Papers in Language Testing and Assessment, 2018, Volume 7, Issue 1, 1-31

The gap between communicative ability measurements: General-purpose English speaking tests and linguistic laypersons' judgments

Takanori Sato
Center for Language Education and Research,
Sophia University, Tokyo, Japan

The assessment criteria for general-purpose speaking tests are normally produced from test developers' intuition or communicative competence models. Therefore, the ideal second language (L2) communication in general-purpose speaking tests reflects language specialists' perspectives. However, neglect of the views of non-language specialists (i.e., linguistic laypersons) on communication is problematic since these laypersons are interlocutors in many real-world situations. This study (a) investigated whether L2 speakers' results on general-purpose speaking tests align with linguistic laypersons' judgments of L2 communicative ability and (b) explored performance features that affect these judgments. Twenty-six post-graduate students of non-linguistic disciplines rated 13 speakers' communicative ability on general-purpose speaking tests and provided verbal explanations of the performance features affecting their ratings. Their ratings were compared with the speakers' test results, and the features that determined their ratings were examined. Although these ratings were not completely different from the test results, some speakers' test results did not align with their ratings. The linguistic laypersons' judgments were affected not only by features that the general-proficiency tests assessed but by other factors as well. The findings of this study will deepen our understanding of real-world interlocutors' views on communication and contribute to the development of authentic criteria for general-purpose speaking tests.

Key words: general-purpose speaking test; linguistic laypersons; indigenous assessment criteria

Email address for correspondence: taka-sato@sophia.ac.jp

Introduction

General-purpose English speaking tests aim to measure second language (L2) speakers' overall communicative ability, and the test scores are regarded as evidence of test-takers' ability to communicate in English in various unspecified real-life domains (Davies et al., 1999). For example, the General English Proficiency Test (GEPT) (Roever & Pan, 2008) and Graded Examinations in Spoken English (GESE) (Trinity College London, 2009) are designed to measure English proficiency applicable to a wide range of language domains, including leisure, education, and business. They can be contrasted with language tests for narrower purposes (Green, 2014), including language for specific purposes (LSP) tests such as the Occupational English Test (OET) (McNamara, 1996) and English for academic purposes (EAP) tests such as the Test of English as a Foreign Language (TOEFL).

General-purpose speaking test assessment criteria are determined by language specialists or theoretical communicative competence models (Bachman & Palmer, 2010; Fulcher, 2010), and the perspectives of non-language specialists or *linguistic laypersons* have rarely been incorporated into general-purpose language assessments. Neglect of their views on communication is problematic as in reality, these laypersons will be the ultimate arbiters of the L2 learners' oral performance in many situations (Elder, McNamara, Kim, Pill, & Sato, 2017). When L2 speakers use English outside the testing context, their interlocutors are often not language specialists (i.e., language teachers, linguists, and accredited raters of language tests); instead, their interlocutors are likely to be linguistic laypersons, who have no experience in teaching and judging language as their occupation, including students majoring in science, history teachers, medical doctors, and sales clerks. Therefore, Brindley (1991) argues that linguistic laypersons' perspectives on communication could profitably be incorporated into the assessment criteria of speaking tests.

This study investigated linguistic laypersons' judgments of L2 performance on general-purpose speaking tests and examined if the test results reflect their judgments. Findings of this study will deepen our understanding of real-world interlocutors' views on communication and contribute to the development of assessment criteria for general-purpose speaking tests that capture performance features valued outside the testing milieu.

Literature Review

Communicative Ability Measured by General-purpose Speaking Tests

General-purpose speaking tests' assessment criteria reflect the construct measured by the tests. Because they are usually based either on language specialists' intuitions

(Fulcher, 2010) or on theories of L2 communication ability (Bachman & Palmer, 2010), the multiple components of language knowledge and interactional competence (Canale & Swain, 1980; Kramsch, 1986) are typically the main focus of the assessment. Davies et al. (1999) list the assessment criteria typically used by oral proficiency tests as pronunciation, fluency, accuracy, and appropriateness; the Common European Framework of Reference (CEFR: Council of Europe, 2001) regards range, accuracy, fluency, interaction, and coherence as the components of spoken language. For example, one such general-purpose test, Pearson Test of English General, which aims to measure test-takers' communicative ability and provide evidence of proficiency in practical language skills, uses the analytic rating criteria of fluency, interaction, range, accuracy, and phonological control (Pearson Education, 2012).

McNamara (1996) calls tests that focus mainly on assessing the quality of language performance *weak performance tests*, as they do not necessarily reflect the criteria on which successful performance is judged in everyday interactions. *Strong performance tests*, on the other hand, assess test-takers' performance using real-world criteria and non-linguistic factors that contribute to communicative success. McNamara (1996) goes on to state that that the majority of general-purpose speaking tests, including the American Council on the Teaching of Foreign Language (ACTFL) Oral Proficiency Interview, are weak performance tests. Even performance tests that simulate real-world tasks may not necessarily focus on the criteria actually used in the real-life target language use (TLU) domain and are still likely to assess language knowledge components and interactional competence as defined by theoretical models. Harding (2014) argues that while much mainstream language testing utilizes real-life tasks and is considered communicative in this sense, non-linguistic factors are not regarded as an important component of the construct measured by these tests.

Linguistic Laypersons' Views on Communicative Ability

The perspectives of non-language specialists on communication have been rigorously addressed in LSP testing. Numerous studies have investigated linguistic lay judges' views by comparing their judgments of L2 speakers' communication and scores on various LSP tests; the linguistic laypersons included health professionals (Lumley, 1998), tour guides (Brown, 1995), non-language subject teachers (Elder, 1993), and undergraduate students (Bridgeman, Powers, Stone, & Mollaun, 2012; Kang, 2012). These studies indicated that language specialists and non-language specialists judge L2 communicative effectiveness differently. For example, Brown (1995) found that non-language specialists (tour guides) were more tolerant of test-takers' weaknesson some linguistic aspects such as grammar, expression, vocabulary, and fluency. In contrast, Kang (2012) demonstrated that undergraduate students without language

teaching experience rated international teaching assistants' English language skills more harshly. Furthermore, Elder (1993) investigated non-language subject teachers' evaluation of the communicative effectiveness of teacher trainees' pedagogical practices and found that the subject teachers focused on behaviours that facilitate student learning rather than trainees' linguistic form. Similarly, Lumley (1998) found that linguistically-oriented criteria used by the OET (intelligibility, fluency, comprehension, appropriateness, grammar, and expression) may underrepresent health professionals' perspectives such that the exam does not fully reflect the domain's communicative needs (see also Wette, 2011).

Other studies in LSP testing have addressed linguistic laypersons' unique perspectives on communication using a different approach: exploratory research on indigenous assessment criteria. Indigenous assessment criteria are criteria 'used by subject specialists in assessing the communicative performances of apprentices in academic and vocational fields' (Douglas, 2000, p. 68). Researchers have sought various non-language specialists' unique perspectives on communication within specific communities (Abdul Raof, 2011; Douglas & Myers, 2000; Fulcher, Davidson, & Kemp, 2011; Jacoby, 1998; H. Kim & Elder, 2015; Pill, 2016; Plough, Briggs, & Van Bonn, 2010). Table 1 summarizes examples of their indigenous assessment criteria. Note that the assessment criteria derived from such non-language specialists are considerably different from the conventional linguistically-oriented speaking test criteria used by weak performance tests. Although linguistic aspects of speakers' oral output, such as pronunciation and interaction, also influence non-language specialists' judgments, they are not their sole or central concerns. In contrast, non-linguistic factors, including non-verbal aspects of performance, affect their perspectives when evaluating L2 communicative ability.

Table 1. Indigenous Assessment Criteria Derived from Various Non-language Specialists

Research	Non-language specialists	Example assessment criteria
Jacoby (1998)	Physicists	Timing, significance of research, content accuracy, linguistic errors in visual aids
Douglas and Myers (2000)	Veterinary	Demeanor, knowledge base, timing, coverage, appearance
Plough et al. (2010)	Faculty evaluators of graduate student instructors	Listening comprehension, pronunciation
Abdul Raof (2011)	Engineering specialists	Delivery, organization, subject matter
Fulcher et al. (2011)	Market customers	Rapport
Kim and Elder (2015)	Pilots and air traffic controllers	Interactions, comprehension, pronunciation
Pill (2016)	Health professionals	Clinical management, management of interaction

In contrast to the volume of work on specific-purposes tests, fewer studies have investigated general-purpose speaking tests with regard to linguistic laypersons' judgments. Nonetheless, some research has been conducted in this area. Researchers have investigated linguistic laypersons' perspectives on communicative ability by comparing speakers' scores on general-purpose speaking tests and ratings given by laypersons. Barnwell (1989) compared L2 Spanish speakers' scores on ACTFL oral interviews and ratings given by untrained native Spanish speakers who used the ACTFL rating scales. Although linguistic laypersons exhibited similar ratings to test scores, the criteria on which they based their judgments of communicative ability appeared to be different from those of the speaking tests. Comments collected after their rating revealed that they attended to factors the tests did not assess, such as speakers' efforts to convey a message. Similarly, Galloway (1980) asked linguistic lay Spanish native speakers to rate 3.5-minute monologues by Spanish L2 learners using pre-defined rating criteria. Comments given by the lay judges after ratings indicated that they attended to the content of the monologues. The features explored by these studies—efforts and content—are often considered beyond the scope of general-purpose language tests and regarded as construct-irrelevant variance (Messick, 1989). However, these non-linguistic features can be relevant and influential to linguistic laypersons' judgments of L2 communicative ability.

Investigating Linguistic Laypersons' Judgments of L2 Performance on General-purpose Speaking Tests

As discussed above, many LSP testing studies have investigated various linguistic laypersons' views on communication and demonstrated that there is dissonance between their views and traditional linguistically-oriented criteria, but few studies have addressed whether linguistic laypersons' judgments of L2 speakers' communicative ability align with scores on general-purpose speaking tests. Moreover, much is unknown about the criteria used by linguistic laypersons when assessing L2 speakers' oral performances on these tests. Since non-language specialists are the ultimate arbiters of communicative effectiveness in many real-world contexts (Elder et al., 2017), their judgments of L2 speakers' performance on general-purpose proficiency tests need to be addressed to confirm whether the ability measured by the tests reflects their perspectives on communicative effectiveness. Although several studies have touched upon this issue (Barnwell, 1989; Galloway, 1980; Hadden, 1991), their participants were asked to assess test-takers' oral performances using the rating scales used by general-purpose tests. Thus, their judgments were largely affected by the pre-established assessment criteria and may differ from how they would assess performance in real-life situations. It is necessary to elicit judgments of L2 communicative ability without the constraint of pre-established criteria in order to

investigate linguistic laypersons' interpretation of this ability more accurately (Zhang & Elder, 2011).

The Current Study

The current study investigated whether the results of general-purpose speaking tests reflect the intuitive judgments of linguistic laypersons. This study also focused on non-criterion features that affect these judgments since they are possible causes of dissonance between test results and rater judgments. The research questions (RQs) were as follows:

1. Do general-purpose speaking test results align with linguistic laypersons' intuitive judgments of L2 communicative ability?
2. What non-criterion features affect linguistic laypersons' judgments of L2 communicative ability?

Methodology

Participants

Raters

Twenty-six linguistic laypersons participated in this study as raters. Linguistic laypersons were defined as individuals without (a) specialized applied linguistics knowledge, (b) training in language assessment and teaching, and (c) formal experience of rating and teaching L2 learners. The raters were a convenience sample. The researcher first approached post-graduate students in non-linguistic disciplines at an Australian university. Then the researcher asked whether they had any language teaching training and teaching/rating experience. Only those who had no language teacher/rater training and formal language teaching/rating experience were included as rater participants. Following this sampling procedure, 26 participants were recruited. However, three raters were excluded from the study because after data collection, it was found that two of them had misunderstood the researcher's instructions and one of them had failed to disclose her extensive language teaching experience. As a result, 23 raters were ultimately included in the study.

This study investigated views of not only native English speakers (NESs) but also non-native English speakers (NNESs) with a variety of first languages (L1s), because English is currently used as an international lingua franca by a substantial number of NNESs (Jenkins, 2015; Seidlhofer, 2011), and neglect of their views on communication

was deemed problematic. To elicit perspectives from lay judges with various L1 backgrounds, the 23 raters represented all of Kachru's (1988) concentric circles (see Table 2). The raters from the Expanding and Outer Circles exhibit a range of English proficiency (IELTS 6.5 to 8.5) and distinct English learning experiences. Certain raters' language proficiency was similar to that of the L2 speakers they were asked to evaluate. However, this study included them as raters because they had been studying in Australian post-graduate programs for more than six months, and so it was assumed that they could understand the L2 performances that they were asked to judge. This study did not separate the views of raters with different language backgrounds because that comparison was not the purpose of this study. However, previous research has found that trained raters from different circles may assess oral performance from different perspectives, although their ratings do not widely differ (e.g., Y.-H. Kim, 2009; Wei & Llosa, 2015; Zhang & Elder, 2011).

Table 2. Genders and Nationalities of the Raters (N=23)

Gender	Female: 15 Male: 8	
Nationality	Expanding Circle:	Chinese, French, Indonesian, Iranian, Italian/German, Japanese, Nepalese, Spanish, Vietnamese (N=10: Raters A to J)
	Outer Circle:	Bangladeshi, Filipino, Indian, Kenyan, Malaysian, Singaporean (N=6: Raters K to P)
	Inner Circle:	American, Australian, British, Canadian, New Zealander (N=7: Raters Q to W)

Although including only post-graduate students as raters may inevitably limit the generalizability of findings, they were selected because the data collection procedure (verbal protocol methods) was cognitively demanding and required high verbal skills. As Lumley (2005) claims, verbal protocol participants should be selected carefully since 'one requirement is that they *can* articulate their thoughts at a metacognitive level' (p. 74) [emphasis in original]. It was assumed that post-graduate students had sufficient cognitive ability and vocabulary to generate metacognitive verbal protocols.

The raters were asked to rate their general familiarity with foreign accents using a five-point scale (1=not familiar, 5=very familiar). The average rating was 3.7, indicating that the raters were at least moderately familiar with NNESs' English. Raters showed different levels of familiarity with the specific English accents of the L2 speakers in the study, and some raters had learned the speakers' first language (see Appendix A). This study did not control for raters' familiarity and language learning experience, although these factors affect the comprehensibility of the speaker messages (e.g., Gass & Varonis, 1984; Harding, 2011; Huang, Alegre, & Eisenberg, 2016). This indicates that individuals who are more or less familiar with the accents of the speakers might perceive their communicative ability differently.

Speakers

This study used 13 speakers' oral performances from two general-purpose speaking tests: the College English Test-Spoken English Test (CET-SET) and the suite of Cambridge English Examinations. These tests were selected because their constructs, assessment criteria, and elicitation tasks are similar to many general-purpose speaking tests. This study used monologues given by seven test-takers of the CET-SET and dialogues given by six test-takers of the Cambridge English Examinations. It was assumed that raters pay attention to different features when judging oral performances with and without an interlocutor. For example, it was considered that interactional competence features (e.g., turn-taking, topic initiation) would affect rater judgments in dialogues but not in monologues. The videotaped performance data were provided by the National College English Committee and Cambridge English Language Assessment. The test sites were not identified.

The CET-SET is the spoken component of the College English Test, which is administered to Chinese college students. Oral performances are assessed by the following analytic rating criteria: (a) *Accuracy*, (b) *Range*, (c) *Size* (of contribution), (d) *Discourse Management*, (e) *Flexibility* (in dealing with various situations and topics), and (f) *Appropriateness* (Zhang & Elder, 2009, 2011). Trained raters give a score from 1 to 5 on each criterion, and the total score is calculated with the following equation: total score = (a + b) × 1.2 + (c + d) × 1.0 + (e + f) × 0.8. The total score is then converted into one of seven levels (from high to low: A^+, A, B^+, B, C^+, C, and D). The test employs a face-to-face group format, where three test-takers sit in a group. Furthermore, the test consists of three parts. In Part 1, test-takers answer warm-up questions on a topic used throughout the entire test. In Part 2, each test-taker is asked to make a 1.5-minute presentation on a given prompt after a one-minute preparation period. Following this, three test-takers are asked to discuss the presentation topic. In Part 3, each test-taker is asked further questions regarding the discussion topic.

The study used video-taped presentation data from Part 2 of the exam, because each presentation lasted more than one minute and was considered sufficient for the raters to judge speakers' monologues. Seven speakers were selected from the pool of CET-SET data (see Table 3). They had different overall scores based on the entire test and were chosen to collect the raters' judgments on a range of proficiency levels. Given that students take the CET-SET after completing mandatory English courses in their university and obtaining an advanced-level score on the written component of the CET (Zhang & Elder, 2009), the speakers were considered roughly low-intermediate to advanced English learners.

Table 3. Seven Speakers of Monologues

Speaker	Gender	Presentation topic	Duration (min.)
CET A+	Female	How to cope with air pollution	1:27
CET A	Male	Causes of air pollution	1:41
CET B+	Female	Consequences of air pollution	1:26
CET B	Male	Causes of air pollution	2:01
CET C+	Female	How to cope with air pollution	1:29
CET C	Female	Consequences of air pollution	2:00
CET D	Male	Causes of air pollution	1:20

Note. These speakers are all Chinese nationals.

Although the CET-SET data also included dialogues from the other parts of the test, this study did not use them because they lacked interaction between the test-takers and appeared more like short monologues. Thus, in addition to the monologues from the CET-SET, the study used test-takers' interactive performances from the suite of Cambridge English Examinations. This study used the speaking sections from three intermediate tests: the Certificate in Advanced English (CAE), the First Certificate in English (FCE), and the Preliminary English Test (PET). Among the tests, the CAE is the most advanced, and the PET is the least advanced. The FCE and PET employ the same five criteria: *Grammar and Vocabulary*, *Discourse Management*, *Pronunciation*, *Interactive Communication*, and *Global Achievement* (University of Cambridge Local Examinations Syndicate, 2009). The CAE assessment criteria are similar to those of the FCE and PET except that the *Grammar and Vocabulary* section is divided into two sections: *Grammatical Resource* and *Lexical Resource*. Fluency is not independently assessed and is evaluated under *Discourse Management*. The tests' speaking section employs a peer-peer paired format. The tests include a conversation between two test-takers and an examiner, an individual test-taker's picture description, and a discussion about issues related to the picture description.

A pair of speakers was selected from each of the CAE, FCE, and PET exams (see Table 4). Each speaker took only one Cambridge Examination, not all three. In the CAE and FCE, test-takers are given a visual stimulus with several pictures and then instructed to discuss each picture. They are required to negotiate with each other and make a decision related to the visual stimulus; the test tasks are designed to elicit a number of different functions, such as sustaining the interaction, exchanging ideas, justifying opinions, agreeing or disagreeing, suggesting, speculating, evaluating, and reaching a decision through negotiation (University of Cambridge Local Examinations Syndicate, 2007). In the PET, two test-takers are asked to engage in a casual conversation on a previously introduced theme, allowing test-takers to discuss

various topics such as their preferences, experiences, and habits. Test-takers are expected to talk about their interests and give reasons for their views. All the speakers whose data were used in this study passed the tests with a satisfactory score and received certificates of equivalent CEFR level. Accordingly, it was assumed that their test levels reflected their English language ability. There were no large test score discrepancies within each pair (see Appendix B).

Table 4. Six Speakers of Dialogues

Speaker (CEFR)	Gender	Nationality	Conversation topic	Duration (min.)
CAE 1 (C1)	Female	Mexican	Weather conditions in the world	3:24
CAE 2 (C1)	Female	Swiss		
FCE 1 (B2)	Female	Italian	Jobs in the Olympic Games	3:41
FCE 2 (B2)	Female	French		
PET 1 (B1)	Male	Saudi	Place they live and furniture	3:38
PET 2 (B1)	Male	Korean		

Data Collection

Rater judgments on 13 speakers' oral performances were collected. First, raters watched the video recording of the test-takers' performances. The recordings were presented to each rater in a randomized order. Next, they indicated their intuitive judgment of each speaker's communicative ability on a seven-point semantic differential scale ranging from poor (=1) to excellent (=7); the mid-points were unspecified. A pilot study showed that the trial participants were able to apply the seven-point scale with ease. Furthermore, the mean single rater-rest of the raters (SR/ROR) correlations was .85, which is considered high and indicates that the raters generally agreed on the ranking of the speakers' performances (Myford & Wolfe, 2004). The written prompt given to the raters was 'While watching each performance, please judge the speaker's COMMUNICATION ABILITY.' Because some raters' experience of formal speaking assessment and contextual factors in the research could have led them to assume that the speakers' English ability, rather than all features pertinent to communicative abilities, was the central concern of this study, the following verbal explanation was given: 'You can look at any aspects of the performances as you like. It doesn't have to be related to their English. It's all up to you what you pay attention to.'

Immediately after rating, the raters provided metacognitive verbal protocols (Ericsson & Simon, 1993), which required them to justify their ratings. The researcher sat behind the raters without interrupting their monologue so they could freely express their impressions. The elicited data (hereafter summary statements) were expected to show the raters' immediate overall reaction to each performance.

Finally, immediately after justifying the rating, the raters reviewed the same performance through stimulated recall (Gass & Mackey, 2017). They were asked to stop the video clip whenever they found something that had affected their judgments and to verbalize it. In this stage, raters were expected to elaborate on the reasons for their ratings by providing detailed descriptions of performance features that affected their judgments. The researcher sat behind the raters and did not intervene while they were engaged in the recall task. The summary statements and stimulated recall data were audio-recorded.

Data Analysis

To answer the first RQ, 'Do general-purpose speaking test results align with linguistic laypersons' intuitive judgments of L2 communicative ability?', the ratings from the 23 raters were statistically analyzed. A many-facet Rasch analysis was conducted using FACETS 3.68.1 (Linacre, 2011). Two facets—raters and speakers—were included in the analysis, and the raters were centered. This study focused on the rank order of speakers' estimated ability, which was measured to examine the correlation between the raters' judgments and the speakers' exam results.

To answer the second RQ, 'What non-criterion features affect linguistic laypersons' judgments of L2 communicative ability?', the summary statements and stimulated recall data were analyzed as follows. First, the data were transcribed and segmented using the C-unit (Loban, 1976). A C-unit consists of an independent clause with its modifiers or any sentence fragments occurring due to hesitation and false starts. This analysis was employed to ensure a highly consistent segmentation process such that the data interpretation was independent from the researcher. Second, the researcher examined all the segments and took notes on the emerging themes, which were used to develop coding categories. A single set of coding categories was developed for the monologic and dialogic tasks since a majority of the themes were identical. Third, each segment was coded into the initially developed categories. The majority of the segments consisted of one theme, and were assigned a single code; however, multiple codes were assigned if a single segment contained more than one theme. During this stage, the initial coding categories were modified by merging several categories or adding new categories. After all the segments were sorted into categories, a second coder coded 10% of the data, and the inter-coder agreement was confirmed (κ=.75). The second coder and the researcher discussed segment coding whenever a disagreement occurred, and the final categories were established through these discussions.

Results

Test Results and Rater Judgments

Table 5 shows the speaker measurement report, which includes the ability measure of each speaker and fit statistics.

Table 5. Speaker Measurement Report

Speaker	Measure	Error	Infit MnSq	Outfit MnSq	Infit *t*	Outfit *t*
CAE 1	2.03	0.19	0.94	0.94	-0.1	0.0
CET A^{+}	1.63	0.19	0.77	0.76	-0.6	-0.6
PET 1	1.17	0.17	1.42	1.42	1.2	1.1
CET A	1.12	0.17	0.68	0.75	-0.9	-0.6
CET B^{+}	1.00	0.17	0.64	0.61	-1.1	-1.1
PET 2	0.98	0.16	1.06	1.27	0.2	0.8
FCE 1	0.54	0.15	0.92	0.87	-0.1	-0.3
CAE 2	0.25	0.14	1.02	0.91	0.1	-0.1
FCE 2	0.19	0.14	0.96	1.09	0.0	0.3
CET B	-0.27	0.14	0.81	0.78	-0.5	-0.6
CET C^{+}	-0.35	0.14	1.26	1.42	0.9	1.2
CET C	-1.23	0.16	0.88	0.97	-0.2	0.0
CET D	-2.22	0.19	0.52	0.50	-1.8	-1.6
Mean	0.37	0.16	0.92	0.95	-0.2	-0.1
SD	1.13	0.02	0.24	0.28	0.8	0.8

Table 5 is arranged by the speakers' ability from the most highly rated speaker to the least highly rated speaker. The order of perceived communicative ability of CET-SET test-takers aligned with their test results. However, some speakers' abilities were not clearly distinguished by the raters. The differences in the estimated measure between CET A and CET B^{+} and between CET B and CET C^{+} were marginal at 0.12 and 0.08 logits, respectively. Furthermore, the rater judgments of the Cambridge Examinations test-takers did not necessarily align with their test results (note that CAE is the most advanced test and PET is the least advanced). Among the six speakers, CAE 1 received the highest ratings from the raters (2.03 logits). However, her interlocutor CAE 2 received ratings lower than FCE 1, PET 1, and PET 2. Furthermore, the two PET test-takers were evaluated more highly than the FCE test-takers, who were supposed to be more advanced than PET examinees. Two speakers with different English proficiencies—CAE 2 and FCE 2—were perceived to be almost identical by the raters since with a difference of only 0.06 logits.

Features Affecting Rater Judgments

Seven categories of performance features were explored in the summary statements and stimulated recall data: (1) English language features, (2) Overall communicative success, (3) Content, (4) Interaction, (5) Non-verbal behaviour, (6) Composure/Attitude, and (7) Other (see Appendix C).

Table 6 shows how many CET-SET and Cambridge Examinations segments were categorized within the main categories. The raters made the largest number of references to speakers' English features, which accounted for 36.7% and 31.2% of all the comments on the two tests, respectively. This finding indicates that the same criteria assessed by the general-purpose speaking tests, including the speakers' fluency, pronunciation, vocabulary, and grammar, influenced the raters' judgments. The qualitative data indicate that the raters' judgments were affected by non-criterion features in addition to language features.

Table 6. Frequency of Segments on Each Main Category

	CET-SET		Cambridge Exams	
Main category	Frequency	%	Frequency	%
1. English language features	1,375	36.7	773	31.2
2. Overall communicative success	746	19.9	438	17.8
3. Content	565	15.1	339	13.7
4. Interaction	28	0.7	303	12.2
5. Non-verbal behaviour	351	9.4	238	9.6
6. Composure/Attitude	213	5.7	157	6.3
7. Other	465	12.4	228	9.2
Total	3,743	100.0	2,476	100.0

Non-criterion Features: Monologues (CET-SET)

Overall Communicative Success

The raters frequently commented on whether the speakers successfully conveyed their message or whether their message was comprehensible. This was categorized as overall communicative success because conveying the speakers' intended thoughts was considered the main goal of informative speeches (Beebe & Beebe, 2009). Message conveyance is closely related to the speaker's language knowledge and fluency since these factors are indispensable in producing comprehensible messages (Hulstijn, 2015). Thus, a large part of this category may be measured by the tests' usual assessment criteria. In fact, many of the raters' comments in this category were accompanied by evaluations of the output's linguistic features. Excerpt 1 illustrates how rater's comprehension can be impeded by the speakers' linguistic problems.

Throughout this paper, the raters' summary statements and stimulated recall data are presented as they were recorded.

1. I found it quite difficult to understand what he was saying because things were rolling off his tongue a lot differently (Rater L, CET D, Summary statement, Categories 1 and 2)

The number of positive comments on this category correlated with the CET-SET test results. However, the raters' judgments of overall communicative success were not completely dependent on the speakers' linguistic quality. Although raters noticed lexicogrammatical and phonological problems, the speakers' communication was regarded as successful when the intended message (information about air pollution) was clearly expressed and transmitted. Excerpt 2 shows that language errors did not necessarily lower raters' judgments of overall communicative success.

2. Even though his sentences like the grammar isn't perfectly constructed or anything, he's still conveying really clearly what he's trying to say by just being confident, speaking loudly, engaging through eye contact and hand gestures. (Rater Q, CET A, Stimulated recall, Categories 1, 2, 5, and 6)

Content

The raters consistently referred to the content of speakers' speech, including the quality of their ideas, elaboration on the ideas, relevance to the topic, and the organization of the speech. The number of positive comments on this category correlated with the test results; CET A^+, CET A, and CET B^+ received numerous positive comments on the content of their speech. The raters mentioned that their arguments were strong and convincing. For example, CET B^+ discussed the seriousness of air pollution by contrasting the skies in different locations worldwide. The real-world examples drew the raters' attention, and her speech was deemed descriptive (see Excerpt 3).

3. I found myself interested in what her thoughts were on the consequences of air pollution aside from just trying to give some feedback on her English language skills. (Rater W, CET B^+, Summary statement, Category 3)

The raters also focused holistically on the schematic organization of the speech. Speakers who included introductions and conclusions were rated higher in the CET-SET presentations. CET B^+'s summary at the end of her speech was judged positively because it helped raters remember the content and helped organize her speech (see Excerpt 4).

4. She also rounds off nicely and ends up her speech, which gives the whole thing some sort of structure, which is good. (Rater C, CET B+, Stimulated recall, Category 3)

In contrast, the raters judged the content of CET B's speech negatively because he repeatedly addressed a single issue. Although he attempted to mention four reasons for air pollution, he did not clearly distinguish them. Moreover, the raters felt that CET B failed to elaborate on the topic sufficiently and frequently moved to different points (see Excerpt 5).

5. Okay "space and blah blah is limited." So what? He looks like he's gonna explain something on that. But then he changes to the life of human beings. (Rater G, CET B, Stimulated recall, Category 3)

Composure/Attitude

The raters also commented on the speakers' confidence, relaxation, and anxiety levels, all of which were categorized under the Composure/Attitude category. Overall, comments on this category correlated with the test results. Some comments indicate that perceived confidence directly affected the raters' impression of L2 communicative ability (see Excerpts 6 and 7).

6. I judge her communication ability very good because at the very beginning she seems confident and she looks totally aware of the whole things that she is going to talk about. (Rater A, CET A+, Summary statement, Categories 1, 3, and 6)

7. She become nervous. And yeah, it caused that her communication ability may seems to not really good. (Rater I, CET C, Summary statement, Category 6)

Non-verbal Behaviour

Raters' comments were not restricted to test-takers' verbal features. The raters seemed to assess confidence and anxiety levels based on non-verbal behaviour, including gestures, eye contact, posture, and facial expressions. They positively perceived gestures that facilitated message conveyance and negatively evaluated body movements that were not connected to the speech's content. In Excerpt 8, Rater F suggested that the speaker could use gestures to indicate a change in her argument.

8. She could have just put her ((raises his index and middle fingers)) "All right. Second argument." using her hands and her arms. (Rater F, CET B+, Stimulated recall, Category 5)

The speakers' eye contact also affected the raters' judgments. The raters frequently commented on whether the speakers were looking at the person or audience they were addressing. Furthermore, eye contact was generally considered to indicate confidence and comfort, whereas its absence was thought to signal nervousness or unfamiliarity with the subject (see Excerpt 9).

9. He gives eye contact when he is putting his point across, which means to me well … or indicates to me that he is obviously confident in the fact that he knows what he is trying to say (Rater L, CET D, Summary statement, Category 5)

Other

This category included a range of miscellaneous features, such as the test-takers' apologies, preparedness, time management, memorization of scripts, recitation of notes, translation, and familiarity with presentations.

Non-criterion Features: Dialogues (Cambridge Examinations)

Overall Communicative Success

The raters frequently mentioned whether Cambridge Examinations test-takers successfully conveyed their message. This was categorized as *Overall Communicative Success* because the goal of the paired interactions was to convey intentions. As Rickheit, Strohner, and Vorwerg (2010) state, in terms of the purpose of message production, 'speakers (as well as writers and signers) produce language in order to convey certain ideas to their interlocutors' (p. 27). In fact, the raters' evaluation of this category did not necessarily correlate with the speakers' language proficiency level, since some lower proficiency speakers received a positive evaluation of their communicative success, as shown in Excerpts 10 and 11.

10. It's easy to understand what he's trying to say even if there are mistakes. (Rater D, PET 1, Stimulated recall, Categories 1 and 2)

11. I think this, it can be effective conversation because both of them can express their thoughts (Rater H, PET 1 and PET 2, Summary statement, Category 2)

Content

The raters addressed the content of the interactions, including elaboration on the ideas, descriptiveness of the message, coherence of the argument, and sufficiency of the discussion. The evaluation of this category did not necessarily align with the

speakers' language proficiency. PET 1 received many positive comments on his content, whereas the other Cambridge Examinations test-takers received larger numbers of negative comments. In particular, the raters frequently commented positively on PET 1's detailed description of his house and how he elaborated on his thoughts, although his speech was not necessarily considered coherent and concise. Excerpt 12 shows a positive comment on his content.

> 12. In terms of contents, PET 1 conveys more clear, a clearer and more detail depiction of his hou… his home and the things he talked about (Rater J, PET 1, Summary statement, Category 3)

In contrast, the speech content of CAE 2, FCE 1, FCE 2, and PET 2 was often judged negatively since they failed to justify their ideas sufficiently. For example, CAE 2 frequently reiterated her partner's points and failed to contribute new ideas to the discussion. Likewise, the FCE examinees failed to elaborate on their opinions with justifications, though they were expected to discuss the advantages and disadvantages of several jobs in the Olympic Games and to state their preference. Excerpt 13 presents comments on the content of CAE 2's speech.

> 13. The poor girl CAE 2, she can't, she seems, she's trying to, she's touching on the surface still. She's not going deeper. (Rater M, CAE 2, Stimulated recall, Category 3)

Composure/Attitude

The raters mentioned the speakers' confidence, relaxation, anxiety, and willingness to communicate. They made frequent positive comments on this factor for the PET test-takers, who were perceived to be confident and relaxed, affecting the raters' overall impressions. The perceived confidence seemed to facilitate message conveyance and compensate for linguistic limitations. Excerpt 14 shows positive feedback on PET test-takers' composure.

> 14. Just in this part I can find so many places with wrong quite easy. But he speaks it quite confidently as well as … It's easy to understand what he's trying to say even if there are mistakes. (Rater D, PET 1, Stimulated recall, Categories 1, 2, and 6)

However, some of the more proficient speakers were perceived to be anxious during their performances and less enthusiastic about communicating (see Excerpts 15 and 16).

15. Anxiety is demonstrated by the fact that she is scratching her nails and fingers and has a much lower voice than her friend CAE 1. (Rater F, CAE 2, Summary statement, Category 6)

16. Overall I think again in their pair conversation, none of them have the passion to communicate with others. Maybe they are not confident enough. (Rater H, FCE 1 and FCE 2, Stimulated recall, Category 6)

Non-verbal Behaviour

Similar to the previous category, raters had positive perceptions of PET test-takers' non-verbal behaviour, but negatively evaluated CAE 2, FCE 1, and FCE 2. The PET test-takers' non-verbal behaviour not only contributed to their perceived confidence, but also enhanced their comprehensibility. They used hand movements effectively to reinforce their intended meaning, compensating for their relative lack of linguistic proficiency. Excerpt 17, a rater's comment on PET 1 when he discussed his family's unity, demonstrates the importance of gestures.

17. So the fact that he was using his hands to gesture "the family being together" was quite good. It strengthened his message there. (Rater V, PET 1, Summary statement, Category 5)

The raters, however, commented negatively on higher proficiency test-takers when their body movements were irrelevant to their message; this included playing with rings, scratching nails and fingers, rubbing legs, maintaining tense shoulders, and avoiding eye contact. Excerpt 18 suggests that Rater W lowered his rating of CAE 2 because of her non-verbal behaviour.

18. So you can watch right here at 2:55 or so CAE 2's right hand's sort of coming out of her sleeves and she sort of ring her fingers. So this is … I was gonna mark them [CAE 1 and CAE 2] little bit closer and I … this nervous energy I got from her kind of maybe bring her down a little bit more. (Rater W, CAE 2, Stimulated recall, Categories 1, 5, and 6)

Co-constructed Interaction

The Cambridge Examinations' assessment criteria include a section on *Interactive Communication*, which measures the candidate's ability to actively participate in developing the discourse. The raters mentioned the interaction as a whole even though they were asked to evaluate each speaker's communicative ability. The raters positively perceived the PET test-takers' interaction since they deepened the discussion by responding to their partners' utterances. There were multiple short

turns where the examinees commented on previous turns or used conversation features (e.g., latched utterances or overlapping talk). Some raters reacted positively to a humorous exchange between the PET test-takers discussing the benefits of having less furniture (see Excerpt 19).

> 19. There are interactive communication in between the two of them. And also it seems that the two of them can answer to each other about each comment, so it make us as the audience can follow the conversation easily. (Rater I, PET 1 and PET 2, Stimulated recall, Category 4)

In contrast, the FCE test-takers' interaction was judged negatively due to their lack of responses to each other's utterances. They often concentrated on expressing their own thoughts by taking turns rather than deeply discussing a given theme. Accordingly, their performance appeared to be two monologues (see Excerpt 20).

> 20. One person is talking about something, the other person interrupts that person and adds something or maybe give information about another topic. It is less … than a successful two-side conversation. These people maybe they lack interaction between each other as well. (Rater A, FCE 1 and FCE 2, Summary statement, Category 4)

Other

This category contained a range of miscellaneous features, speech features, and speakers' behaviours. For example, while watching the videos, the raters guessed where the speakers were from, based on their performances and appearance.

Discussion

The first RQ examined whether general speaking test results align with linguistic laypersons' intuitive judgments of L2 communicative ability. Overall, the findings indicate that linguistic laypersons' ratings are not completely different from exam results, but there is possibility that test results may not reflect linguistic laypersons' judgments of L2 communicative ability.

With regard to the CET-SET performances, the raters' intuitive judgments and test results aligned in terms of the rank orders. A possible reason is that the linguistic features measured in the test affected the raters' judgments of communicative ability. For the Cambridge Examination test-takers, the raters' judgments did not always correlate with the test results; the communicative abilities of the PET test-takers were judged more highly than those of the CAE test-takers and the FCE test-takers. This

finding suggests that the non-language specialists were tolerant of linguistic errors or did not consider linguistic aspects of performance to be as important as the tests consider them to be (Brown, 1995; Elder, 1993; Hadden, 1991). Alternatively, features not captured by the tests' linguistically-oriented criteria might influence rater judgments, as in the case in some previous studies (Barnwell, 1989; Galloway, 1980; Lumley, 1998).

It must be acknowledged that the raters in this study judged test-takers' performances on a single section of the test, whereas the tests' official raters assessed overall performance on the whole test. Moreover, the difference in elicitation tasks in the Cambridge Examinations was considered to influence their ratings. These are alternative possible reasons for the dissonance between the ratings and the official test results.

The second RQ focused on the non-criterion features that affected linguistic laypersons' judgments of L2 communicative ability. The analysis of summary statements and stimulated recall data has demonstrated that the raters' judgments were affected by factors that were assessed by the tests and those that were not. The raters commented on linguistic resources, pronunciation, and fluency, which suggests that the linguistic quality of the test-takers' performances influenced the raters' judgments of L2 communicative ability. Importantly, however, linguistic features and fluency were not the sole determiners of the raters' judgments. This finding aligns with the literature on how non-language specialists evaluate L2 communicative ability (see Table 1).

The emerging non-criterion features were mostly the same for the monologic and dialogic performances. However, the major difference between the two types of performance was observed in the number of raters' comments on interactional features. In particular, the raters attended to co-constructed interaction itself (Jacoby & Ochs, 1995), and not only each individual's communicative ability, when they judged dialogic performances (discussed below). This is a possible reason why the rater judgments of the Cambridge Examinations test-takers were not necessarily consistent with the test-takers' English proficiency.

The raters were concerned with overall communicative success in terms of how successfully the speaker conveyed message and how comprehensible the speaker's message was, which suggests that linguistic laypersons did not focus solely on specific performance features. A lack of linguistic and phonological accuracy was found to affect speakers' comprehensibility. However, the raters positively evaluated overall performance when speakers found means to express their opinions, regardless of

evident linguistic and phonological errors, in particular in the dialogic performances on the Cambridge Examinations. These errors might therefore be a minor factor in raters' judgments of communicative ability unless they impeded comprehensibility. Alternatively, it is possible that linguistic laypersons in this study were tolerant of errors in lexicogrammar and pronunciation (Brown, 1995) or simply did not pinpoint pronunciation errors.

A major emerging feature was the quality of test-takers' speech content or how convincing, elaborated, coherent, and organized the speaker's presentation was. Similarly, raters attended to speakers' elaboration on the ideas, descriptiveness of the message, coherence of the argument, and sufficiency of the discussion in the dialogues. This result corroborates Galloway's (1980) finding that linguistic lay raters mainly focused on the content of L2 speakers' impromptu speech. In addition, this result is similar to some non-language specialists' focus when they observe peers' presentations, such that the content (e.g., significance of the theme) affected non-language specialists' evaluation of performance (Abdul Raof, 2011; Jacoby, 1998). It may be that, as Hughes and Reed (2017) argue, 'In naturally occurring, spontaneous speech, interlocutors do not focus on the mechanics of their interaction but on the ideas/emotions/information being conveyed' (p. 93).

The results also showed that test-takers' non-verbal behaviour, including body movement, eye contact, posture, and facial expression, influenced linguistic laypersons' judgments of their communicative ability. These non-verbal features were addressed in both the monologues and dialogues, although eye contact was perceived to be more important in the former than the latter. Linguistic laypersons' judgments thus depended not only on test-takers' oral performance but also on visual information, as Douglas and Myers (2000) indicated. Particularly, non-verbal behaviours that were positively perceived included gestures that reinforced verbal messages and facilitated message conveyance. The type of gesture perceived as helpful in the performance was *iconic gestures*, which 'present images of concrete entities and/or actions' (McNeill, 2005, p. 39). This finding aligns with that of Gullberg (1998), who discovered that gestures used when discussing concrete objects or actions played an important role in raters' judgments of L2 communicative performances.

Raters also considered speakers' composure and attitude when judging communicative ability in both monologic and dialogic performances. In fact, relaxation and composure have been regarded as components of interpersonal competence (Morreale, Spitzberg, & Barge, 2013). The role of confidence in rater judgment was explored by Zhang and Elder (2011); they found that perceived confidence (or lack thereof) influenced language teachers' evaluation of CET-SET test-

takers. Composure was judged mostly through the speakers' non-verbal behaviour. This observation is largely supported by interpersonal competence models (Morreale et al., 2013), which regard non-verbal behaviour (e.g., body movement, posture, gaze) as skills that reflect relaxation and confidence.

Finally, in the dialogic performances on the Cambridge Examinations, the raters addressed interactions as a whole rather than focusing on individual test-takers' behaviour. More specifically, the rater comments were related to the speakers' interactional pattern or how collaboratively both speakers co-constructed their discourse. The interaction that was positively perceived (the PET test-takers' interaction) was *collaborative*, meaning that the speakers engaged with each other's ideas by responding to their partner's utterances and by introducing new ideas (Galaczi, 2008). In contrast, *parallel* interaction was negatively perceived (the FCE test-takers' interaction); here, each speaker initiated new topics but minimally engaged with each other's ideas. This finding supports the work of Galaczi (2008), who found that interaction co-construction or observed interaction patterns impact raters' judgments of interactive effectiveness.

Conclusion

This study revealed that weak performance tests (McNamara, 1996) may not necessarily align with linguistic laypersons' judgments of L2 communicative ability. It also explored features that affected laypersons' judgments despite being considered construct irrelevant by many general-purpose speaking tests. General-purpose proficiency tests are used for various educational purposes, one of which is to inform stakeholders about the test-takers' language knowledge. If the test's purpose is restricted to assessing this aspect of communicative competence, non-linguistic features are construct-irrelevant variances. However, if speaking tests aim to measure overall communicative ability as valued in real-world situations, developing assessment criteria based on linguistic laypersons' perspectives will lead to a more valid and authentic evaluation of real-life communicative skills. The non-criterion features explored by this study are possible additional criteria that could be incorporated into current linguistically-oriented assessment criteria for general-purpose speaking tests. It is worth noting that LSP tests' assessment criteria already contain test-takers' rapport, confidence, non-verbal behaviour, and eye contact (Douglas, 2000; Fulcher et al., 2011; Jungheim, 2001).

However, incorporating linguistic laypersons' unique perspectives, in particular the non-linguistic criteria, into general-purpose speaking tests poses a practical challenge

for language testing. General-purpose speaking tests have employed linguistically-oriented criteria because language knowledge is easier to describe and has been emphasized in various communicative competency theories (Luoma, 2004; McNamara, 1996). In contrast, Hymes's (1972) *ability for use* (which includes volitional and cognitive factors) is more difficult to grasp and is rarely discussed in applied linguistics (McNamara, 1996). Consequently, the role of non-linguistic features in communication has not been clearly understood or explicated. Without a clearer understanding of non-linguistic features, assessment of the features is more subjective than assessment of linguistic features. Moreover, there is the issue of the scores' generalizability. Since general-purpose speaking tests need to generalize the assessment of performance samples to a broad range of unspecified contexts, the assessment criteria must be relevant to a wide range of TLU domains. While language knowledge and fluency are almost always required for successful verbal communication (Hulstijn, 2015), some non-linguistic criteria derived from this study are context-embedded and not necessarily relevant to certain local contexts (e.g., composure, non-verbal behaviour). Further research should address the fairness and practical issues of assessing non-criterion features in general-purpose speaking tests.

This study has several limitations. First, it investigated only a small number of highly educated linguistic laypersons' perspectives on communication, which might limit the generalizability of the findings. Furthermore, the influence of their familiarity with the test-takers' accents was not examined. Second, this study did not examine how strongly the emerging factors affected linguistic laypersons' overall judgments. Although the frequency of comments was considered in this study, commenting on a certain feature does not necessarily indicate its importance (Barkaoui, 2011). Third, this study only used two types of elicitation tasks. Since influential features differ according to the type of communicative task (Chalhoub-Deville, 1995), other factors may have emerged in rater judgments had other tasks been used. Fourth, this study is further limited by the data collection procedure (i.e., retrospective verbal protocol and stimulated recall). Although this method is regarded as one of the most viable methods to probe raters' cognition (Lumley, 2005), raters' comments might not accurately reflect what they actually paid attention to when watching the videos (Gass & Mackey, 2017). Additionally, the raters were not interlocutors and only evaluated communicative ability as observers. It is recommended that further studies consider these limitations when investigating linguistic laypersons' judgments of L2 communicative ability.

Acknowledgements

This article is derived from my PhD thesis, and I would like to thank the thesis supervisors, Tim McNamara and Catherine Elder, for their support throughout the research project. This article reports on research using examination data provided by the National College English Committee and Cambridge English Language Assessment. This research was supported by the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment from Educational Testing Service. My thanks also go to the editors and the anonymous reviewers of PLTA for their insightful comments and suggestions.

References

Abdul Raof, A. H. (2011). An alternative approach to rating scale development. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 151-163). Hampshire: Palgrave Macmillan.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Beebe, S. A., & Beebe, S. J. (2009). *Public speaking: An audience-centered approach* (7th ed.). Boston: Pearson Education.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91-108.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO Regional Language Centre.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

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

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning, 45*(2), 251-281.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge: University of Cambridge Local Examinations Syndicate.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*(3), 235-254.

Elder, C., McNamara, T. F., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialists can tell us. *Language and Communication, 57*, 14-21.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English Examination. *Language Assessment Quarterly, 5*(2), 89- 119.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The Modern Language Journal, 64*(4), 428-433.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York: Routledge.

Gass, S. M., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-89.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Oxon: Routledge.

Gullberg, M. (1998). *Gestures as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund, Sweden: Lund University Press.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41*(1), 1-24.

Harding, L. (2011). *Accent and listening assessment*. Frankfurt am Main: Peter Lang.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186-197.

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters'accent familiarity on speaking assessment. *Language Assessment Quarterly, 13*(1), 25-41.

Hughes, R., & Reed, B. S. (2017). *Teaching and researching speaking* (3rd ed.). New York: Routledge.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins Publishing Company.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Middlesex: Penguin Books.

Jacoby, S. (1998). *Science as performance: Socializing scientific discourse through the conference talk rehearsal*. Unpublished PhD thesis, University of California Los Angeles, Los Angeles.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171-183.

Jenkins, J. (2015). *Global Englishes* (3rd ed.). Oxon: Routledge.

Jungheim, N. O. (2001). The unspoken element of communicative competence: Evaluating language learners' nonverbal behavior. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 1-34). Honolulu: University of Hawaii Second Language Teaching and Curriculum Center.

Kachru, B. B. (1988). The sacred cows of English. *English Today, 16*(1), 3-8.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants'oral performance. *Language Assessment Quarterly, 9*(3), 249-269.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing, 32*(2), 129-149.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366-372.

Linacre, J. M. (2011). Facets Rasch measurement computer program (Version 3.68.1). Chicago: Winsteps.com.

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347-367.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. Essex: Pearson Education.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Morreale, S. P., Spitzberg, B. H., & Barge, J. K. (2013). *Communication: Motivation, knowledge, skills* (3rd ed.). New York: Peter Lang.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460-517). Maple Grove, MN: JAM Press.

Pearson Education. (2012). *PTE General Score Guide*. London: Pearson Education.

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33*(2), 175-193.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Rickheit, G., Strohner, H., & Vorwerg, C. (2010). The concept of communicative competence. In G. Rickheit & H. Strohner (Eds.), *Handbook of communication competence* (pp. 15-62). Berlin: De Gruyter Mouton.

Roever, C., & Pan, Y.-C. (2008). GEPT: General English Proficiency Test. *Language Testing, 25*(3), 403-418.

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Trinity College London. (2009). *Graded Examinations in Spoken English (GESE) syllabus from 1 February 2010*. London: Trinity College London.

University of Cambridge Local Examinations Syndicate. (2007). *First Cerficiate in English handbook for teachers*. Cambridge: University of Cambridge ESOL Examinations.

University of Cambridge Local Examinations Syndicate. (2009). *Preliminary English Test handbook for teachers*. Cambridge: University of Cambridge ESOL Examinations.

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly, 12*(3), 283-304.

Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly, 8*(2), 200-210.

Zhang, Y., & Elder, C. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET-SET solution. *Language Assessment Quarterly, 6*(4), 298-314.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Appendix A

Raters' Familiarity with Foreign Accent and Language Learning Experience

Rater	Familiarity	Familiar Accent(s)	Language Learning Experience
A	3	Iranian, Turkish	Arabic
B	3	Vietnamese, Singaporean, Chinese, Japanese, Swiss	N/A
C	4	Italian, German, French	French, Russian
D	4	Nepalese, Chinese, Indian	Mandarin, Hindi
E	4	Vietnamese, Swiss, Singaporean	Chinese, Spanish
F	4	French, German	German
G	5	Spanish, Most Latin American, Most EU, Indian, Chinese	French
H	4	Korean, Bangladeshi, Vietnamese, Japanese	N/A
I	4	Singaporean, Filipino, Vietnamese, Japanese, Chinese, Thai	Japanese
J	3	Korean	N/A
K	4	Filipino, Chinese, Korean, Singaporean, Malaysian, Indonesian, Thai, Mexican, Colombian, Brazilian, French	Spanish
L	3	Indian, Mexican	French
M	1	Singaporean, Malaysian, Filipino	N/A
N	5	Nigerian, Indian	French, Hindi
O	4	Malaysian, Singaporean	Cantonese, Malay
P	4	Indian, American, Chinese	N/A
Q	3	N/A	German, Japanese
R	4	Chinese, Taiwanese	German, French
S	4	Ethiopian	French
T	4	German, Austrian	German
U	3	N/A	N/A
V	5	Chinese, Vietnamese, Malaysian, Korean, Singaporean, German	French
W	4	Malaysian	Spanish

Note. The 'Familiarity' column presents raters' self-report on the degree of their familiarity with foreign accents in general (1=not familiar; 5=very familiar).

Appendix B

The Speakers' Test Scores

Speaker	GR	LR	GV	DM	P	IC	GA
CAE 1	3.5	3.5	N/A	3.5	3.5	4.0	4.0
CAE 2	3.0	3.0	N/A	3.5	3.0	3.5	4.0
FCE 1	N/A	N/A	3.5	3.0	3.5	3.5	3.5
FCE 2	N/A	N/A	3.0	3.0	3.5	3.0	3.0
PET 1	N/A	N/A	3.5	4.0	4.0	4.0	4.0
PET 2	N/A	N/A	3.5	4.0	3.5	4.0	4.0

Note. GR=grammatical resource; LR=lexical resource; GV= grammar and vocabulary; DM=discourse management; P=pronunciation; IC=interactive communication; GA=global achievement. The score range is from 0 to 5.

Appendix C

The Seven Main Categories and Subcategories

Main category	Subcategories
1. English language features	Overall English performance; Fluency; Pronunciation; Linguistic resources
2. Overall communicative success	Overall performance and global ability; Overall message conveyance; Overall comprehensibility of message
3. Content	Ideas; Framing of ideas; Topical knowledge
4. Interaction	Interaction and engagement; Interactional pattern
5. Non-verbal behaviour	Body movement; Eye contact; Posture; Facial expression
6. Composure/Attitude	Confidence; Relaxation; Anxiety; Attitudes; Willingness to communicate
7. Other	Miscellaneous features; Comments unrelated to speaker behaviours; Rater's general belief