LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 16, No: 2, July – December 2023

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Effects of Raters' L1, Assessment Experience, and Teaching Experience on their Assessment of L2 English Speech: A Study Based on the ICNALE Global Rating Archives

Shin'ichiro Ishikawa

iskwshin@gmail.com, IPHE, Kobe University, Japan

APA Citation:
Ishikawa, S., (2023). Effects of raters' L1, assessment experience, and teaching txperience on their assessment of L2 English speech: A study based on the ICNALE Global Rating Archives. *LEARN Journal: Language Education and Acquisition Research Network, 16*(2), 411-428.

Received 18/01/2023

Received in revised form 08/03/2023

Accepted 15/05/2023

ABSTRACT

TESOL practitioners, especially in Asia, tend to believe that reliable assessment of students' L2 English speech can be done solely by L1 English native speakers with sufficient teaching and assessment experiences. Such a belief, however, may need to be reconsidered from a new perspective of "diversity and inclusivity." This study used data from the ICNALE Global Rating Archives, a newly compiled assessment dataset, to examine the degree of the effects of raters' L1, assessment experience, and teaching experience on their assessment of Chinese, Japanese, and Thai learners' L2 English speech. The quantitative analyses showed that (1) raters' L1 significantly influenced the assessment scores for all three learner groups, but the difference between native speaker raters and non-native speaker raters was not clear, (2) raters' assessment experience influenced the scores only for Japanese learners, and (3) raters' teaching experience did not significantly influence the assessment scores for any of the three learner groups. These findings, which cast doubt on the dependence on native speaker raters alone, suggest the need to involve a greater variety of raters in L2 speech assessment.

Keywords: L2 speech assessment, effects of rater background variables, diversity and inclusion in L2

Introduction

Assessing the linguistic quality of learners' L2 English speeches in a consistent and reliable manner can be a highly demanding task. Compared to essay assessment, speech assessment concerns a much greater number of aspects, including pronunciation, intonation, fluency, pragmatics, and interaction, in addition to language (grammar) and content. Hall and Hope (2016) suggest that speeches are essentially "ephemeral" and "[e]ven if a speech sample is recorded, it is often difficult to identify the individual strengths and weaknesses that characterise a test-taker's overall speaking abilities."

Thus, some of the L2 English teachers in Asia, especially those in regions where English is taught as a foreign language (EFL) rather than a second language (ESL), seem to feel hesitant to assess student speeches by themselves in high-stakes testing situations. This attitude may be corroborated by their firm belief that reliable L2 speech assessment can and should only be done by native or near-native English speakers (NENS) with enough experience in teaching the language and assessing student outputs. For example, the Tokyo Metropolitan Board of Education implemented an L2 English-speaking test as part of the capital's public high school entrance exam in 2022, the first in the nation. What should be noted here is that the assessment of student responses was conducted not in Japan but in the Philippines, where local staff with BA degrees graded them in pairs (Honda & Tsuchidate, 2022). Why the speech of Japanese students was evaluated not in Japan but abroad in the Philippines has not been explained by the Board of Education.

However, such total dependence on NENS raters with experience in assessment and teaching and the consequent exclusion of non-native speaker (NNS) raters with relatively fewer experiences may be ungrounded, especially when considering the fact that (i) an increasing number of people in Asia use English as a lingua franca, a tool for communication between those with different L1 backgrounds, (ii) "native-speakerism" (Holliday, 2005), or an ideal ENS teacher model, has been critically reconsidered in recent applied linguistics, and (iii) the principle of "diversity and inclusivity" is becoming mainstream in every field of ELT. Regarding the last point, the American Council on the Teaching of Foreign Languages (ACTFL) released a special comment in 2019: "ACTFL values diversity and strives for inclusion across world language teaching and learning contexts." ACTFL also emphasises that "[n]o individual should experience marginalisation of their contributions or talents because of their unique attributes" such as language identity, ethnicity, national origin, and race as well as age, belief system, disability status, gender (identity/expression), sexual orientation, and socioeconomic status. Undoubtedly, this principle should also be applied to L2 assessment tasks. Chau et al. (2022) mention that there is "increased attention to the need for promoting diversity and inclusion in language education," and they touch upon the possibility that the studies of global Englishes and translanguaging lead to "the acceptance and celebration of different language backgrounds, cultures, beliefs, and values of students" in ELT.

These findings suggest that more NNS teachers should be involved in L2 speech assessment tasks. However, to realise this, we need to confirm that the assessment by NNS raters is sufficiently reliable and does not deviate much from the assessment by experienced NS raters.

Therefore, this study analyses the data from the ICNALE Global Rating Archives (Ishikawa, 2020), a newly compiled large-scale assessment dataset, and discusses the features of the assessments of L2 English speech—utterances in the L2 persuasion roleplays—of college students from China, Japan, and Thailand by various rater subgroups based on L1s, assessment and teaching experience.

Literature Review

Reliability in Assessment

Reliability, or the agreement of assessment results across different assessment conditions, is indispensable for a good assessment. Bachman and Palmer (1996) state that reliability is "an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure" (p. 20). Reliability is a multi-faceted construct. Price et al. (2017) mention that reliability concerns three aspects: consistency over time (test-retest reliability), consistency across items (internal consistency), and consistency across different researchers (inter-rater reliability) (p. 67).

Inter-rater reliability is a particularly crucial factor in rater-mediated assessments. Wind and Peterson (2018) conducted a systematic literature review of 259 language assessment studies and reported that many discussed assessment quality in terms of inter-rater reliability. When a group of raters is consistent with one another, they should "reach consensus in their ratings on the same test takers" and "produce the same or similar ratings for each test taker" (Yan & Fan, 2022).

Inter-rater reliability is usually measured using Cronbach's alpha coefficient, which is usually expected to be higher than .8, meaning that 80% of the total variance is true score variance, whereas only 20% is error variance (Brown, 2022).

Rater Background Effects

To discover what type of rater backgrounds affect L2 speech assessments, previous studies have discussed several aspects of rater bias. In this study, we reviewed three articles. First, Han (2016) summarised that previous studies have focused on rater variables, such as language backgrounds (L1 and L2), assessment expertise, and rater training, to explain why raters differ, although Fan stated these effects remain understudied. Dalman and Kang (2019) suggest that "human raters are vulnerable to be[ing] impacted by listeners' bias" and mentioned factors such as first language status, exposure to different varieties of English, educational background, linguistic knowledge, prior language teaching/tutoring experience, linguistic stereotyping, and attitude, as potential sources for bias. They also emphasised the potential value of rater training. Noh et al. (2021) surveyed 43 representative studies and reported that the most widely discussed speech/essay rater variables were assessment experience (13 studies), L1s (12 studies), familiarity with test-takers' L1s (11 studies), rater training (11 studies), and teaching experience (8 studies).

Although the findings in the previous studies are not necessarily consistent, some of them suggest the possibility that (1) raters with different L1 backgrounds may give more or less importance to each assessment criteria and apply a different range of assessment scales, (2) ENS raters may tend to be better at holistic assessment and may pay more attention to the overall organisation, while NNS raters may be better at analytical assessment, and may pay more attention to grammatical accuracy, (3) when raters are familiar with learners' L1s as their first or second language, they may be more lenient in their assessment, (4) experienced raters may use rubrics more effectively and assess learner outputs more consistently and reliably, (5) rater training may help lessen the variance in assessment severity and consistency, and (6) raters with teaching experience may present better assessment performance. Among these variables, this study focuses on three key factors: L1, assessment experience, and teaching experience.

L1

Many studies have suggested that ENS/NNS raters focus on different aspects of a learner's speech. For example, Zhang and Elder (2014) analysed the assessments given by ENS and Chinese raters to Chinese college students' L2 English speech on the national test and

reported that ENS raters often focused on interaction as a critical factor of communication, whereas NNS raters had a narrower view and relied solely on linguistic accuracy in speech. Saito and Shintani (2016) reported the opposite finding. Analysing the differences between Canadian and Singaporean raters' assessments of L2 English speeches of L1 Japanese learners, they reported that Canadian raters paid attention exclusively to phonological accuracy and fluency of learner speeches and assigned severer scores in comprehensibility, whereas Singaporean raters paid more attention to various lexicogrammatical aspects and assigned more lenient comprehensibility scores. They explained that this is because Canadian raters used "only North American English in a monolingual environment," whereas Singaporean raters were accustomed to plural English models in their multilingual context.

Meanwhile, some studies have suggested similarities in the assessments by ENS/NNS raters. Gui (2012) compared the scores and comments ENS and L1 Chinese raters gave to the speech of Chinese college students in a college speech contest and revealed that both showed a fairly high degree of agreement on the competition winners and the scores, although ENS raters offered varied critical comments, whereas Chinese raters offered more positive comments on the students' pronunciation, usage of English expressions, and speech delivery. Wei and Llosa (2015) also found similarities in the assessment scores of ENS and NNS raters. They analysed how ENS (American) and Indian raters assessed Indian students' oral skills in the TOEFL speaking task and revealed no significant differences in terms of the use of scoring criteria, attitude toward Indian English, internal consistency, and severity in scoring.

In addition, several studies suggest that a common linguistic background between a rater and a test taker may lead to lenient scoring. Winke et al. (2011) analysed data from the TOEFL speaking task and revealed that raters with Spanish as an L2 were lenient toward L1 Spanish test-takers' speech, and raters with Chinese as an L2 were lenient toward Chinese test-takers' speech. In contrast, Lee (2017) compared the assessments of Korean college students' L2 English speech in the TOEFL speaking task with three kinds of student raters: (a) non-native Korean speakers (e.g. international students), (b) Korean/English bilingual speakers (i.e. L1 Korean speakers with equal proficiency in English), and (c) native Korean speakers, who were told to assess the accentedness and comprehensibility of the other students' speech. The analysis showed that the average assessment scores were highest for (a), followed by (b), and the lowest for (c). In this case, limited familiarity with test-takers' L1 seemed to lead to a more lenient assessment.

Prior Experience in Speech Assessment

In addition to L1, prior experience in L2 speech assessment may also influence assessment performance. It is usually expected that prior assessment experience, which includes participation in a rater training program, offers raters an opportunity to become accustomed to L2 English accents, leading to better assessment performance and greater lenience in assessment. Huang et al. (2018) compared the Common European Framework of Reference for Languages (CEFR)-based assessments of speech samples by two college teachers who had attended a rater standardisation training course with the assessment of four other teachers who had not, and revealed that the trained raters agreed in 44% of cases, and the correlation between their scores was very high (ρ= .89). Thus, they concluded that the training helped enhance the raters' overall assessment performance. Xi and Mollaun (2009) compared the assessment performance of Indian raters who received two regular training sessions where they assessed the speech of a variety of test takers with raters who received a single regular session and an additional special session focusing on assessing the speech of Indian test-takers. It was revealed that after attending the first regular session, Indian raters, some of whom had complex feelings about Indian English, improved the accuracy and stability of their assessment of the oral performances of Indian and non-Indian test takers. They also showed that a special training session helped Indian raters assess the oral performance of Indian test-takers more consistently and reliably.

However, the relationship between assessment experience and assessment performance may not be clear. Isaacs and Thomson (2013) examined how differently experienced and novice raters assessed the L2 speech of new foreign residents of Canada using a 5-point and 9-point scale. The results yielded no statistically significant group differences in either rater experience or scale length.

Prior Experience in L2 Teaching

The effects of prior experience in L2 teaching have also been widely discussed. It is usually expected that raters' experience of teaching English to non-native speakers offers them an opportunity to become accustomed to learner accents and to understand the hardship of speaking in L2, which usually leads to lenience in assessment. Analysing undergraduate students' subjective judgements of the oral skills of non-native teaching assistants and the objective prosodic features of their speeches, Kang (2012) showed that students' language teaching experience, as well as their native-speaker status, explained 7–9% of the variance in their judgement, although prosodic variables explained 18–19% of it. It was also suggested that students tended to be more lenient in their assessment if they had previous language-teaching experience. Hsieh (2011) compared the assessment of the oral skills of non-native teaching assistant candidates by American undergraduate students and ESL teachers. The analysis showed that ESL teachers, who were linguistically more sophisticated, tended to adopt an analytical approach in their assessments and assigned more lenient scores than students.

Meanwhile, Kang et al. (2019) did not observe a significant correlation between the quantity of teaching experience and their holistic rating scores as well as rating severity, and reported that 20% of untrained raters' score variance could be explained by their background (especially native speaker status) and attitudinal elements.

As with L1 effects, the effects of raters' assessments and teaching experiences on their assessment performance are ambiguous. Noh and Matore (2022) analysed 164 English teachers' assessment of lower secondary school students' speech based on three criteria: vocabulary, grammar, and communicative competence. The data analysis proved that both assessment and teaching experience were related to severity rather than lenience in the assessments, whereas the experience of attending rater training did not significantly affect assessment performance.

Needs for Further Research

As briefly surveyed, some studies suggest that raters' L1s, assessment and teaching experience affect their L2 speech assessment behaviours. However, the findings of different studies are inconsistent. Thus, whether such effects really exist, and if they do, whether they lead to lenience or severity in the assessments, and whether they lead to improvement or deterioration in the overall assessment performance remains unclear. This may be because many previous studies were based on relatively small assessment datasets.

When discussing the effect of a particular rater variable on assessment, it should be noted that it can be much more unstable than generally expected. It is quite difficult to say whether assessment experience, for example, may lead to severity or lenience in assessment. If a rater has much experience in assessing high-proficiency learners' speech before, they naturally come to be more demanding in the assessment, whereas if they have regularly assessed novice learner speeches, they may become more lenient. This suggests that a finding from the analysis of a smaller dataset —the assessment that a few raters gave to the outputs of a few learners, for instance— is least likely to be generalisable.

Bearing this in mind, Ishikawa (2023) analysed the assessment more than 50 raters gave to the same set of 140 learners' spoken and written outputs and revealed that none of the raters' L1, nationality, sex, assessment experience, or teaching experience significantly influenced their overall assessment scores. Although this seemed to be an interesting finding, the analysis was

only preliminary because it analysed 140 output samples from a variety of learners as a whole set, without considering the internal variance in terms of learners' L1s and their L2 proficiency levels. Due to this limitation, whether raters' assessment performance is truly independent of their background variables is still unclear. In this study, we aim to re-examine this issue.

Methodology

Aim and RQs

As mentioned above, this study aims to reconsider the appropriateness of privileging experienced NENS-teacher raters and exclude other raters by quantitatively examining the possible effects of three types of rater background variables (L1s, assessment experience, and teaching experience) on their L2 speech assessments.

To solve the methodological limitations of Ishikawa (2023), this study focused on the assessment data only for the speech of college students from three Asian EFL regions— China, Japan, and Thailand— at roughly the same L2 proficiency levels. The analysis was conducted on each of these three datasets, and whether a stable trend applicable to all three learner groups can be found was examined.

The research questions (RQs) for this study are shown as follows:

RQ1 To what degree is L2 speech assessment influenced by raters' L1s?
RQ2 To what degree is L2 speech assessment influenced by raters' assessment experience?
RQ3 To what degree is L2 speech assessment influenced by raters' teaching experience?

Data

This study used assessment data taken from the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023). ICNALE consists of four output modules: Spoken Monologues, Spoken Dialogues, Written Essays, and Edited Essays. It is one of the largest learner corpora focusing on Asian learners and includes 4,400 monologues, 4,250 dialogue pieces, 5,600 essays, and 1,300 edited essays. Approximately 4,300 college students, including some graduate students, from ten regions in Asia and 370 native English speakers (including students, teachers, and others) participated in the project. The dataset analysed in this study is part of the ICNALE Global Rating Archives (ICNLAE GRA) (Ishikawa, 2020), which is the latest addition to the ICNALE. It Includes rubric-based assessments of 140 speeches and the same number of essays chosen from two ICNALE modules: Spoken Dialogues (Ishikawa, 2019) and Written Essays (Ishikawa, 2013). This study focused only on speech assessment data. The following subsections introduce what types of speech were assessed, what types of raters were recruited, and how the assessment was conducted.

Speech Samples

We analysed the assessment data for the speech of 60 college students, 20 from each of the three Asian EFL regions: China, Japan, and Thailand.

All participating students were classified into four CEFR-linked proficiency levels (A2, B1 low, B1 upper, and B2+) based on their scores on standard proficiency tests such as TOEFL, TOEIC, and IELTS or on the common vocabulary size test (See Ishikawa, 2013) as shown in Table 1.

Table 1

Proficiency Levels of Three Learner Groups

	A2	B1_1	B1_2	B2+	Sum
Chinese	2	6	6	6	20
Japanese	5	5	5	5	20
Thai	6	6	6	2	20
Total	13	17	17	13	60

Although the numbers of Chinese students at the A2 level and Thai students at the B2+ level were somewhat smaller due to the imbalance in the number of participants in the original corpus, proficiency levels were largely controlled between the three learner groups.

These students joined an oral performance interview, which included a persuasion roleplay based on the topic of a part-time job. In the roleplay, participants were requested to persuade their stubborn college supervisor who believed that students should focus on their studies to allow them to continue working part-time.

Part of the transcribed speech of CHN_001 is shown below, where [S] and [T] stand for student and teacher, respectively.

[S] Okay. Um, I think, um, this part-time job is very useful for me. It's not just as money. Of course, money is very important.
[T] Mm-hmm?
[S] But I think the part-time job improve my, uh, Japanese ability.
[T] I see, but actually, as your teacher, I think that you are really a good student and that if you spend more time, I believe that you can do a very good research, so that's why I want you to stop working, you know, as soon as possible. And then you mentioned the Japanese proficiency. Okay, no problem. Even if you do not work outside, you can develop your Japanese ability. Now, you are talking to me and then now you are talking – you can talk with your friends in the seminars, so even if you do not work outside, you can develop your Japanese ability. Don't mind. It's okay.
[S] Yes, of course, I can improve my Japanese in school but, uh, uh, part-time job is not like school. Uh, it's a society environment.
[T] Mm-hmm.
[S] Means people will not think you are student. They think you as a staff, and in that situation, you can be more, um, stronger – your heart will be more stronger.
…

Although the roleplays usually last for a few minutes, we prepared audio files that included only the initial 90 seconds of utterances, which were anonymised and sent to the raters for assessment.

Raters

The ICNALE project team hired a variety of raters from several regions in Asia, Canada, and the US. All raters were asked to respond to an online background survey, which enquired about raters' age, sex, L1, nationality, countries where they had lived for more than ten years, highest degree (e.g. BA, MA, Ph. D.), major at college (e.g. English, humanities, sciences, etc.), current job (e.g. English teacher, teacher of other subjects, non-teacher, etc.), past jobs, English proficiency (e.g. B2, C1, C2, near-native), past experience of using English for professional purposes, delivering English presentations, writing English reports, joining English discussions,

assessing student essays, and assessing student speech. This study focuses on three variables: L1, speech assessment experience, and teaching experience.

The backgrounds of the 60 speech raters discussed in the current analysis are summarised in Table 2.

Table 2

Rater Backgrounds

Background	Subtypes
L1s	English (2), Filipino (15), Chinese (10), Japanese (9), Thai (5), Others (19)
Assessment experience	never (7), some (17), many (36)
Teaching experience	English teachers (32), Other teachers (7), Non-teachers (21)

In terms of L1s, we classified the 60 raters into five subgroups: NENS or L1 English and Filipino raters (17), L1 Chinese raters (10), L1 Japanese raters (9), L1 Thai raters (5), and other NNS raters (19), including L1 Lao, Indonesian, and Korean speakers, among others. Three NNS rater subgroups (L1 Chinese, Japanese, and Thai raters) shared L1 backgrounds with one of the three learner groups.

In terms of L2 speech assessment experience, we classified raters into three subgroups—never: 0 times (7), some: 1–5 times (17), and many: 6+ times (36)— based on the results of the self-report survey.

In terms of teaching experience, the raters were then classified into three subgroups: English teachers (32), who teach English at high schools, colleges, and language schools; non-English teachers, many of whom teach a variety of subjects (economics, mathematics, etc.) at college; and non-teachers (21), who include businesspeople, graduate students, and people with no regular jobs.

Assessment Policy

To elicit high-quality assessment data from a variety of raters, the project team prepared a detailed assessment guide and asked all raters to assess each of the 140 speech samples using two assessment methods: a holistic assessment (/100) and an analytical assessment (/100). The latter covers ten assessment criteria (/10 for each), as shown in Table 3.

Table 3

Analytical Assessment Criteria

Language	Content	Attitude
Intelligibility	Comprehensibility	Willingness to communicate
Complexity	Logicality	Involvement
Accuracy	Sophistication	
Fluency	Purposefulness	

Notes: See Ishikawa (2023) for a detailed definition of each assessment criterion.

The team also regulated the scoring policy. To prohibit raters from assigning scores that are too high or too low or assigning almost identical scores to all the speech samples, the team asked all the raters to make the average assessment scores fall between 40 and 60% and to make the standard deviation fall between 20 and 30%. These values were automatically calculated on the spreadsheet so that raters could check them quickly. When the assessments did not meet these requirements, raters were asked to adjust their scores.

All raters were requested to carefully read an assessment guide, which explained all of the above in great detail, and then take a check test based on the content of the guide. Those who could not pass the test were requested to reread the guide and take a re-test. Only those who successfully passed the test were allowed to begin their assessment. Therefore, common rater training was offered to all participating raters.

Data Analysis

The analysis is based on the total sum of a holistic assessment score (/100) and the sum of ten kinds of analytical assessment scores (/100) assigned by raters, hereafter referred to as total rating score or TRS (/200). The mean TRS was calculated for each rater subgroup and for each of the three learner groups.

Assessments by different rater subgroups were compared from four viewpoints: (i) trend of change in TRS (whether a steadily increasing or decreasing trend between different rater subgroups can be seen), (ii) significance of the effect of a particular rater-related variable (whether a variable significantly influences the assessment and whether a significant difference can be observed between different rater subgroups), (iii) assessment sensitivity (whether a particular rater subgroup assesses learner speeches with greater precision), and (iv) inter-rater reliability (whether a group of raters assign similar scores to the same set of learner speech samples). This four-stage verification helps to achieve a more careful discussion of the matter. In all cases, we concluded that a meaningful effect caused by a particular rater variable exists only when the difference is observed consistently in the assessments of all three learner groups.

For (i), we first examined the mean TRS values assigned by different rater subgroups. If a target rater variable (L1, assessment experience, and teaching experience) is effective, a stable increasing or decreasing trend between the different rater subgroups in the assessments of all three learner groups should be found.

Regarding (ii), we applied two-way ANOVA tests to confirm whether a target variable had a significant effect on the assessment scores for all three learner groups. When the interaction was not significant, we focused on the main effect of the target variable; when it was significant, we focused on the significance of the simple main effect of the target variable. If the effect was significant, we conducted a post-hoc test (Holm) to determine whether there was a significant difference between the different rater groups.

Regarding (iii), we analysed the coefficients of variation (CV). CV, which is obtained by dividing the standard deviations by the mean values, represents the relative data dispersion. If a rater is sufficiently sensitive and can discern the minutest differences in the quality of the learner outputs, the value is expected to increase.

Finally, regarding (iv), we examined Cronbach's alpha, which is calculated using the formula: $cN/ [v + c (N\text{-}1)]$ (N: the number of items [i.e. raters], c: the average inter-item covariance among the items, v: the average variance). As a measure of internal consistency among raters, it represents the degree to which a set of assessments are related as a group. If a particular rater subgroup assesses the same set of learner speech samples in a reliable (i.e. consistent) manner, it is naturally expected to increase. A value of .70 or higher is usually regarded as "acceptable" in most social science research (UCLA Advanced Research Computing, n.d.).

We examined each RQ from the four perspectives mentioned above. First, for RQ1 (effects of L1), we compared five rater subgroups: L1 English/Filipino (ENG/FIL)speakers (i.e. NENS), L1 Chinese speakers, L1 Japanese speakers, L1 Thai speakers, and other NNS. Next, for RQ2 (effects of assessment experience), we compared three rater subgroups: raters who have never assessed learner speeches before (never), raters who have assessed learner speeches 1-5 times (some), and raters who have done it more than six times (many). Finally, for RQ3 (effects of teaching experience), we compared three rater subgroups: English teachers, other teachers, and non-teachers.

Our initial hypotheses, which are based on the major findings in the literature, are TRS values steadily change between NENS and NNS raters, between those with and without assessment experience, and between those with and without teaching experience, and NENS raters with more assessment and teaching experience present more sensitive and reliable assessment performance than NNS raters with limited assessment and teaching experience.

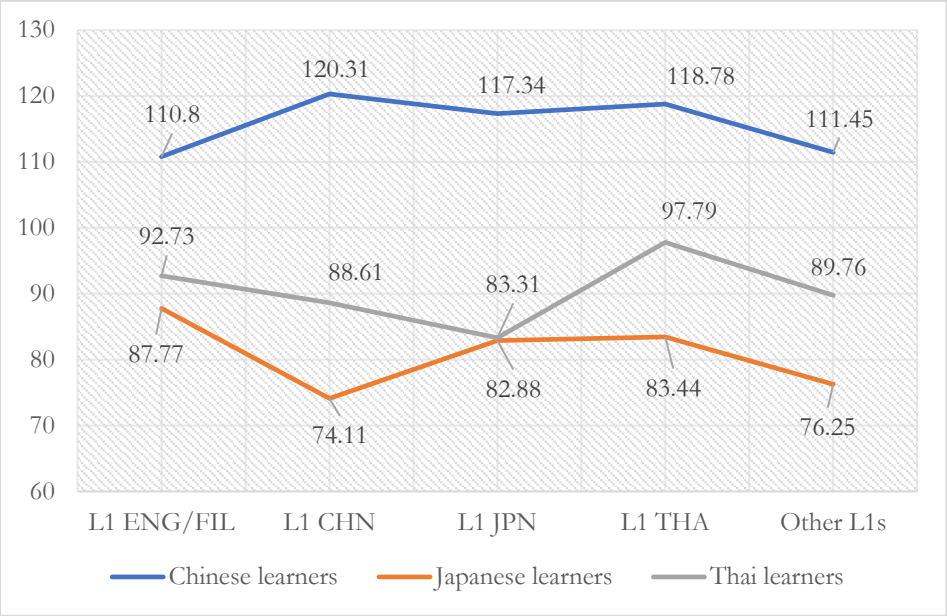
Findings and Discussions

RQ1 Effects of Raters' L1 Background

First, regarding the trend of change, Figure 1 represents the mean TRS values of the five rater subgroups based on L1s assigned to the speech samples of the three learner groups.

Figure 1

Mean TRS Assigned by Five Rater Subgroups Based on L1s



Among the five rater subgroups, L1 ENG/FIL or NENS raters assigned the lowest scores to Chinese learners (110.80 vs 111.45–120.31), the highest scores to Japanese learners (87.77 vs 74.11–83.44), and the second highest scores to Thai learners (92.73 vs 83.31–97.79). It was not shown that NENS raters consistently gave higher or lower scores than NNS raters. Rather, different patterns were observed in each of the three learner groups. Thus, the data did not support the hypothesis that raters' L1, especially the gap between NENS and NNS, was a decisive factor in L2 speech assessment behaviour, and they presented different assessment patterns.

In addition, unlike the suggestions in some studies, it was not proven that raters always assigned higher scores when assessing learners with common L1 backgrounds. L1 Thai raters assigned higher scores to Thai learner speeches (97.79 vs 83.31–92.73) than other raters, but a similar trend was not observed with L1 Chinese and Japanese raters.

Second, regarding the significance of the effect of L1s as a rater variable, the two-way ANOVA tests proved that the interaction between the rater L1s and the learner groups was significant ($F(8, 3585)=4.32$, $p< .001$), and the simple main effects of rater L1s were all significant for learners from China ($F(4, 3585)=3.11$; $p=.015$), Japan ($F(4, 3585)=5.97$; $p< .001$), and Thailand ($F(4, 3585)=2.92$; $p=.020$), meaning that raters' L1s significantly influence the

assessment scores. However, the Tukey test proved no significant differences between NENS and L1 Chinese raters (p= .60), NENS and L1 Japanese raters (p= .68), and NENS and L1 Thai raters (p= .75), although a significant difference was observed between NENS and other NNS raters (p= .04). Thus, statistical tests did not support a clear difference between NENS and NNS raters in the TRS.

Next, the CV and Cronbach's alpha values are shown in Table 4.

Table 4

Assessment Quality for Five Rater Subgroups Based on L1s

Index	Rater Subgroups	Learner Groups		
		Chinese	Japanese	Thai
CV (%)	L1 English/Filipino	30.71	42.40	43.25
	L1 Chinese	26.46	47.09	45.02
	L1 Japanese	31.02	50.12	55.06
	L1 Thai	32.19	41.82	41.78
	Others	32.37	49.35	47.01
Cronbach's alpha	L1 English/Filipino	0.82	0.82	0.92
	L1 Chinese	0.92	0.93	0.96
	L1 Japanese	0.93	0.93	0.95
	L1 Thai	0.66	0.54	0.91
	Others	0.91	0.93	0.96

Third, regarding assessment sensitivity, NENS raters presented middle-level CV values among the five rater subgroups. In the case of the assessment of Chinese learners' speech, for instance, the CV for NENS raters was 30.71, which was between the maximum value (Others: 32.37) and the minimum value (L1 Chinese raters: 26.46). This suggests that NENS raters are not necessarily sensitive to differences in the quality of a learner's speech.

Finally, regarding inter-rater reliability, NENS raters also presented middle-level values among the rater subgroups. For instance, in the assessment of Japanese learner speeches, Cronbach's alpha value for NENS raters was 0.82, just between the maximum value (L1 Chinese/Japanese raters and Others: 0.93) and the minimum value (L1 Thai raters: 0.54). This exemplifies that raters' individual differences did not necessarily have a smaller effect for NENS raters than for NNS raters. Thus, the hypothesis that NENS raters present more sensitive and reliable assessment performance than NNS raters, which many TESOL practitioners have taken for granted, was not clearly supported in our data.

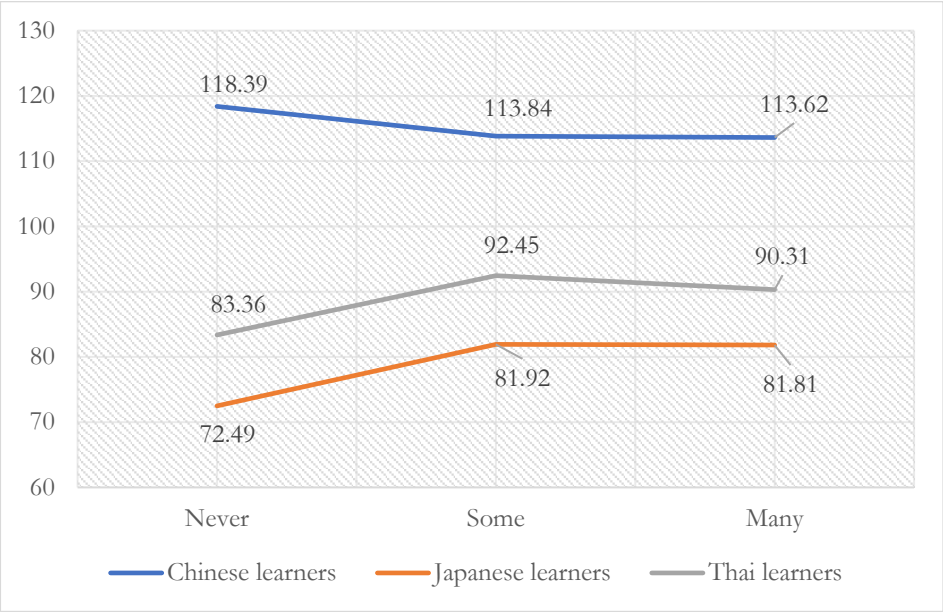
Another notable finding is that the inter-rater reliability was sufficiently high not only for NENS raters but also for most NNS raters. Except for L1 Thai raters, for whom the values were below 0.7 in the assessments of Chinese/Japanese learners, all NNS raters presented a satisfactory level of assessment consistency, which seems to rationalise the inclusion of a greater number of NNS raters in L2 speech assessment.

RQ2 Effects of Raters' Assessment Experience

Regarding the trend of change, Figure 2 represents the mean TRS values of the three rater subgroups based on L2 speech assessment experience assigned to the speech of the three learner groups.

Figure 2

Mean TRS Assigned by Three Rater Subgroups Based on Assessment Experience



Among the three rater subgroups, raters with much assessment experience assigned the lowest scores to Chinese learners (113.62 vs 113.84–118.39) and the middle scores to Japanese (81.81 vs 72.49–81.92) and Thai learners (90.31 vs 83.36–92.45), whereas raters with no assessment experience assigned the highest scores to Chinese learners (118.39), and the lowest scores to Japanese (72.49) and Thai learners (83.36). No clear contrast was observed between the two rater subgroups. Thus, the hypothesis that TRS steadily increases or decreases as raters have more assessment experience was refuted.

Second, regarding the significance of the effect of assessment experience as a rater variable, the two-way ANOVA tests proved that the interaction between the raters' assessment experience and the learner groups was significant ($F(4, 3591)=2.58$, $p= .036$), and the simple main effect of the former was significant for learners from Japan ($F(2, 3591)=3.69$; $p=.025$), but not significant for learners from China ($F(2, 3591)=0.93$; $p=.393$) and Thailand ($F(2, 3591)=2.81$; $p= .060$), meaning that raters' assessment experience does not always influence assessment scores. In addition, the Tukey test proved no significant differences both between "never" and "some" ($p=.080$) and between "some" and "many" ($p=.830$), which refutes the possibility of a consistent change of TRS according to the amount of the assessment experience. Thus, our hypothesis that assessment experience plays an important role in L2 speech assessment, which much of the literature has hinted at before, was not supported. In comparison to the L1 effect, the effect of assessment experience was only slight.

Next, the CV and Cronbach's alpha values are shown in Table 5.

Table 5

Assessment Quality for Three Rater Subgroups Based on Assessment Experience

Index.	Rater Subgroups	Learner Groups		
		Chinese	Japanese	Thai
CV (%)	Never	26.69	49.92	51.14
	Some	31.76	46.90	43.97
	Many	31.19	46.03	46.62

Cronbach's alpha	Never	0.91	0.93	0.96
	Some	0.84	0.88	0.93
	Many	0.95	0.96	0.98

Third, with regard to assessment sensitivity, raters with much assessment experience presented a middle-level CV value for Chinese learners (31.19 vs 26.69–31.76), the lowest value for Japanese learners (46.03 vs 46.90–49.92), and the middle-level value for Thai learners (46.62 vs 43.97–51.14), whereas raters with zero assessment experience presented the lowest value for Chinese learners (26.69) and the highest values for Japanese (49.92) and Thai learners (51.14). This indicates that there were no clear contrasts between these two rater subgroups. Thus, our hypothesis that more assessment experience leads to a greater level of sensitivity in assessment was refuted.

Finally, regarding inter-rater reliability, raters with much assessment experience presented the highest Cronbach's alpha values in the assessments of all three learner groups (Chinese: 0.95, Japanese: 0.96, and Thai: 0.98), suggesting that raters' individual differences may tend to have a smaller effect for experienced raters. However, raters with zero experience did not present the lowest values (Chinese: 0.91, Japanese: 0.93, and Thai: 0.96). The relationship between assessment experience and inter-rater reliability does not seem to be linear, if it exists. Thus, the hypothesis that experienced raters would show a higher level of reliability was supported only partially.

Cronbach's alpha values were higher than .7, even for those who had no or little assessment experience. There seems to be no clear reason to exclude novice raters from assessment tasks. As mentioned before, some teachers tend to avoid assessing a student's L2 speech because they do not have enough assessment experience, but such a preconceived idea may not be grounded.

The data presented a somewhat mixed picture regarding the merits of assessment experience. This may be partially explained by the data collection method adopted in the ICNALE GRA project, in which a detailed assessment guide was prepared, and all raters were asked to take a check test before commencing the assessment task, which might have helped neutralise the possible effect of past assessment experience.

RQ3 Effects of Raters' Teaching Experience

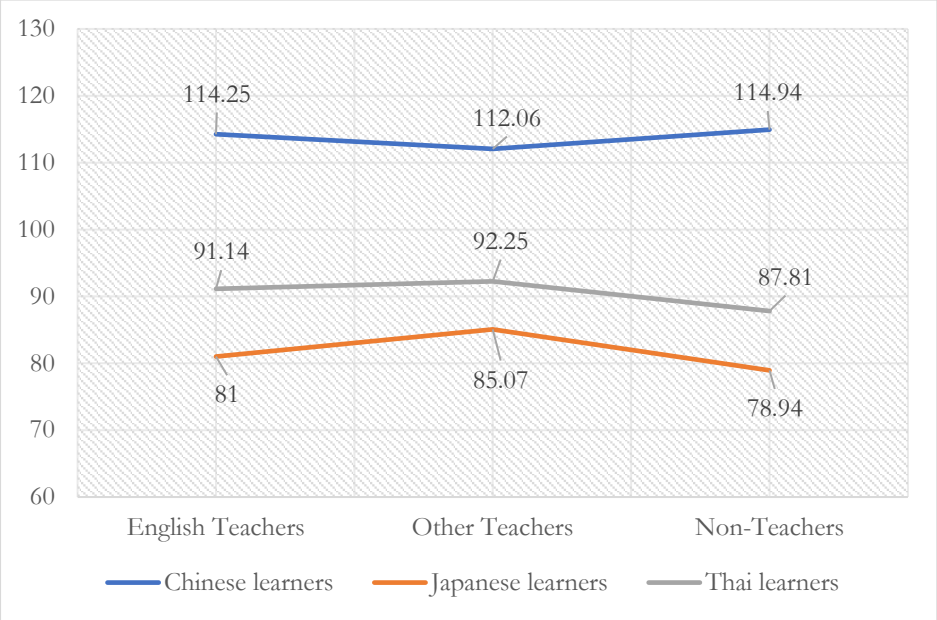
First, regarding the trend of change, Figure 3 represents the mean TRS values of the three rater subgroups based on teaching experience assigned to the speech samples of the three learner groups.

Among the three rater subgroups, English teachers assigned the middle-level scores to all three learner groups. In the case of the assessment of Thai learner speeches, for example, the TRS was 91.14, which was between the maximum value (Other teachers: 92.25) and the minimum value (Non-teachers: 87.81). The hypothesis that the mean TRS steadily changes in the order of English teachers, other teachers, and non-teachers was not supported by the current dataset.

Second, regarding the significance of the effect of teaching experience as a rater variable, the two-way ANOVA tests proved that the interaction between the raters' teaching experience and the learner groups was not significant ($F(4, 3591)=0.95$, $p= .436$), and the main effect of the raters' teaching experience was not significant either ($F(2, 3591)=0.96$; $p=.383$). Thus, the hypothesis that English teacher raters assess speech samples in a different manner from others was not supported.

Figure 3

Mean TRS Assigned by Three Rater Subgroups Based on Teaching Experience



Next, the CV and Cronbach's alpha values are shown in Table 6.

Table 6

Assessment Quality for Three Rater Subgroups Based on Teaching Experience

Index	Rater Subgroups	Learner Groups		
		Chinese	Japanese	Thai
CV (%)	English Teachers	32.45	46.30	45.95
	Other Teachers	30.07	49.27	42.53
	Non-Teachers	28.53	46.49	48.41
Cronbach's alpha	English Teachers	0.94	0.95	0.99
	Other Teachers	0.66	0.84	0.82
	Non-Teachers	0.94	0.93	0.97

Third, regarding assessment sensitivity, English teachers presented the highest CV value for Chinese learners (32.45 vs 28.53–30.07), the lowest value for Japanese learners (46.30 vs 46.49–49.27), and the middle value for Thai learners (45.95 vs 42.53–48.41). Non-teachers showed the lowest value for Chinese learners (28.53), the middle value for Japanese learners (46.49), and the highest value for Thai learners (48.41). Thus, no stable relationship between raters' TESOL backgrounds and assessment sensitivity was found.

Finally, regarding inter-rater reliability, English teachers presented the highest Cronbach's alpha values when assessing all three learner groups (0.94–0.99). This seems to suggest a positive effect of the TESOL background, but considering that even non-teachers showed quite high values (0.93–0.97), the effect is subtle.

Notably, Cronbach's alpha values were higher than .7 for both other teachers and non-teachers, except when other teachers assessed Chinese learner speeches (0.66), which underpins the idea of inviting people with a variety of occupational backgrounds, including business people, to L2 speech assessments rather than depending solely on the judgement of English teachers. Considering that English is a practical tool for global communication, "good speech" should be

evaluated not only from an English teacher's viewpoint but also from the viewpoint of various people who participate in such communication.

Conclusion

Using data from the ICNALE GRA, this study quantitatively re-examined the assessments of Chinese, Japanese, and Thai learners' L2 English speech by a variety of rater subgroups to clarify the degree of the effects of three different rater-related variables (L1s, assessment experience, and teaching experience) on their assessment performance. Table 7 summarises the findings from the analyses with a focus on (i) the trend of change in the total rating score, (ii) the statistical significance of the effect of a particular rater-related variable, (iii) the degree of variation or sensitivity in the assessment scores, and (iv) inter-rater reliability.

Table 7

Summary of the Findings in the Current Analyses

	Change in TRS	**Effect of a Variable**	**Assessment Sensitivity**	**Inter-rater Reliability**
RQ1 (L1)	No clear gaps were seen between NENS/NNS raters	Significant for all three learner groups, but the difference between NENS/NNS raters was not significant.	The order of NENS>NNS was not confirmed.	The order of NENS>NNS was not confirmed, and the latter also showed a value of > .7.
RQ2 (Assessment Experience)	No clear gaps were seen between experienced and less experienced raters	Significant only for one learner group, and the difference between subgroups was not significant.	The order of experienced> less experienced was not confirmed.	The order of experienced> less experienced was suggested, and the latter also showed a value of > .7.
RQ3 (Teaching Experience)	No clear gaps were seen between Eng. teachers, other teachers, and non-teachers	Not significant for all three learner groups.	The order of Eng. teachers > Others was not confirmed.	The order of Eng. teachers> others was suggested, and the latter also showed a value of > .7.

As mentioned before, English teachers and researchers, especially in Asian EFL regions where opportunities to use English for oral communication have been relatively scant, often believe that being an NS and having sufficient experience in L2 speech assessment and English language teaching are prerequisites for a reliable speech assessment. However, the present study, which uses a larger assessment dataset and adopts a careful approach to data analysis by controlling for the possible interference of learner-related variables (L1s and L2 proficiency levels), exemplified that none of the raters' L1s, assessment experience, and English teaching experience significantly influenced their assessment performance. This supports the findings of Ishikawa's (2023) study, which analysed 140 speech samples as a single set without considering their internal variance, and supports the findings of some previous studies suggesting that L1 (Gui, 2012; Wei & Llosa, 2015), assessment experience (Isaacs & Thomson, 2013), and teaching experience (Kang et al., 2019) may not clearly influence raters' assessment performance.

The findings of this study offer two suggestions. First, it has been shown that the quality of L2 speech assessment is likely to be influenced by a complex combination of many rater/learner-related variables. We observed several cases in which a difference between rater subgroups was suggested for one learner group but not for the other learner groups. This means that analysing a variety of learners and raters as a whole without considering the possibility of

their internal variance might be problematic and even misleading. Researchers should be aware that L2 speech assessment performance can easily vary according to the rater and learner type.

Second, the present study proved that if a detailed assessment guide is prepared, an appropriate rater screening is conducted, and statistical score adjustment is made, NNS raters, less experienced raters, and raters without TESOL backgrounds can also assess the student speech samples with a satisfactory level of consistency. This clearly casts doubts on the conventional view to blindly prioritise experienced NS-teacher raters and at the same time to marginalise other raters in L2 assessments, which is closely related to what Holliday (2006) called native-speakerism, "a pervasive ideology within ELT [English language teaching], characterised by the belief that 'native-speaker' teachers represent a 'Western culture' from which spring the ideals both of the English language and of English language teaching methodology." Although such a belief is still deep-rooted, TESOL practitioners in Asia may need to intentionally and strategically put aside the conventional "dominant professional discourses" to readdress this issue "'at the level of the prejudices embedded in everyday practice," which enables understanding of "the meanings and realities of students and colleagues from outside the English-speaking West" (Holliday, 2006). Realising the principle of "diversity and inclusivity" and involving a greater variety of raters in L2 assessment is a first step for future change.

Finally, the present study has four limitations: (i) analysis of persuasion roleplays, (ii) rater training, (iii) score adjustment, and (iv) self-reporting in the background survey. First, it focused exclusively on assessments of learner utterances in L2 persuasion roleplays. Roleplays are regarded as one of the reliable L2 output elicitation measures in that they "represent oral production, full operation of the turn-taking mechanism, impromptu planning decisions contingent on interlocutor input, and hence negotiation of global and local goals, including negotiation of meaning" (Kasper and Dahl, 1991, p. 228), but they do not represent the whole range of learners' speech, including both monologues and dialogues or both casual and formal speech. Second, in the ICNALE GRA project, all raters were given a detailed assessment guide and requested to pass the check test. This means that even novice raters with no prior assessment experience may have had a chance to learn about the basics of L2 speech assessment. Third, all raters were required to adjust their rating scores so that both the mean and standard deviation values fell rigidly within the preset range, which might have neutralised the possible differences between rater subgroups. Fourth, rater variables were investigated based on raters' self-reports. Although we offered a sufficient explanation when conducting a rater background survey, there remains the possibility that some raters might have overrated or underrated their own experience in speech assessment as well as teaching. This suggests the need to be careful about the generalisability of the findings of the present study.

Acknowledgements

This paper is based on the author's oral presentation at the International Hybrid Conference on Diversity and Inclusivity in English Language Education (9–10 December 2022) hosted by the Language Institute of Thammasat University, Thailand. This study is supported by the Japan Society for the Promotion of Science (JSPS)'s Kakenhi grant (20H01282).

About the Author

Shin'ichiro Ishikawa: A professor of applied linguistics at Kobe University, Japan. His research interests include corpus linguistics, SLA, and applied linguistics. He is a principal researcher in the ICNALE learner corpus project.

References

ACTFL (2019). *Diversity and inclusion in world language teaching & learning.* https://www.actfl.org/advocacy/actfl-position-statements/diversity-and-inclusion-world-language-teaching-learning

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, J. D. (2022). Classical test theory. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 323–335). Routledge.

Chau, M. H., Lie, A., Jacobs, G. M., & Renandya, W. A. (2022). Introduction: Promoting diversity and inclusion in language education through research and practice in global Englishes and translanguaging. *TESL-EJ*, *26*(3). https://doi.org/10.55593/ej.26103a0

Dalman, M., & Kang, O. (2019). Listener background in L2 speech evaluation. In N. Feza (Ed.), *Metacognition in learning* (pp. 1-14). Intech Open. DOI: 10.5772/intechopen.89414

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*, 186–203.

Hall, C. S., & Hope, A. K. (2016). Tips for testing speaking. *TESOL Connections 2016 April*, 1–4. http://newsmanager.commpartners.com/tesolc/issues/2016-04-01/3.html

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(1), 1–24. https://doi.org/10.7916/salt.v16i1.1261

Holliday, A. R. (2005). *The struggle to teach English as an international language*. Oxford University Press.

Holliday, A. R. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387. https://doi.org/10.1093/elt/ccl030

Honda, Y., & Tsuchidate, S. (2022, November 28). *Tokyo includes 1st speaking test for English in entrance exam*. The Asahi Shimbun. https://www.asahi.com/ajw/articles/14779198

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 47–74.

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, *8*(1), 1–17. https://doi.org/http://dx.doi.org/10.1186/s40468-018-0069-0

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgements of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Ishikawa, S. (2019). The ICNALE Spoken Dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching*, *74*(4), 153–177.

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, *5*, 121–144.

Ishikawa, S. (2023). *The ICNALE Guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

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.

Kang, O., Rubin, R., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, *12*, 215–247.

Lee, H. (2017). The effects of rater's familiarity with test taker's L1 in assessing accentedness and comprehensibility of independent speaking tasks. *SNU Working Papers in English Linguistics and Language*, *15*, 93–111.

Mohd Noh, M.F., & Mohd Matore, M.E.E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, *13*: 941084. https://doi.org/10.3389/fpsyg.2022.941084

Mohd Noh, M.F., Mohd Matore, M.E.E., Niusila Faamanatu-Eteuati, & Norhidayu Rosman (2021). Rating quality in rater mediated language assessment: A systematic literature review. *Journal of Contemporary Issues in Business and Government*, *27*(2), 6096–6116. https://doi.org/10.47750/cibg.2021.27.02.606

Price, P. C., Jhangiani, R. S., Chiang, I-C., A., Leighton, D. C., & Cuttler, C. (2017). *Research methods in psychology*. (3rd American ed.). Pressbooks. https://opentext.wsu.edu/carriecuttler/

Saito, K., & Shintani, N. (2016). Foreign accentedness revisited: Canadian and Singaporean raters' perception of Japanese-accented English. *Language Awareness*, *25*(4), 305–317. https://doi.org/10.1080/09658416.2016.1229784

UCLA Advanced Research Computing (n.d.). *What does Cronbach's alpha mean? SPSS FAQ*. https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/

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

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Winke, P., Gass, S., & Myford, C. M. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series 2011(2)*, i–67. https://doi.org/10.1002/j.2333-8504.2011.tb02266.x

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL *i*BT speaking section and what kind of training helps? *ETS Research Report Series 2009(2)*, i–37. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x

Yan, X., & Fan, J. (2022). Reliability and dependability. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 477–494). Routledge.

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306–325. https://doi.org/10.1080/0969594X.2013.845547