop *a priori* rating criteria based on theorizations of pragmatic competence, and recruit what we refer to as *expert raters* who are trained in a closely related academic field (e.g., pragmatics, applied linguistics). Such expert raters have been found to exhibit considerable variability in scoring severity (e.g., Liu & Xie, 2014; Youn, 2007); in interpretation of the substantive meaning of rating criteria (e.g., Li et al., 2019); in scoring bias towards examinees, pragmatic features, and/or assessment items (e.g., Youn, 2007); and in prioritization of certain rating criteria over others (e.g., Taguchi, 2011; Walters, 2007). However, the common practice in the field of only having expert raters evaluate pragmatic performance can be problematic because other potentially relevant stakeholders are left out. Such stakeholders include, for example, native speakers of the target language who are not equipped with the kind of academic training and/or teaching experiences that expert raters have. Such native speakers are the people that L2 learners are supposed to interact with outside the language classroom (e.g., consider the study abroad context), and they are the people who evaluate learners' performance in real world contexts. Hence, it is critical to include such *non-expert raters* into pragmatics assessment.

To date, with only the exception of Taguchi (2011), very little is known about how non-expert raters would assess L2 pragmatic performance and what evaluation criteria they would adopt. Answers to these questions would have boundary-crossing implications: they would allow us to gauge the generalizability of existing findings regarding expert raters' scoring behavior and cognition; such information would also inform L2 instruction and learning by understanding which aspects of linguistic performance are deemed important by potential stakeholders. This study intends

to contribute to this line of research by investigating how non-expert native Chinese speakers evaluate the production of speech acts and pragmatic routines in L2 Chinese.

2 Literature Review

In this section, we start with a brief review of the rating criteria used to evaluate L2 pragmatic performance. We then discuss quantitative studies on raters' scoring behavior. In the spirit of paradigmatic boundary crossing, the review of quantitative research is complemented by a discussion of qualitative studies on rater cognition, because raters' cognitive processes during scoring have been found to be related to their scoring behavior. This section ends with a critique of the existing literature from a boundary crossing perspective.

2.1 Rating Criteria in L2 Pragmatics Assessment

Rating scales with descriptors of evaluation criteria have been widely used to assess L2 pragmatic performance (Taguchi & Li, 2021). Such criteria have developed over time to reflect the evolving theorizations of pragmatic competence (for a recent review, see Li, 2021). In the early stage of rating criteria development, researchers resorted to the understanding of pragmatic competence as consisting of pragmalinguistic and sociopragmatic components. The former refers to the connections between linguistic forms and their pragmatic functions, and the latter concerns the sociocultural rules underlying linguistic behavior of a particular speech community (Leech, 1983; Thomas, 1983). Pragmatics rating criteria during that period mainly addressed considerations of appropriateness, directness, and politeness; other criteria adopted in the early stage of pragmatics assessment research included realization of communicative intention, use of formulaic expressions (for assessing pragmatic routines), and amount of speech/information (e.g., Hudson et al., 1995; Liu, 2006).

While the broadly defined notion of appropriateness has remained in all pragmatics assessment research since Hudson et al.'s (1995) foundational project, researchers have later incorporated additional criteria into assessment. One notable addition was the inclusion of linguistic accuracy in evaluating pragmatic performance (e.g., Chen & Liu, 2016; Grabowski, 2013; Li et al., 2019; Taguchi, 2012), which reflects the close relationship between grammatical and pragmatic competencies (Bardovi-Harlig, 2003). More recent, and significant, additions to pragmatics assessment criteria have been informed by theorizations of interactional competence (e.g., Young, 2011) and discursive pragmatics (Kasper, 2006). Under these perspectives, pragmatic competence is not considered as an individual trait, but rather as an ability that emerges in the process of co-constructing meaning in interaction (Taguchi, 2019). Interactional skills such as turn-taking, topic management,

and repair are thus critical for understanding pragmatic competence and have been incorporated into pragmatics assessment (Timpe, 2013; Youn, 2015).

As discussed above, appropriateness, linguistic accuracy, and interactional skills constitute the major dimensions of pragmatics assessment criteria. Because speech acts have been the main focus in the field, existing rating criteria are best at serving the purpose of assessing speech acts rather than other pragmatic constructs (e.g., pragmatic routines). In actual assessment practice, raters are typically provided with a set of *a priori* rating criteria. While such rating criteria can orient raters to the major dimensions to consider during the scoring process, they do not delineate specific linguistic features that may lead to a higher or lower score. An example is Li et al.'s (2019) study that assessed the production of compliment responses, refusals, and requests. The score descriptor of Band 4 (there were six scoring levels) reads "target communicative function somewhat realized; expression somewhat appropriate for a given scenario (e.g., verbosity, somewhat more direct and/or indirect than needed, use of uncommon semantic formula) as judged by native speaker raters; syntactic and/or lexical errors tend to interfere with meaning and/or appropriateness" (p. 293). Such general descriptions of benchmark performance for this score band leaves plenty of room for interpretation by raters. For example, exactly what linguistic features in examinees' productions constitutes "somewhat appropriate" performance may be quite different across raters' minds. Variability in rater cognition may, in turn, influence their scoring behavior (discussed below).

2.2 *Raters' Scoring Behavior in Assessing L2 Pragmatics: Quantitative Studies*

Quantitative research on raters' scoring behavior has mainly focused on understanding whether raters exhibit similar or different levels of severity, whether they perform scoring consistently (e.g., being consistent in scoring severity), and whether they demonstrate any bias in scoring (i.e., being particularly harsh or lenient for certain examinees, assessment items, and/or pragmatic features). These issues are typically investigated by using the Rasch model, which is a psychometric model widely adopted in L2 performance assessment (McNamara et al., 2019). Based on raw scores, the Rasch model estimates rater severity, examinee ability, and difficulty of test items on a *logit scale*. The logit scale is an interval scale centered at the zero point and extending to positive and negative infinity. The measurement unit on the logit scale is called a *logit*. A larger (or positive) logit value indicates greater severity of raters in scoring, higher ability of examinees, and a higher difficulty level of assessment items, and vice versa. Moreover, the Rasch model outputs separation indices to indicate the number of statistically distinct levels of rater severity, examinee ability, and item difficulty. It also calculates *fit* statistics (called *Mean Square*, or MnSq) to reveal the extent to which the response patterns of individual raters, examinees, and test items conform to the model's expectations. An acceptable range

of MnSq to indicate good model fit is 0.5–1.5 (Wright & Linacre, 1994), although some researchers have also used the more conservative range of 0.7–1.3 (e.g., Liu, 2006). Finally, the Rasch model allows bias/interaction analysis among the variables. For example, it can tell whether a rater gives particularly harsh scores to specific examinees and/or specific items.

Existing studies have predominantly focused on scoring behavior of what we previously referred to as *expert raters*. For example, Youn (2007) studied three expert raters' behavior in scoring speech act production (including apologies, refusals, and requests) in L2 Korean based on Hudson et al.'s (1995) rating criteria (discussed in the previous section). The raters were all native Koreans with graduate training in applied linguistics, and two of them also had relevant teaching experience. Results showed that the raters' scoring behavior conformed to the expectations of the Rasch model, but they significantly differed in scoring severity. The raters showed different bias patterns in assigning scores to individual examinees. Similar findings were reported by Liu and Xie (2014), who recruited both native and non-native English speaker raters (who were all college English instructors) to evaluate written production of apologies by Chinese EFL learners. The raters showed biases in scoring certain examinees, which was likely due to differences in prioritizing certain criteria during the scoring process. For example, some raters considered grammatical knowledge to be critical, but others attached more importance to how examinees realized apologies. Raters' differential interpretation of rating criteria was also reported in Li et al.'s (2019) study, where two expert native speaker raters with shared academic, cultural, linguistic, and professional backgrounds evaluated speech act production (including compliment responses, refusals, and requests) in L2 Chinese.

Collectively, findings of the small number of existing studies suggest that expert raters, in assessing speech act production with a set of *a priori* rating criteria, are generally able to assign scores consistently (i.e., their scoring patterns meet the expectation of the Rasch model), but they often show considerable variation in scoring severity and may exhibit scoring biases towards examinees or assessment items. It is unclear whether non-expert raters without relevant academic training or instructional experience would demonstrate similar scoring behavior. Moreover, raters' scoring bias and varied severity in scoring may be related to their individualized cognitive processes, as demonstrated in Liu and Xie's (2014) study. Variability in rater cognition is an issue often examined in qualitative studies, which are reviewed in the next section.

2.3 Variability in Rater Cognition: Qualitative Studies

Research on rater cognition in L2 pragmatics assessment typically analyses raters' protocols detailing their cognitive processes during scoring in order to investigate which aspect(s) of examinee performance they attend to. The small body of literature has focused on the effects of native speaker status and varied native language

backgrounds on rater cognition, and only one study included non-expert raters (Taguchi, 2011).

Two studies compared rater cognition between native and non-native expert raters (Alemi & Tajeddin, 2013; Walters, 2007). In Walters' (2007) study, two expert raters (native and non-native English speakers) trained in conversation analysis (CA) were recruited. Both raters evaluated one ESL learner's role play of two CA-informed constructs (i.e., assessment, pre-sequence) and one speech act (i.e., compliment) in terms of the level of realization. The two raters discussed discrepancies in their ratings through a series of dialogues, which revealed considerable differences in how they interpreted the same performance. Whereas the non-native rater (who shared the same native language as the learner) cited L1 transfer as a possible explanation of the learners' non-native-like performance, the native rater relied on his intuition for evaluation. Moreover, the non-native speaker also paid attention to fluency and clarity in pronunciation, but the native speaker did not.

While it may be difficult to attribute Walters' findings to native status because of the small sample size of his study, Alemi and Tajeddin's (2013) study demonstrated rater variability between native and non-natives with a larger group of rater participants. The researchers recruited 50 native English raters (who were ESL faculty) and 50 non-native raters (who received M.A. training in applied linguistics and had multiple years of teaching experience) to evaluate refusals in L2 English elicited through a written DCT. The raters evaluated overall appropriateness of the refusal responses and wrote down their scoring rationale. Results showed that the non-native raters were more lenient than their native counterparts. Regarding rater cognition, while the native raters resorted to 11 criteria during the scoring process, the non-native raters only referred to six criteria. Moreover, the two rater groups differed in their predominant evaluation criteria: whereas politeness was the most important consideration among the non-native raters, provision of appropriate reasoning and explanation was the leading criterion among the native raters.

As the aforementioned studies show, expert raters sometimes employ criteria that may not be incorporated in theory-informed pragmatics assessment literature (see the first section of this literature review), such as fluency and pronunciation. This tendency is more clearly shown in Sydorenko, Maynard, and Guntly's (2014) study. Three expert raters (with ESL teaching experience and familiarity with the speech act literature) listened to ESL learners' oral production of multiple-turn requests, evaluated the level of overall appropriateness, and explained their scoring rationale. Results showed that the raters paid attention to the sequential organization of requests, noting the follow-up moves (e.g., thanking, closing) after a request was delivered. The raters also considered the specific contexts in which request utterances occurred, as well as intonational patterns, repetitiveness of speech, and cultural misunderstanding.

While the above studies all focused on expert raters, Taguchi's (2011) study is the only one that included non-expert raters. Similar to Sydorenko et al.'s findings, Taguchi's non-expert raters also paid attention to various aspects of speech act production during scoring, and there were considerable variations in individual raters' cognition. Taguchi's raters were all native English speakers but differed in cultural

backgrounds: there were one African American male, one Australian white male and one female, and one Japanese American female. The raters evaluated appropriateness of two speech acts (i.e., request and opinion) in L2 English, and shared their scoring rationales through individual introspective interviews. Results showed that the raters tended to focus on different dimensions of learner performance as the basis for scoring: some prioritized linguistic forms but others paid more attention to semantic content and strategies. The raters also varied in their level of tolerance for the same aspect of performance. Finally, some raters also resorted to personal experience to support scoring decisions.

2.4 A Boundary Crossing Critique of the Literature, and This Study

Our literature review so far has shown that the mainstream practice of L2 pragmatics assessment typically relies on expert raters and adopts theory-informed, *a priori*, rating criteria that mainly focus on broadly defined dimensions of appropriateness, linguistic accuracy, and interactional skills. Existing studies suggest that such predetermined rating criteria are often open to individualized interpretations when expert raters evaluate specific instances of pragmatic performance. During the scoring process, expert raters are also likely to prioritize certain criteria over others, may attend to features that are not typically assessed in the literature (e.g., fluency), and factor in their personal experiences and/or expectations. Such variability in rater cognition may influence raters' scoring behavior (e.g., severity in scoring, bias in scoring).

From a boundary crossing perspective, several issues need to be addressed. First, the field's predominant focus on expert raters (except for Taguchi's study discussed above) artificially creates a "rater eligibility boundary" in pragmatics assessment based on professional training/knowledge, which underestimates the importance of other potential stakeholders of L2 pragmatics assessment. As Sydorenko et al. (2014) contended, criteria for assessing L2 pragmatic performance should not come exclusively from experts in pragmatics research or experienced language professionals, but also from people who are most likely to interact with the targeted examinee population. We would argue that such people include what we previously referred to as *non-expert raters*, who are not savvy in linguistics or pedagogical theories and may not have rich experiences in interacting with L2 speakers. Such non-expert raters should be included in the practice of pragmatics assessment because, arguably, they are the most likely interlocutors for L2 learners outside the classroom. Second, the field's almost exclusive focus on speech acts (except for Walters' study reviewed above) in understanding rater behavior and cognition reflects and reinforces a "target construct boundary" that is still in place in the larger field of interlanguage pragmatics, as speech acts have long been the most extensively researched pragmatic feature (Taguchi & Roever, 2017). Such a target construct boundary in L2 pragmatics assessment research tends to restrict our

understanding of whether and how raters may adjust their cognitive processes according to different pragmatic features. Incorporating pragmatic features in addition to speech acts is thus in order. Finally, existing research on rater cognition has exclusively focused on English as the target language. This "target language boundary" needs to be crossed given the considerable cultural and linguistic variations among world languages.

Inspired by the boundary crossing spirit of G. Richard Tucker (see Zhang & Miller, this volume), the present study aimed to address the aforementioned issues by focusing on the rating behavior and cognition of non-expert raters who evaluated both speech acts and pragmatic routines in L2 Chinese. We adopted quantitative and qualitative methodological approaches to answer the following research questions.

RQ1. What are the patterns of non-expert raters' scoring behavior?

RQ2. What criteria do non-expert raters adopt to evaluate speech acts and pragmatic routines?

3 Method

3.1 Examinees

Examinee data came from 51 American learners of Chinese recruited from a study abroad program in China. There were 22 males and 29 females, with a mean age of 20.41 years (SD = 0.96). At the time of data collection, the examinees were just starting their study abroad semester. Prior to going abroad, they had received, on average, 2.22 years of formal instruction in Chinese (SD = 1.18). The examinees took the New HSK for placement. The HSK test is a standardized Chinese proficiency test suite consisting of separate tests for six proficiency levels for the written part (tapping listening, reading, and writing) as well as separate tests for three oral proficiency levels (elementary, intermediate, and advanced) (see Peng et al., 2021 for a review of the test). The examinees took the HSK Level 4 written test (score range: 0–300) and the intermediate-level speaking test (score range: 0–100). The mean of the combined test scores was 229.03 (SD = 51.84, range: 142.25–328.75), suggesting that the examinees had roughly intermediate-mid to advanced-mid level of proficiency.

3.2 Instrument

All examinees responded to a 12-item computerized oral DCT consisting of six speech act items representing request (k = 2), refusal (k = 2), and compliment response (k = 2), as well as six pragmatic routine items. The Appendix shows a list of these scenarios. These items came from a larger project assessing pragmatic development in L2 Chinese (e.g., Li et al., 2019; Taguchi et al., 2016; Xiao et al.,

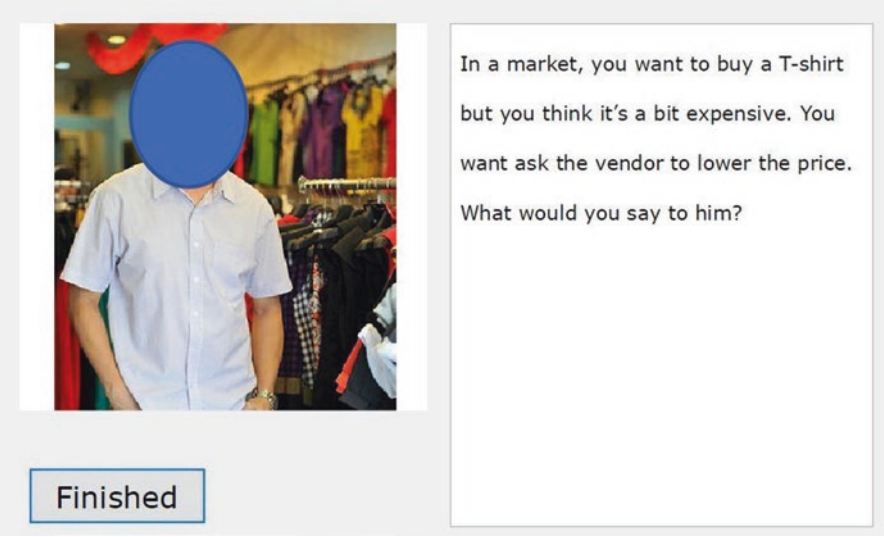

Fig. 1 A screenshot of an Oral DCT item

2019). In responding to the oral DCT, examinees first heard a scenario description in English and at the same time saw a picture illustrating the scene. After the audio was done, a beep reminded the examinees to start saying what they would say in that scenario. There was no time limit for the assessment items. Examinees' oral responses were recorded in the computer, and the audio files were evaluated by a group of non-expert raters (see above Fig. 1).

3.3 Non-expert Raters

A total of 101 non-expert raters were recruited from a major south-eastern city in the US. They were all native Chinese speakers coming from Mainland China and were enrolled in colleges and universities at the time of this study. The mean length of stay in the U.S. was 30.58 months (*SD* = 32.68). There were 64 males and 47 females, with a mean age 24.19 years (*SD* = 4.65 years). None of the raters were in the fields of linguistics or applied linguistics, and all reported no or highly limited experience of interacting with learners of L2 Chinese.

3.4 Rating Procedures

The oral responses of the 51 examinees were assigned to 15 batches. Each batch contained the data of nine examinees, including the same six examinees shared across all batches (which served as anchors for linking different raters' performance

when performing the Rasch analysis) and three different, randomly selected examinees. The 101 raters were randomly assigned to evaluate one of the 15 batches of examinee data; each batch of examinee data was evaluated by seven, eight or nine raters. They saw the same DCT scenarios as the examinees did, listened to the examinees' oral productions one by one, and performed a Yes/No binary judgment on whether an oral response fulfilled the communicative goal as required by each scenario. The raters were also encouraged, but not required, to verbalize their scoring rationale in Chinese after making each judgment. No specific guidelines were given to the raters and they were free to comment on any aspect of the oral responses. The raters' verbal protocols were recorded and later transcribed for analysis.

Before starting their judgment and verbalization task, the raters received a brief warmup exercise, during which they tried several practice items and were familiarized with the verbalization procedure. Data collection was conducted individually for each rater in a quiet room on campus.

3.5 Data Analyses

To answer RQ 1, we collected a total of 10,908 binary judgments (101 raters × 9 examinees for each rater × 12 DCT items for each examinee). All "Yes" judgments were converted to the score of "1" and all "No" judgments the score of "0." Due to two missing data points, the total number of judgments for statistical analysis was 10,906. Out of the 10,906 judgments, 8289 (or 76%) were "Yes", and the remaining 2617 (or 24%) were "No". We built a three-facet Rasch dichotomous model including raters ($n = 101$), examinees ($n = 51$), and oral DCT items ($k = 12$). The quantitative analysis was performed with the software FACETS Version 3.71.3.

To answer RQ 2, due to an unexpected loss of a portion of the verbal protocol data, analysis was based on the data from 81 raters who evaluated 48 examinees. Out of the 8748 potential verbal protocols (i.e., 81 raters x 9 examinees x 12 scenarios), the 81 raters provided 2753 verbal protocols (a 31.47% response rate). Based on these protocols, the three researchers of this study followed a data-driven approach (Youn, 2015) to developing our coding scheme. This involved a bottom-up, iterative procedure by reviewing all verbal protocols in order to extract and refine our codes and the entire coding scheme. The finalized coding scheme included 16 first-order codes (rating criteria), which were grouped into three major categories: *holistic meaning expression*, *linguistic expressions*, and *interaction*. The total instances of coding were 2945. The three researchers went through and discussed all instances of coding together to reach consensus. Following is the coding scheme that provides definitions of each first-order code (rating criteria) with representative examples from our data. Due to space limit, only English translations are provided for the examples. All coding was performed through NVivo Version 12.

1. Holistic meaning expression

   1.1 Comprehensibility of meaning

   Definition: Overall comprehensibility of an utterance, ease of understanding the speaker's intention

   Example #1, for Scenario #1 (Cashier) (see scenario description in the Appendix)

   He has already expressed that he wants to buy a jacket, therefore I could understand it clearly when he was asking where he is going to pay.

   1.2 Incomprehensibility of meaning

   Definition: Overall incomprehensibility of meaning, difficulty in understanding the speaker's intention

   Example #2, for Scenario #7 (Wrong phone call)

   I only understood the part "I am" and didn't understand what he said afterwards.

   1.3 Misunderstanding

   Definition: an utterance that may cause misunderstanding.

   Example #3, for Scenario #5 (Bargain)

   What she said was "this T-shirt is too expensive, and I don't have money", which makes people think that she might not mean to ask the peddler to lower the price; instead, she might not want to buy this T-shirt. Just a little bit like, her expression could cause the peddler to misunderstand what she means and therefore is not willing to continue the conversation with her anymore.

   1.4 Incomplete meaning

   Definition: an utterance that does not fully express the intended meaning by leaving out important information (i.e., lacking semantic formula)

   Example #4, for Scenario #3 (Presentation)

   He didn't express his refusal; he just said that he was sorry, which could be counted as a half refusal, but he did not provide any reasons.

2. Linguistic expressions

   2.1 Code switching

   Definition: an utterance that includes the use of English words/phrases

   Example #5, for Scenario #8 (Restaurant)

   Restaurant waiters, with their English proficiency, won't be able to understand the meaning of "carry away".

   2.2 Word choice

   Definition: an utterance that includes wrongly used word(s)

   Example #6, for Scenario #6 (Photo)

   She chose the wrong verb and said "to make a photo", which expresses a completely different meaning as for "taking photos".

   2.3 Key expression

   Definition: production (or lack thereof) of keyword(s) that renders success (or lack thereof) in meaning expression

Example #7, for Scenario #1 (Cashier)

He mentioned the keywords "to pay", and the salesperson should be able to understand what he meant.

2.4 Incomplete utterance

Definition: an utterance that is syntactically incomplete due to a lack of linguistic knowledge

Example #8, for Scenario #8 (Restaurant)

Because he didn't know how to express the meaning "taking the food away". He just came up directly and asked the waiter, but he didn't know what to say next.

2.5 Grammar

Definition: syntactic or morphosyntactic features of an utterance that may interfere with or enhance meaning expression

Example #9, for Scenario #6 (Photo)

There are some problems with his word order. The adverbial is not put in the correct position, and there is no preposition in the sentence; but we could understand him in communication.

2.6 Pronunciation

Definition: clarity and accuracy of pronunciation that may interfere with or enhance meaning expression

Example #10, for Scenario #1 (Cashier)

First of all, her pronunciation is not accurate. She said that "I want to sell this", "where I can buy it". If I were the clerk, I would ask what you want to sell, and what you meant by saying where to buy. I don't understand what you are talking about.

2.7 Intonation

Definition: intonational features that may interfere with or enhance meaning expression and/or politeness

Example #11, for Scenario #4 (Essay)

"Do you think it is very interesting?". It may not sound very polite to use a rhetorical question in Chinese.

2.8 Fluency

Definition: temporal features that may interfere with or enhance meaning expression

Example #12, for Scenario #10 (Cell phone)

Because she speaks intermittently, I couldn't hear what she was talking about.

2.9 Nativelikeness

Definition: an utterance or expression that may or may not conform to native speakers' intuition

Example #13, for Scenario #5 (Bargain)

What he wants to express should be that the T-shirt is a little expensive, wishing it to be cheaper. He said, "Why is it so expensive", meaning that "this is a little expensive". However, the way he said it is quite different from what we are used to, and we might not be able to understand what he wants to express the moment we hear it in actual communication.

2.10 Politeness

Definition: level of politeness that may interfere with or enhance effectiveness in interaction

Example #14, for Scenario #4 (Essay)

Professor Xiao praised his travel essay, but he said "thank you, do you understand this?" Although he might not intend to offend Professor Xiao, this expression is very offensive to the interlocutor.

3. Interaction

3.1 Turn management

Definition: an utterance that is perceived to connect well or poorly with prior or subsequent turns

Example #15, for Scenario #4 (Essay)

He did not respond to the interlocutor's comments, what he said was a totally different thing from what the interlocutor had said. The interlocutor already asked to discuss this essay with him and expressed that this essay is interesting. But he is still asking "Can I discuss this essay with you?"

3.2 Contextualization

Definition: visualization specific context of communication

Example #16, for Scenario #5 (Bargain)

He expresses that it is very expensive while holding a T-shirt, therefore the peddler should understand that he wants to buy this T-shirt at a cheaper price.

4 Results

4.1 RQ1. Non-expert Raters' Scoring Behavior

RQ1 focused on non-expert raters' behavior of scoring L2 pragmatic performance. Figure 2 is an output graph of the Rasch model. The first column on the left represents the logit scale, on which rater severity, examinee ability, and item difficulty are measured. The second column shows the distribution of rater severity with each asterisk (*) representing two raters and each dot (.) one rater. Harsher raters appear in higher positions than more lenient raters. The third column displays the ability distribution of the 51 examinees. A higher position on the logit scale corresponds to a higher ability level, and vice versa. The last column indicates the distribution of items in terms of difficulty level. More difficult items occupy higher positions on the scale than easier items (e.g., Item #1, Cashier, was the most difficult one).

Rasch calibrated statistics showed that rater severity measures spread across 3.41 logits (i.e., from 1.76 to −1.65 logits), indicating variability in scoring severity among the raters. Indeed, the corresponding rater separation index was 1.92 (or 2.90 strata) with a reliability coefficient of .79, meaning that the raters can be grouped according to three statistically distinct levels of severity. Importantly, all (100%)

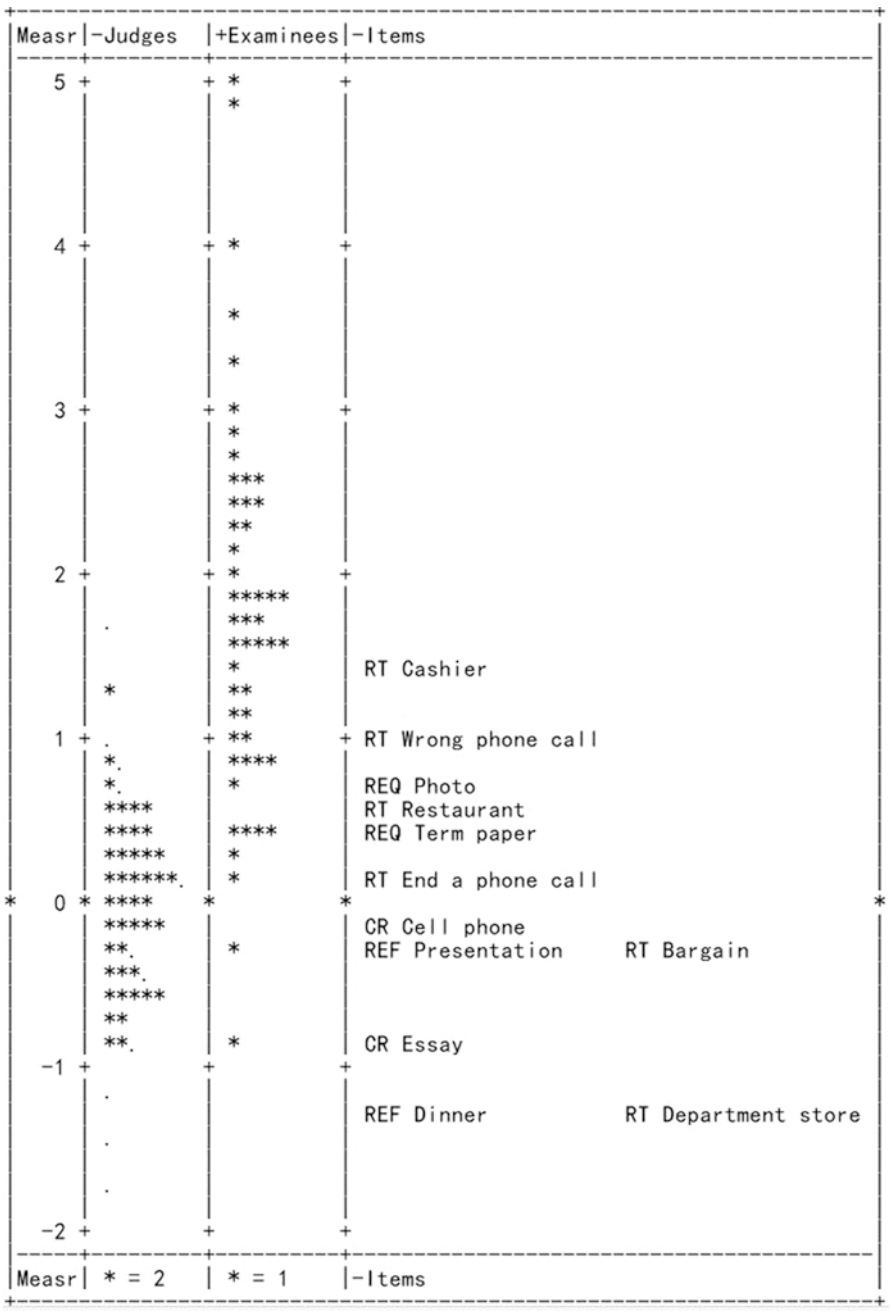

Fig. 2 Wright map. (*Note.* CR Compliment response, REQ Request, REF Refusal, RT Routine)

raters' infit MnSq values fell within the 0.7–1.3 range, suggesting satisfactory model fit, i.e., all raters' judgments conformed to the expectations of the Rasch model. This means that the raters' judgments were consistent. It is also relevant to briefly report the examinee statistics. Examinee ability measures spanned across 6.08 logits (from 5.27 to −0.81 logits), with an average of 1.74 logits (*SD* = 1.19). The examinee separation index was 2.50 (or 3.67 strata) with a reliability coefficient of .86. This means that the examinees could be reliably grouped into more than three distinct ability levels. Moreover, the infit MnSq statistics of 50 out of the 51 examinees (or 98%) fell within the 0.7–1.3 range, suggesting satisfactory model fit of individual examinees' item responses. Turning to the item statistics, the item difficulty measures spread by 2.82 logits (from 1.47 to −1.35 logits).

We further conducted two sets of bias/interaction analyses to examine: (1) whether the raters were more or less severe in scoring individual examinees, and (2) whether they were more or less severe in scoring according to individual items. For the rater x examinee bias/interaction analysis, the purpose was to test the null hypothesis that "there is no statistically discernible bias in each rater's ratings towards individual examinees." Out of 906 bias/interaction terms, only one (or 0.11%) reached statistical significance. This is substantially below the commonly accepted 5% misfit ratio. For the rater x item bias/interaction analysis, we were interested in testing the null hypothesis that "there is no statistically discernible bias in each rater's ratings towards individual items". Out of 1212 interaction terms, 25 (or 2.06%) were statistically significant, which is also below the commonly accepted 5% threshold.

In summary, the non-expert raters' scoring behavior met the expectations of the Rasch model in terms of scoring consistency, yet the raters varied significantly in scoring severity. The raters as a group showed very limited instances of bias in scoring towards individual examinees and/or according to assessment items.

4.2 RQ2. Non-expert Raters' Rating Criteria

RQ 2 examined the criteria that our non-expert raters drew on to evaluate task fulfillment of different speech acts and pragmatic routines. Table 1 displays the frequencies of all first-order codes (i.e., criteria) grouped according to three major categories (i.e., *holistic meaning expression*, *linguistic expressions*, and *interaction*) and for speech acts and routines, respectively. In presenting the findings, we will refer to the examples in the coding scheme (see the Method section).

The percentage statistics in Table 1 show similarities and differences in raters' criteria for assessing speech acts and pragmatic routines. Regarding similarities, *holistic meaning expression* was the most frequently referenced among the three major categories, accounting for 55.97% of the total instances of codes for speech acts and 55.06% for routines. Raters often commented holistically on whether an utterance's meaning was comprehensible (i.e., *comprehensibility of meaning*, see Example #1 in the coding scheme) or incomprehensible (i.e., *incomprehensibility of*

Table 1 Distribution of first-order codes (criteria) between speech acts and pragmatic routines

Categories	First-order codes (criteria)	Speech acts		Routines	
Holistic meaning expression	Misunderstanding	86	6.22%	55	3.52%
	Incomplete meaning	135	9.76%	90	5.76%
	Comprehensibility of meaning	397	28.71%	442	28.30%
	Incomprehensibility of meaning	156	11.28%	273	17.48%
	Subtotal	*774*	*55.97%*	*860*	*55.06%*
Linguistic expressions	Code switching	62	4.48%	133	8.51%
	Word choice	27	1.95%	57	3.65%
	Key expression	49	3.54%	181	11.59%
	Incomplete utterance	16	1.16%	43	2.75%
	Grammar	25	1.81%	12	0.77%
	Pronunciation	72	5.21%	103	6.59%
	Intonation	4	0.29%	4	0.26%
	Fluency	29	2.10%	28	1.79%
	Nativelikeness	10	0.72%	16	1.02%
	Politeness	19	1.37%	15	0.96%
	Subtotal	*313*	*22.63%*	*592*	*37.90%*
Interaction	Turn management	272	19.67%	32	2.05%
	Contextualization	16	1.16%	47	3.01%
	Subtotal	*288*	*22.82%*	*79*	*5.06%*
Uncoded		8	0.58%	31	1.98%
Total		1383	100.00%	1562	100.00%

meaning, see Example #2). To a far lesser extent, raters also based their judgments on whether an utterance might lead to *misunderstanding* (see Example #3) and whether an utterance fully expressed the intended communicative function expected in a specific scenario (i.e., *incomplete meaning*, Example #4).

On the other hand, the non-expert raters differentially drew on the other two larger categories of criteria according to the targeted pragmatic features. As Table 1 shows, raters commented on aspects of *linguistic expressions* more frequently when assessing pragmatic routines (36.94%) than speech acts (21.26%). A closer examination of the individual criteria within this category revealed a nuanced picture. To begin with, the differences between speech acts and routines mainly came from five criteria, and three of these criteria were about vocabulary knowledge: *code switching* (Example #5), *word choice* (Example #6), and *key expression* (Example #7). Table 1 shows that raters referred to *code switching* and *word choice* nearly twice as frequently in assessing routines as in assessing speech acts; the difference in *key expression* was even larger. Another criterion that was used with higher frequency in assessing routines than speech acts was *incomplete utterance* (Example #8). Still another criterion with notable difference between routines and speech acts was *grammar* (Example #9); but this time the frequency was higher for speech acts than for routines. Different from the previous five criteria, the raters showed little difference between routines and speech acts for the following criteria: *pronunciation*

(Example #10), *intonation* (Example #11), *fluency* (Example #12), *nativelikeness* (Example #13), and *politeness* (Example #14). Among these criteria, raters commented on *pronunciation* more frequently than the other criteria.

In terms of the larger category *interaction*, it carried heavier weight for assessing speech acts (22.20%) than for routines (6.02%). Among the three criteria within this category, *turn management* (Example #15) is where there was a major gap between routines (2.05%) and speech acts (19.67%). As it turned out, out of the 272 references to *turn management* under speech acts, compliment response accounted for 94.85%, whereas refusal and request took 5.15% and 0%, respectively. For the remaining criterion *contextualization* (Example #16), raters referred to it more than twice as frequently for routines as for speech acts.

In summary, our non-expert raters relied on three major categories of criteria to evaluate fulfillment of pragmatics tasks involving speech acts and routines. They predominantly focused on the criteria under the larger category *holistic meaning expression* for evaluating both speech acts and pragmatic routines. They appeared to prioritize the criteria under the larger category of *linguistic expressions* when scoring pragmatic routines; meanwhile, they paid more attention to the criteria under the larger category of *interaction* when evaluating speech acts.

5 Discussion

RQ 1 focused on non-expert raters' scoring behavior. The raters varied significantly in scoring severity, yet their scoring performances were highly consistent. Moreover, there were only very limited instances of scoring bias towards individual examinees or items. These findings echo existing research on the scoring behavior of expert raters in assessing L2 pragmatics (e.g., Liu & Xie, 2014; Youn, 2007). Different from previous studies where expert raters were given predetermined rating criteria and received training on scoring, the non-expert raters in this study were not given any uniform, *a priori*, assessment criteria, nor did they receive training on scoring pragmatic performance. Instead, our non-expert raters were free to utilize their own criteria to judge examinees' fulfillment of the pragmatics tasks. In previous studies, expert raters were typically asked to score pragmatic performance according to multiple score bands, which is arguably more cognitively complicated than the binary judgments that our non-expert raters did in this study. Because the binary judgments were relatively straightforward, the raters probably did not need specialized knowledge or training, and could instead rely on their native-speaker intuitions to make judgments. Hence, it was likely that the straightforwardness of the judgment task contributed to the high level of scoring consistency in this study. It would be interesting to examine non-expert raters' scoring consistency based on a rating scale with multiple score bands.

Severity in scoring, on the other hand, showed considerable variation among the 101 non-expert raters, with the rater severity measures spanning across 3.41 logits with a separation index of 1.92 (or 2.90 strata). Because each rater scored only a

subset of the examinees, large individual differences in scoring severity tended to have a major impact on examinee scores. In L2 performance assessment, a typical threshold is that the range of examinee ability is roughly twice (or more) as wide as the range of rater severity; when this threshold is met, the impact on individual raters' scoring severity on examinee test scores is considered as acceptable (Myford & Wolfe, 2000). In pragmatics research focusing on rater behavior, only a few studies reported both rater and examinee statistics, and there are variations across studies in meeting this criterion. For example, Youn (2007) reported a rater severity range of 0.52 logits and an examinee ability range of 0.51 logits, which is way below the threshold; yet Li et al. (2019) found a rater severity range of 0.56 logits and an examinee ability range of 3.75 logits, which is clearly above the threshold. In this study, the examinees' ability range was 6.08 logits, which is nearly twice the range of rater severity (i.e., 3.41 logits). We suspect that a lack of rater training (which was intentional in this study) and raters' personality attributes (i.e., being harsher or more lenient) may have resulted in the variability in rater severity in this study. Our non-expert raters, unlike the expert raters in previous studies (e.g., Li et al., 2019; Youn, 2007), did not have an opportunity to discuss and calibrate their scoring criteria as a group (and it was logistically impractical to do so given the large number of raters recruited). It would be interesting to examine the extent to which non-expert raters' scoring severity can be homogenized by introducing appropriate rater training sessions, which may help ameliorate the influence of personality traits on scoring severity.

RQ 2 examined the non-expert raters' cognitive processes during scoring, focusing on the similarities and differences as they evaluated two different types of pragmatic features, i.e., speech acts and routines. Verbal protocol analysis showed that the raters predominantly oriented towards criteria related to *holistic meaning expression* regardless of pragmatic features. This finding makes sense because the raters were instructed to judge task fulfillment, i.e., whether the intended meaning was conveyed in a specific scenario, which clearly depends on the success in conveying the intended meaning.

Our non-expert raters also paid attention to various criteria under the larger categories of *linguistic expressions* and *interaction*, where there were notable differences between speech acts and pragmatic routines. While Li et al. (2019) demonstrated that individual raters' scoring behavior varied according to different pragmatic features, the results to be discussed here complement their findings by uncovering how raters' underlying cognitive processes may vary based on different pragmatic features. Specifically, under *linguistic expressions*, raters commented on vocabulary knowledge (i.e., criteria of *key expression*, *code switching*, and *word choice*) more frequently for assessing routines than speech acts; the pattern was revised for the criterion *grammar*, which was cited more frequently for evaluating speech acts than routines. These differences likely reflect the unique characteristics of the two pragmatic features and echo existing findings on the acquisition of speech acts and routines (discussed below).

To begin with, speech acts such as requests and refusals typically entail the coordination of various semantic formulae (e.g., providing justifications, thanking, and the focal request/refusal expression *per se*) and the production of syntactically

complex forms. For example, the examinee expression associated with Example #9 was 你可以拍照片我吗*?*(*Could you take my picture?*). This grammatically incorrect utterance that involves the question structure可以…吗*?* (*Could* … question particle?) lacks a complex preposition structure, as the rater pointed out in Example #9. In contrast, pragmatic routines, being fixed or semi-fixed linguistic expressions, are syntactically simpler and semantically less complicated than speech acts. This means that each word in a routine expression plays an important role; oftentimes, one keyword or one short expression could determine the success or failure of producing a pragmatic routine, as Examples #5 and #7 can show. In addition, previous studies on the acquisition of Chinese pragmatic routines reported that an important strategy that learners employed to develop their ability to produce routines was to use core lexical items (e.g., Bardovi-Harlig & Su, 2018; Li et al., in press; Taguchi et al., 2013). Regarding speech acts, researchers have reported that L2 Chinese learners often experienced difficulty in incorporating morphosyntactic and lexical devices into request utterances (e.g., Li, 2014; Wen, 2014), and that learners gradually developed the ability to produce more semantically sophisticated refusals (e.g., Tang et al., 2021). In this study, the non-expert raters were able to intuitively adjust their evaluation criteria and orient to different aspects of examinees' performance according to speech acts and pragmatic routines, and they did so without knowledge of relevant pragmatics theories and/or research findings.

Under the larger category of *linguistic expressions*, our non-expert raters also paid attention to *pronunciation*, *intonation*, *fluency*, *politeness*, and *nativelikeness*. These results corroborate prior research on what expert raters focus on when evaluating speech acts in L2 English (Alemi & Tajeddin, 2013; Liu & Xie, 2014; Sydorenko et al., 2014; Taguchi, 2011; Walters, 2007). While existing studies typically featured only a small number of expert raters, the relatively large number of non-expert raters in this study, along with its focus on Chinese as the target language and on two types of pragmatic features, can add to the generalizability of existing research findings. It is encouraging to know that non-expert raters are able to orient to aspects of L2 pragmatic performance just like expert raters do.

The third larger category that emerged from our protocol data was *interaction*, where there were also considerable differences between speech acts and pragmatic routines. There was a large gap in frequency of reference regarding *turn management* (i.e., 19.67% for speech acts and 2.05% for routines). The relatively frequent comments on *turn management* for speech acts was a bit surprising at first glance, because this study adopted a single-turn oral DCT, which did not allow turn taking or meaning negotiation. In hindsight, this finding was likely due to the characteristics of the speech acts under investigation in this study. In particular, the scenarios involving compliment responses and refusals, by nature, involved a responding turn rather than an initiating turn (which was the case for the request scenarios). Clearly, *turn management* is a key skill in scenarios involving compliment responses and refusals because task fulfilment depends on examinees' ability to produce a turn that connects naturally and sensibly to the previous turn, as Example #15 can show. Indeed, 94.85% of our raters' comments on turn management were found in the compliment response scenarios and 5.15% in the refusal scenarios. On the other hand, none of the routine scenarios necessitate a responding turn, which could

explain why *turn management* carried much lighter weights in our raters' evaluation of routines.

The last criterion where there was a notable difference between speech acts and routines was *contextualization* (under the larger category of *interaction*). Raters cited this criterion more than twice as frequently in assessing routines (3.01%) as in assessing speech acts (1.16%). As Example #16 can show, even though the scenario description and the accompanying photo (Fig. 1) do not indicate or show a speaker holding a T-shirt in hand, trying to bargain a sales price with the street vendor, the rater was able to mentally visualize the specific scene by drawing on personal experiences and/or observations. Comparatively speaking, *contextualization* was more prominent among routine scenarios than speech act scenarios. This is likely because routines, by definition, are tied to specific contexts of communication, i.e., there is a relatively fixed connection between a routine expression and a particular scenario – Kecskes (2016) even coined the term *situationally bound utterances* to refer to pragmatic routines. In comparison, pragmalinguistic forms of speech acts can often be used across different scenarios, thus the connection between pragmalinguistic form and context is weaker than that between routines and context. It would therefore be easier for raters to visualize a specific scene for routines than for speech acts.

6 Conclusions and Boundary Crossing Implications

In crossing the boundaries regarding rater eligibility, target construct, and target language, as identified in the literature review section, this study represented an initial effort to examine non-expert raters' scoring behavior and cognition involved in assessing pragmatics in L2 Chinese. Concerning scoring behavior, despite considerable variability in judgment severity, the non-expert raters performed scoring consistently, with very limited instances of scoring biases. Concerning rater cognition, the raters were primarily oriented to *holistic meaning expression* in judging examinees' task fulfillment regardless of pragmatic features. However, they focused more on criteria related to *linguistic expressions* (notably those related to vocabulary knowledge) in evaluating pragmatic routines than speech acts, and more on criteria related to *interaction* (notably the criterion of *turn management*) when assessing speech acts than routines. Such variability in rater cognition according to targeted pragmatic features can be explained by the characteristics of speech acts and pragmatic routines.

By crossing multiple boundaries, this study can have practical implications for pragmatics assessment and L2 teaching in general. In crossing the rater eligibility boundary by focusing on non-expert raters (in contrast to previous studies' predominant focus on expert raters), this study demonstrates that untrained native speakers (of Chinese), as important stakeholders of pragmatics assessment, are actually able to evaluate L2 pragmatic performance and achieve satisfactory scoring quality, provided that the scoring task is relatively straightforward and that the stakes of the intended pragmatics assessment are relatively low. Including

non-expert native speaker raters in the process of pragmatics assessment should enhance the validity of score interpretation in the assessment context of the target language community, because these are the people that L2 learners are most likely to interact with. Furthermore, in light of the potential of Chinese becoming a lingua franca (Gil, 2021), it would be worthwhile to further cross the rater eligibility boundary by investigating the feasibility of including non-expert, non-native speakers of Chinese who use the language for their professions.

Moreover, in crossing the target construct boundary by including pragmatic routines in addition to speech acts, results of this study indicate that non-expert raters adjust evaluation criteria according to different pragmatic features. Hence, if a goal of L2 pragmatics assessment is to inform examinees of their strength and weakness when they interact with potential interlocutors, it would be important to develop evaluation criteria according to targeted pragmatic features and, perhaps also adjust the weights of such criteria accordingly. These issues would not have surfaced in this study, and would not inform future studies, if we or L2 pragmatics researchers alike were limited by the target construct boundary and focused predominantly on speech acts. Future research can continue to cross the target construct boundary by including more varied pragmatic features in investigating rater cognition and scoring behavior.

Finally, implications of our findings can also cross the boundary of pragmatics assessment to inform L2 (Chinese) teaching in general. The fact that our non-expert raters paid predominant attention to criteria under *holistic meaning expression* highlights the importance of focusing on communicative function (i.e., expressing intended meaning) in L2 instruction. While formal aspects of linguistic expressions do matter in the evaluation of task fulfillment, vocabulary knowledge and, to a lesser extent, pronunciation skills appear to be more important than grammatical knowledge based on our rater protocol analysis. While Chinese language instructors often tend to emphasize grammatical structures in instruction, our findings suggest that grammatical accuracy may only play a very minor role in determining the success of getting one's message across. Moreover, skills such as turn management, which is often not emphasized in (Chinese) language classrooms or in textbooks, should be highlighted to various degrees according to instructional targets (e.g., speech acts vs. pragmatic routines).

Appendix: List of 12 Scenarios

Item numbers indicate order of appearance in the Oral DCT.

Speech act scenarios

Compliment response items

#4 (Essay) You wrote an essay about your travel experience and submitted it to Professor Xiao's class. Today, you meet him in the hallway and you start to

talk to each other. During your conversation, Professor Xiao says: "Oh, by the way, I read your essay and it is really interesting." What would you say to him?

#10 (Cell phone) You meet your friend Xiao Wang in the hallway. Xiao Wang sees your newly purchased cell phone and says: "Is this your new cell phone? It looks really fancy!" How would you respond to Xiao Wang?

Refusal items

#2 (Dinner) You meet your friend Xiao Li after class. Xiao Li invites you to dinner with his friend but you don't want to go. What would you say to Xiao Li?

#3 (Presentation) You come to Professor Li's office to ask a few questions. Before you leave, she asks you to do your presentation one week earlier than you originally scheduled. However, you don't want to do that. What would you say to Professor Li?

Request items

#6 (Photo) You meet your friend Xiao Li at a party today. You want to ask Xiao Li to take your picture. What would you say to him?

#12 (Term Paper) Today is the deadline for submitting your term paper, but you don't have it finished because you were sick. So you want to ask Professor Sun for an extension. Now you come to Professor Sun's office. What would you say to him?

Pragmatic routine scenarios

#1 (Cashier) At a department store, you cannot find where the cashier is. You want to ask this shop assistant for this. How would you ask him?

#5 (Bargain) In a market, you want to buy a T-shirt but you think it's a bit expensive. You want to ask the vendor to lower the price. What would you say to him?

#7 (Wrong phone call) When you answer your phone, you hear a young man's voice. Obviously, he dialed your number by mistake. What would you say to him?

#8 (Restaurant) In a restaurant, you want to take the leftovers with you. What would you say to this waitress?

#9 (End a phone call) You and your friend are talking on the phone. It seems that you both have said all you want to say, so you would like to end your conversation. What would you say to her?

#11 (Department store) In a department store, a shop assistant asks whether you would like to buy anything. You do not intend to buy anything. What would you say to her?

References

Alemi, M., & Tajeddin, Z. (2013). Pragmatic rating of L2 refusal: Criteria of native and nonnative English teachers. *TESL Canada Journal, 30*(7), 63–81.

Bardovi-Harlig, K. (2003). Understanding the role of grammar in the acquisition of L2 pragmatics. In A. Martinez-Flor, E. U. Juan, & A. F. Guerra (Eds.), *Pragmatic competence and foreign language teaching* (pp. 25–44). Universitat Jaume I.

Bardovi-Harlig, K., & Su, Y. (2018). The acquisition of conventional expressions as a pragmalinguistic resource in Chinese as a foreign language. *The Modern Language Journal, 102*(4), 732–757.

Chen, Y., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDCT and e-mail tasks. *Language Assessment Quarterly, 13*(3), 231–250.

Gil, J. (2021). *The rise of Chinese as a global language: Prospects and obstacles*. Palgrave Macmillan.

Grabowski, K. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149–171). Palgrave Macmillan.

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Working Papers in TESOL & Applied Linguistics, 16*(1), 1–24.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics (technical report #7)*. University of Hawaii, Second Language Teaching & Curriculum Center.

Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K. Bardovi-Harlig, J. C. Felix-Brasdefer, & A. S. Omar (Eds.), *Pragmatics and language learning: Vol. 11* (pp. 281–314). University of Hawai'i at Manoa: National Foreign Language Resource Center.

Kecskes, I. (2016). Situation-bound utterances in Chinese. *East Asian Pragmatics, 1*(1), 108–126.

Leech, G. (1983). *Principles of pragmatics*. Longman.

Li, S. (2014). The effects of different levels of linguistic proficiency on the development of L2 Chinese request production during study abroad. *System, 45*, 103–116.

Li, S. (2021). Pragmatics assessment in English as an international language (EIL). In Z. Tajeddin & M. Alemi (Eds.), *English as an international language: Pragmatic pedagogy (pp. 191–211)*. Routledge.

Li, S., Taguchi, N., & Xiao, F. (2019). Variations in rating scale functioning in assessing pragmatic performance in L2 Chinese. *Language Assessment Quarterly, 16*(3), 271–293.

Li, S., Taguchi, N., & Xiao, F. (in press). Effects of proficiency on the development of pragmatic routine production in L2 Chinese. In F. Xiao (Ed.), Second language Chinese development: A longitudinal perspective. Lexington Books.

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of Chinese EFL learners*. Peter Lang.

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing, 4*(1), 50–65.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

Myford, C., & Wolfe, E. (2000). *Monitoring sources of variability within the test of spoken English assessment system* (Vol. 2000, pp. i–51). Educational Testing Service.

Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, mulit-purpose proficiency test. *Language Testing, 38*(2), 326–337.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189.

Sydorenko, T., Maynard, C., & Guntly, E. (2014). Rater behaviour when judging language learners' pragmatic appropriateness in extended discourse. *TESL Canada Journal, 32*(1), 19–41.

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics, 21*(3), 453–471.

Taguchi, N. (2012). *Context, individual differences, and pragmatic competence*. Multilingual Matters.

Taguchi, N. (2019). Second language acquisition and pragmatics: An overview. In N. Taguchi (Ed.), *The Routledge handbook of second language acquisition and pragmatics* (pp. 1–14). Routledge.

Taguchi, N., & Li, S. (2021). Contrastive pragmatics and second language pragmatics: Approaches to assessing L2 speech act production. *Contrastive Pragmatics, 2*(1), 1–23.

Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford University Press.

Taguchi, N., Li, S., & Xiao, F. (2013). Production of formulaic expressions in L2 Chinese: A developmental investigation in a study-abroad context. *Chinese as a Second Language Research, 2*(1), 23–58.

Taguchi, N., Xiao, F. & Li, S. (2016). Development of pragmatic knowledge in L2 Chinese: Effects of intercultural competence and social contact on speech act production in a study abroad context. *The Modern Language Journal, 100*(4), 775–796.

Tang, X., Taguchi, N., & Li, S. (2021). Social contact and speech act strategies in a Chinese study abroad context. *Study Abroad Research in Second Language Acquisition and International Education, 6*(1), 3–31.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*, 91–111.

Timpe, V. (2013). *Assessing intercultural communicative competence*. Peter Lang.

Walters, S. F. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24*(2), 155–183.

Wen, X. (2014). Pragmatic development: An exploratory study of requests by learners of Chinese. In Z. Han (Ed.), *Studies in second language acquisition of Chinese* (pp. 30–56). Multilingual Matters.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370–371.

Xiao, F., Taguchi, N., & Li, S. (2019). Effects of proficiency sub-skills on pragmatic development in L2 Chinese study abroad. *Studies in Second Language Acquisition, 41*(2), 469–483.

Youn, S. (2007). Rater bias in assessing the pragmatics of KFL learners using facets analysis. *Second Language Studies, 26*(1), 85–163.

Youn, S. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in language learning and teaching* (pp. 426–443). Routledge.