Exploring the Relationships among Raters' Attitudes toward Accentedness, Tolerance of Ambiguity, and Rating Behaviors in Speaking Assessment

Yanping DENG

Introduction

A rater effect in speaking assessment refers to "an inconsistency introduced into the rating process by the raters themselves" (Winke, Gass, & Myford, 2012, p. 234). Statistically, it is defined as the variance which is related to raters themselves but unrelated to test-takers' speaking ability. Thus, this effect threatens the validity of the test. To shed light on raters' inconsistent rating behaviors or individual rating patterns, such as severity or leniency, researchers have attempted to investigate the sources of rater effects in speaking tests (e.g., Gui, 2012; Isaacs & Thomson, 2013; Winke & Gass, 2013) from the perspective of rater language backgrounds such as familiarity with test-takers' first languages (L1s) and rating experience.

Among these background variables, rater attitudes have been investigated to interpret raters' rating processes and results through employing qualitative analysis methods (e.g., interviews, think-aloud protocols, and written comments) in recent years. While the findings of these studies are inconclusive, some researchers have argued that the varied attitudes toward accented English may explain raters' idiosyncratic rating patterns (Kang, 2008; Wei & Llosa, 2015). Carey, Mannell, and Dunn (2011) also proposed that raters' attitudes toward accentedness should be examined in rater effect studies.

In addition, personality traits are considered to affect individual behaviors. Among them, tolerance of ambiguity (TA) has been found to be related to various variables such as living abroad, language learning experience, and language learning strategies (Dewaele & Wei, 2013; Ely,1989). Some of these backgrounds, such as living abroad and language learning experience, have been frequently examined to understand raters' ratings. However, few studies have investigated the relationship between TA and rater rating behaviors in the language testing field.

To address these research gaps, the author examined Chinese raters' attitudes toward accentedness in the speaking test and the relationship among attitudes toward accentedness, TA as a rater background variable, and rating behaviors. In so doing, this investigation of Chinese raters seeks to achieve a better understanding of raters' idiosyncratic rating behaviors.

Literature Review

Previous studies on rater bias and variability have shown that, even if rating results are consistent, raters may exhibit distinctive rating patterns with variation in severity (e.g., Gui, 2012; Yan, 2014). Some of these studies (e.g., Winke et al., 2012; Winke & Gass, 2013; Weigle, 1998; Xi & Mollaun, 2011; Zhang & Elder, 2011) have suggested that rater language backgrounds may serve as probable cues for rater effects because they may bias rating processes and rating results. Such backgrounds include the experience with rating language tests, rater training, language teaching, living abroad, and familiarity with test-takers' L1s. However, few such studies have examined raters' attitudes, which may play an important role in raters' scoring processes and results (Wei & Llosa, 2015). Furthermore, TA is a psychological trait and has been found to be related to language backgrounds and language attitudes. Thus, this section reviews previous studies on language attitudes as well as TA.

Language and Rater Attitudes

In language attitude studies, native or standard varieties of English tend to be judged more positively than non-standard and accented varieties of English in terms of "status," such as traits of power, competence, social status, or intelligence by native or non-native English speaker judges (Cargile & Giles, 1998; McKenzie, 2008). This phenomenon is prominent in international teaching assistants (ITAs) in the United States, which came into being along with the rise of research universities in it since the 1860s (Minkel, 1987). U.S. college students generally hold negative attitudes toward ITAs' foreign accents and pronunciation. Their complaint that the foreign-accented English affects their understanding the course has been frequently reported although most of the U.S. states have legislated ITAs' English proficiency standards (Monoson & Thomas, 1993), and some of the universities have provided systematic programs of ITA training and English proficiency testing, such as those conducted at Iowa State University reported by Plakans (1997). In contrast, non-standard and accented varieties of English tend to be judged more positively than the standard ones in terms of "solidarity," also called "attractiveness," related to traits of kindness, warmth, honesty, or friendliness (Cargile & Giles, 1998; McKenzie, 2008).

Learners and teachers of English as a second or foreign language (ESL and EFL, respectively) reported the preference for native or standard varieties of English. For instance, Japanese students in Sasayama (2013) and McKenzie's (2008) studies showed that they rated American English higher in power than Japanese English and Japanese English more positively in solidarity compared with American English. Furthermore, Japanese students also expressed their desire to achieve native-like pronunciation (Matsuda, 2003; Sasayama, 2013). A similar preference for native or standard English has been found among Malaysian undergraduates (Lin, Choo, Kasuma, & Ganapathy, 2018) and Chinese college students and teachers (He & Li, 2009) as well.

When it comes to language assessment studies on raters' attitudes toward accentedness, some researchers mentioned that rater attitudes might also explain their rating results and decision-making processes (e.g., Kang, 2008). However, only a few studies have been conducted to examine this issue in speaking proficiency assessment. For instance, Cai (2015) noted that raters' attitudes, which may lead to weighing different features of speeches differently across raters, were not reflected frequently in the actual rating process. In contrast, Wei and Llosa (2015) treated raters' attitudes toward non-native test-takers' accentedness as a dichotomy by counting frequencies of positive and negative comments on raters' attitudes obtained in their interviews. They found that raters' attitudes might affect their rating assignment. However, Wei and Llosa's dichotomous method limits our understanding of raters' attitudes, suggesting the need for more in-depth investigations into this issue from broader perspectives.

Worthy of note here is that the degree of accentedness cannot be ignored when the relationship between raters' attitudes toward accentedness and speaking proficiency ratings is considered. Some previous studies addressed this issue by comparing listeners' attitudes toward accentedness of different degrees (e.g., Cargile & Giles, 1998; McKenzie, 2008). McKenzie (2008) suggested, for example, that the speech of heavily accented Japanese English was rated by Japanese learners lower for competence but rated higher for attractiveness than native English varieties and moderately accented Japanese English. Thus, the different degrees of accentedness should be taken into account in investigating raters' attitudes toward accentedness as well. Currently, it remains unclear as to how raters' attitudes toward English with different degrees of accentedness affect accentedness and speaking proficiency ratings in an English-speaking test.

Tolerance of Ambiguity (TA)

TA has been investigated in the fields of psychology, multilingualism, and second language learning (e.g., Dewaele & Wei, 2013; Ely, 1989). Herein, TA is defined according to Budner (1962) as "the tendency to perceive ambiguous situations as desirable" (p. 29), and intolerance of ambiguity as "the tendency to perceive (i.e., interpret) ambiguous situations as sources of threat" (p. 29). TA is the reaction along a continuum from rejection to acceptance of unfamiliar, new, complex, or uncertain situations that provide insufficient information (McLain, 1993). Some people who perceive an ambiguous situation (e.g., a new language or unfamiliar accent) to be unthreatening are willing to be tolerant of and accept the unclear situation, showing great adaptiveness to it.

Regarding instruments to measure TA, Budner (1962) devised a 16-item Likert scale to measure the degree of TA. Since the average internal consistency of Budner's scale was only .49, showing weak reliability, Herman, Stevens, Bird, Mendenhall, and Oddou (2010) developed a new measure called the Tolerance for Ambiguity Scale (TAS). The internal consistency of this five-point Likert scale with 12 items reported by Herman et al. was acceptable, at .73 (coefficient alpha). The TAS was widely used across previous studies conducted in different cultures and contexts. For instance, in an ESL context, Behresi, Moulaei, and Motlag (2016) found a significant relationship between TA scores and students' listening comprehension. In the field of multilingualism, Dewaele and Wei (2013) investigated the relationship between TA and linguistic backgrounds among 2158 multilinguals from 204 countries, whose average language levels ranged from low to high. The results showed that participants with experience in living abroad for more than three months had statistically significantly higher TA scores than those without the experience in living abroad. They also confirmed that multilinguals scored significantly higher on TA than bilinguals, who in turn scored significantly higher than monolinguals.

As described above, TA is related to language backgrounds. The effect of language backgrounds on rating results has been frequently investigated in rater effect research. Meanwhile, TA is also related to attitudes toward accentedness, which is confirmed by Dewaele and McCloskey (2015). They used an online questionnaire to investigate 2035 multilinguals' attitudes toward accentedness and found that those tolerant of ambiguity were less bothered by others' accentedness. Thus, there appear to be some complex links among raters' attitudes toward accentedness, rating behaviors, and TA. However, there is a dearth of research on their relationships. Exploring rater attitudes along with their relationships to these variables helps to explain raters' scoring results, which may contribute in turn to developing effective programs for rater training and monitoring rater attitudes.

Research Questions

As an attempt to address the research gaps mentioned above, this study aims to examine raters' attitudes toward accentedness and its relationships to TA and rating behaviors. In detail, three research questions are addressed in this study:

1. How do the raters rate accentedness and overall English-speaking proficiency?
2. How are the raters' TA and attitudes toward accentedness characterized?
3. What is the relationship among scores assigned by raters, TA, and attitudes toward accentedness?

Method

Participants

The raters were 32 Chinese doctoral students (16 females, 16 males) studying in Tokyo, with an age range of 25–33 years. They were from different fields of specialization, such as biology, computer vision, and education. They were defined as inexperienced raters as they reported that they had no previous experience of assessing others' speaking proficiency prior to their participation in this study. Their English proficiency was high according to their self-reported scores on TOEFL, ILETS, CET, or other English proficiency tests. English was the medium of instruction at their universities. Their oral Japanese proficiency was limited to some frequently used fixed expressions. All the raters had the experience of taking the TOEFL iBT test or simulation tests.

Materials

Speech recordings. The TOEFL speaking test was selected for use in this study because most of the raters were familiar with the test. An independent speaking task prompt was chosen from a website called *tuofuxiaozhan* (http://toefl.zhan.com/), which is a Chinese website authorized by ETS to provide past TOEFL speaking topics publicly to test-takers. The prompt used in this study was "Do you agree or disagree with the statement that people's personality never changes?" This topic was not biased for or against anyone and was related to everyone's life. Thus, this topic was deemed suitable for this study. To complete the task, test-takers needed to state their opinions and provide reasons to support their opinions within 45 seconds.

Speech samples employed in this study were collected online from 27 Chinese students of a Chinese university, who voluntarily participated in this study. They were undergraduate students from various fields of specialization. The 27 test-takers' responses were rated based on TOEFL iBT independent speaking rubrics (see below for further details) by two raters specializing in language testing with English teaching experience. Among these 27 responses, 10 completed responses represented a range of TOEFL independent speaking rubrics scores (i.e., score one, two, three, and four) and Chinese accentedness, covering 1–9 points on Munro, Derwing, and Burgess's (2010) accentedness described below and did not include filled pauses (e.g., *ah, eh*) or noticeable silent pauses which may affect comprehensibility and task completion within the required time. Accordingly, these 10 responses were selected for rating. Rating the 10 responses did not trigger raters' fatigue according to some raters' feedback collected during a pilot study conducted prior to this study.

Questionnaires. Two online questionnaires (see Table 1) were devised to investigate raters'

Table 1. The Structure of the Two Questionnaires

Questionnaire	Section	Content	N of items
Questionnaire 1	A	Sociobiographical and language backgrounds	25
	B	Tolerance of ambiguity	12
Questionnaire 2	C	Attitudes toward accent of self	8
	D	Attitudes toward accents of others	7

backgrounds, TA, and attitudes toward accentedness. Section B of Questionnaire 1 concerns raters' TA based on Herman et al.'s (2010) 12-items TAS. Sections C and D of Questionnaire 2 were designed for the purpose of this study to assess attitudes toward self-accent and others' accents based on the results of a pilot study, where seven raters reported that they wanted to get rid of their accents but to be generous with others' foreign accents. Accent has been defined by Derwing and Munro (2009) in two ways. The first definition emphasizes the influence of L1 on L2 phonology while the second focuses on the difference between individuals' non-native speech and a local variety. Considering the two perspectives and the fact that standard American or British English is frequently used as a teaching model in China, this study defined self- and others' accents as self and others' L1-accented English, which is different from a standard English variety such as standard American or British English. In addition, the attitudes toward self- and others' accents were measured by items related to raters' concern about self- and others' L1-accented English, which is different from a standard English variety such as standard American or British English. All items in Questionnaire 2 were on a five-point Likert scale ranging from 1 point (strongly disagree) to 5 points (strongly agree).

Rating scales on accentedness and overall speaking proficiency. Two scales were adopted to assess accentedness and overall speaking proficiency. A nine-point Likert scale for accentedness ranging from "No accent" (1 point) to "Very strong accent" (9 points) was adopted from Munro et al.'s (2010) study on listeners' detection of non-native speaker's status. This scale is widely employed by many researchers. Munro et al. defined accentedness as "the degree to which the pronunciation of an utterance sounds different from an expected production pattern" (p. 112). For rating overall speaking proficiency, the TOEFL iBT independent speaking rubrics with the corresponding descriptors (https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf) were adopted. There were five points in the rubrics ranging from 0, representing that "speaker makes no attempt to respond or response is unrelated to the topic" to 4, showing that "the response fulfills the demands of the task, with at most minor lapses in completeness, and it is highly intelligible and exhibits sustained, coherent discourse." The holistic rating for overall speaking proficiency was based on three dimensions: language use, delivery, and topic development. Language use concerns the effective use of grammar and vocabulary to express the idea clearly. Delivery refers to fluent and intelligible speech with clear and accurate pronunciation and intonation, whereas topic development is related to the fully elaborated and coherent speech with clear ideas.

Procedures

The 32 raters completed three tasks individually in the same quiet university classroom on different days at their convenience. First, they completed the first questionnaire on their backgrounds and TA. Second, they listened to and rated the speech recordings for accentedness and overall speaking proficiency. During this rating session, the raters were instructed first to familiarize themselves with the rating criteria. The researcher explained the rating procedure. Then the raters rated three speech samples selected from the same speech sample pool for practice. They could ask any questions whenever they felt confused, and the researcher answered them. When they had completed the exercise, they were asked to rate the 10 speech samples that were not included in the practice set described above. They could pause anywhere and repeat listening to gain adequate information to make a decision and ask any questions related to their ratings. Finally, after completing the rating task, the raters were instructed to fill out the second questionnaire on their attitudes toward accentedness. It took approximately 30–45 minutes for the raters to complete the three tasks.

Analyses

All analyses presented below were conducted on SPSS (IBM SPSS Statistics). For these analyses, mean ratings across the 10 speech samples were obtained for each rater on accentedness and overall speaking proficiency. The total scores were calculated for Sections B, C, and D of the two questionnaires. As part of preliminary analyses, the author first obtained descriptive statistics to examine the normality of score distributions. Internal consistency reliability estimates (Cronbach's alpha) were examined for Sections B, C, and D as well. To answer Question 1 (How do the raters rate accentedness and overall English-speaking proficiency?), the author examined two grand means of the 32 raters' ratings on accentedness and overall speaking proficiency and their standard deviations. Inter-rater reliability (IRR) was assessed by using a two-way mixed, consistency, average measures intra-class correlations (ICCs) (McGraw & Wong, 1996). A two-way model ICC was used because all speeches were rated by all raters, which means a fully crossed design. Raters were not randomly selected, and the author was not to generalize these ratings to a larger rater population. Thus, a mixed model, where raters were considered to be a fixed effect but speeches were considered random, was appropriate. Moreover, it is important for raters to assign scores similarly in terms of rank ordering instead of the absolute score level; therefore, a consistency type ICC was used. All speeches were rated by the same set of raters, and the author's main interest was in the reliability of the mean ratings provided by all raters. Thus, an average-measure unit ICC was suitable. Then, because the score distributions were normal (see Table 2 below for details), Pearson correlation coefficients were calculated to uncover the relationship between accentedness and overall speaking proficiency ratings.

For Question 2 (How are the raters' TA and attitudes toward accentedness characterized?), the author examined the descriptive statistics for scales of TA, attitudes toward self-accent, and attitudes toward others' accents, obtained from Sections B, C, and D. In addition, percentages and frequencies of the categories for each item in these scales were calculated to achieve a better understanding of raters' TA and language attitudes.

Concerning Question 3 (What is the relationship among scores assigned by raters, TA, and attitudes toward accentedness?), Pearson correlation coefficients were analyzed to shed light on relationships among these variables.

Results

Descriptive statistics on all five variables employed in this study are presented in Table 2. Values of Cronbach's alpha for TA, attitudes toward self-accent, and others' accents were calculated (.20, .66, and .80 respectively), suggesting an acceptable to a high level of reliability (i.e., above .60) except for TA (Nunnally, 1978). The descriptive statistics (mean, standard deviation, skewness, and kurtosis) in Table 2 show that the five variables were normally distributed. Accordingly, it was determined to obtain Pearson correlation coefficients to analyze the relationships among these variables. Detailed results are discussed below in relation to the corresponding research questions.

How Do the Raters Rate Accentedness and Overall English-speaking Proficiency?

As shown in Table 2, the grand mean of mean accentedness ratings across the 32 raters was 3.99 (*SD* = .93). As the nine-point scale of accentedness ranges from "No accent" (1 point) to "Very strong accent" (9 points), this grand mean indicates that the raters generally perceived non-native accents in the 10 speech samples albeit with a relatively low degree of accentedness. The standard deviation was only .93 on the nine-point scale, showing that the raters' ratings for accentedness were not widely varied. As for the ratings of overall speaking proficiency, the mean rating was 2.55 (*SD* = .51). According to the five-point independent speaking rubrics, this score represents that the 10 speech responses were generally intelligible, fluent, and coherent. The standard deviation was .51, showing a narrow distribution of raters' ratings for the overall speaking proficiency. Thus, the raters generally perceived that the responses represented a high level of English-speaking ability.

As to the consistency among ratings assigned by the 32 raters, Table 3 shows the inter-rater reliability assessed by ICCs. According to Cicchetti (1994), higher ICC means higher IRR, and the value of 1

Table 2. Descriptive Statistics for All Variables (*N* = 32)

Variable	Full score	*M*	*SD*	Skewness	Kurtosis	*Min*	*Max*
Accentedness	9	3.99	.93	−.21	.10	1.80	5.90
Overall speaking proficiency	4	2.55	.51	.24	−.40	1.70	3.80
Tolerance of ambiguity	60	38.41	3.55	.14	−.77	32.00	45.00
Attitudes toward accent of self	40	27.28	4.18	.40	−.47	20.00	37.00
Attitudes toward accents of others	35	22.53	4.63	−.61	.57	11.00	32.00

means perfect agreement across the ratings. The ICCs were .96 (95% confidence interval, .91 to .99) for accentedness and .95 (95% confidence interval, .89 to .98) for overall speaking proficiency, indicating a high level of inter-rater consistency in rank-ordering. Finally, regarding the relationship between accentedness and overall speaking proficiency ratings, Table 4 shows a statistically significant and moderate negative relationship with a correlation coefficient of -.49. It suggests that, on average, the lower speaking proficiency ratings assigned by raters were associated with heavier accentedness perceived by them, and vice versa. It should be noted, however, that the result cannot be overinterpreted. One reason is the small sample size of the 10 speech samples used to calculate means to represent rater's severity levels in rating accentedness and overall speaking proficiency. Another is the limited variability in both variables that might be unique to this study sample, which might have weakened the observed correlation between them.

To summarize, concerning Question 1, these 32 inexperienced raters rank-ordered the speech samples consistently on accentedness and overall speaking proficiency. Additionally, the raters tended to give low speaking proficiency scores to speeches with relatively heavy accents.

How Are the Raters' TA and Attitudes toward Accentedness Characterized?

TA. As shown in Table 2, the mean TA total score was 38.41 (*SD* = 3.55). Each item in Section B of Questionnaire 1 was rated on the five-point scale. Thus, the possible total score for the 12 statements in the TAS, calculated by reversing responses to negatively worded items, ranges from the minimum score of 12 to the maximum score of 60 with a midpoint of 36. The mean TA score (38.41), slightly larger than

Table 3. Inter-rater Reliability Estimates and Confidence Intervals

	Accentedness	Overall speaking proficiency
Intra-class correlation	.96 (.91–.99)	.95 (.89–.98)

Note. The numbers in brackets are 95% confidence intervals.

Table 4. Pearson Correlation Coefficients among All Variables

Variable	Accentedness	Overall speaking proficiency	TA	Accent of self	Accents of others
Accentedness	1.00				
Overall speaking proficiency	−.49**	1.00			
TA	−.20	.03	1.00		
Accent of self	.15	.12	−.37**	1.00	
Accent of others	.23	.02	−.20	.56**	1.00

$^*p < .05$; $^{**}p < .01$.

36, suggests that the raters had neither a high nor low tolerance of ambiguity. Table 5 lists the percentage and frequency data for their ratings to the individual items.

According to Items 1, 4, and 5, a large proportion of the raters showed a neutral attitude regarding familiar situations. In addition, the percentages of "Neutral" were not small for most of the items except for Items 3 and 8. Raters also showed contradictory answers to these items. For instance, most of the raters agreed on Items 2, 3, 7, and 8, suggesting that they were tolerant of ambiguity, while they tended to agree on Items 9, 10, 11, 12, indicating that they were not tolerant of ambiguity. Interestingly, although most of the raters admitted that they could be tolerant of people whose values were different from theirs (Item 2, 46.9%), they also expressed that they could not be tolerant of all kinds of people (Item 6, 40.6%). Therefore, a consistent attitude tendency did not emerge based on the responses, which might partly explain the low Cronbach's alpha for this scale.

Attitudes toward self-accent. Table 6 presents the percentage and frequency data for Questionnaire 2. More than half of the participants (59.4%) suggested that they did not want to speak English with

Table 5. Percentage and Frequency Data for the TAS

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1. I avoid setting where people don't share my values.	0% (0)	32.3% (10)	40.6% (13)	25% (8)	3.1% (1)
2. I can enjoy being with people whose values are very different from mine.	3.1% (1)	43.8% (14)	34.4% (11)	12.5% (4)	6.3% (2)
3. I would like to live in a foreign country for a while.	43.8% (14)	46.9% (15)	3.1% (1)	0% (0)	6.3% (2)
4. I like to surround myself with things that are familiar to me.	0% (0)	38.7% (12)	45.2% (14)	12.9% (4)	3.2% (1)
5. The sooner we all acquire similar values and ideals the better.	0% (0)	31.3% (10)	31.3% (10)	28.1% (9)	9.4% (3)
6. I can be comfortable with nearly all kinds of people.	3.1% (1)	28.1% (9)	28.1% (9)	40.6% (13)	0% (0)
7. If given a choice, I will usually visit a foreign country rather than vacation at home.	28.1% (9)	53.1% (17)	15.6% (5)	0% (0)	3.1% (1)
8. A good teacher is one who makes you wonder about your way of looking at things.	31.3% (10)	56.3% (18)	12.5% (4)	0% (0)	0% (0)
9. A good job is one where what is to be done and how it is to be done are always clear.	15.6% (5)	34.4% (11)	25% (8)	25% (8)	0% (0)
10. A person who leads an even, regular life in which few surprises or unexpected happenings arise really has a lot to be grateful for.	9.4% (3)	34.4% (11)	31.3% (10)	25% (8)	0% (0)
11. What we are used to is always preferable to what is unfamiliar.	0% (0)	40.6% (13)	28.1% (9)	31.3% (10)	0% (0)
12. I like parties where I know most of the people more than ones where all or most of the people are complete strangers.	12.5% (4)	43.8% (14)	18.8% (6)	21.9% (7)	3.1% (1)

a Chinese accent (Item 1). Interestingly, compared with the proportion (53.1%) in the situation of communicating with a non-native speaker (Item 3), more raters (68.8%) indicated that they did not want to speak with the Chinese accent when communicating with a native speaker (Item 2).

Corresponding with the result that most of the raters disfavored their Chinese accents, the raters indicated a preference for native-like accents. More than half of the raters agreed that owning a native-like accent was important (Item 4, 62.5%), and approximately half of them (56.3%) showed that a native-like accent was an indicator of one's English speaking proficiency (Item 6). Item 15 showed that 84.4%

Table 6. Percentage and Frequency Data for Questionnaire 2

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Section C: Attitudes toward accent of self					
1. I do not like speaking English with my first language accent.	25% (8)	34.4% (11)	21.9% (7)	15.6% (5)	3.1% (1)
2. When I speak with a native English speaker, I do not want to have my Chinese accent.	18.8% (6)	50% (16)	21.9% (7)	9.4% (3)	0% (0)
3. When I speak with a non-native English speaker, I do not care about my accent.	3.1% (1)	28.1% (9)	15.6% (5)	40.6% (13)	12.5% (4)
4. I think that it is important for me to sound like a native English speaker.	12.5% (4)	50% (16)	18.8% (6)	12.5% (4)	6.3% (2)
6. I believe that the more my speaking sounds like a native speaker, the higher my English-speaking proficiency is.	12.5% (4)	43.8% (14)	31.3% (10)	9.4% (3)	3.1% (1)
13. I think that my Chinese-accented English does not annoy me a lot when I speak with a native speaker.	0% (0)	40.6% (13)	43.8% (14)	15.6% (5)	0% (0)
14. I think that my Chinese-accented English does not annoy me a lot when I speak with a non-native speaker.	3.1% (1)	37.5% (12)	40.5% (13)	18.8% (6)	0% (0)
15. If I have a chance, I want to get rid of my Chinese accent in my English speaking.	25% (8)	59.4% (19)	9.4% (3)	3.1% (1)	3.1% (1)
Section D: Attitudes toward accents of others					
5. I prefer teachers who are from America and Britain.	21.9% (7)	46.9% (15)	18.8% (6)	9.4% (3)	3.1% (1)
7. When I communicate with others in English, I focus more on content than accents.	15.6% (5)	56.3% (18)	15.6% (5)	12.5% (4)	0% (0)
8. Accent should not be regarded as one criterion of judging one's English speaking proficiency.	6.3% (2)	31.3% (10)	34.4% (11)	25% (8)	3.1% (1)
9. I do not like communicating with persons who have heavy accents.	6.3% (2)	43.8% (14)	28.1% (9)	15.6% (5)	6.3% (2)
10. I do not like the English with a heavy accent.	12.5% (4)	53.1% (17)	25% (8)	9.4% (3)	0% (0)
11. I feel more comfortable when I talk with an American than with an Indian.	15.6% (5)	40.6% (13)	37.5% (12)	3.1% (1)	3.1% (1)
12. When I speak to a non-native speaker, I hope she/he has a native-like accent.	3.1% (1)	50% (16)	25% (8)	15.6% (5)	6.3% (2)

wanted to get rid of their Chinese accents.

However, a large proportion of the raters (40.6%) believed that their Chinese accents did not impede their communication with native or non-native speaker interlocutors (Items 13 and 14). Meanwhile, 43.8% and 40.5% of the raters held a neutral attitude regarding whether their Chinese accents affected their communication with native and non-native speaker interlocutors, respectively (Items 13 and 14). Moreover, few raters strongly endorsed these two items, showing that the Chinese accent may bring them some communication difficulties to different degrees.

In sum, the results above show that the Chinese raters tended to prefer a native-like accent and relate accent to speaking proficiency. They also desired to eliminate their Chinese accents as it might bring them some communication difficulties.

Attitudes toward others' accents. In terms of raters' attitudes toward others' accents, a large majority of the raters (71.9%) showed that, when communicating, they paid more attention to the content than accents (Item 7). However, more than half of the raters expressed that they did not like heavily accented English (Item 10, 65.6%). Additionally, 50.1% of the raters expressed that they disliked communicating with the speaker whose English was heavily accented (Item 9). Nearly half of them showed that they liked to speak with a native English speaker (Item 11, 56.2%) and hoped their non-native English speaker interlocutors could speak English with little accent (Item 12, 53.1%). Moreover, regarding accent as a criterion to judge English speaking proficiency (Item 8), 37.6 % of the raters disagreed it, while 34.4% kept a neutral attitude, and 28.1% favored it. Above all, the raters showed a tendency to prefer interlocutors who could speak English with native accents. In addition, they had divergent attitudes toward whether accent should be an assessment criterion of English-speaking proficiency.

What Is the Relationship among Scores Assigned by Raters, TA, and Attitudes toward Accentedness?

Table 4 reports Pearson correlation coefficients among TA, rating behaviors, and accent attitudes. Results indicate that TA, as a personality variable, was not significantly correlated with accentedness, overall speaking proficiency ratings, or attitudes toward others' accents. However, the correlation between TA and attitudes toward self-accent was significantly negative ($r = -.37, p < .01$), suggesting that the lower tolerance of ambiguity of the raters was associated with raters' greater concern about their own accents. In addition, rating behaviors (i.e., rating on accentedness and overall speaking proficiency) did not statistically correlate with raters' concern about self- or others' accents (i.e., accent attitude). In short, the Pearson correlation analysis shows that TA had no significant relationship with the rating behaviors. However, TA was significantly correlated with raters' attitudes toward self-accent instead of their attitudes toward others' accents. Moreover, ratings on accentedness and overall speaking proficiency had no significant relationships with raters' concern about self- and others' accents (i.e., accent attitude).

Discussion

This study investigated raters' attitudes toward accentedness and introduced the TA as one type of rater characteristics that may play a role in rater effect studies. As for Question 1, concerning raters' ratings on accentedness and overall speaking proficiency, the result confirmed that the 32 inexperienced raters' ratings were highly consistent. One possible reason for the high consistency may be the homogeneity of the raters' backgrounds, such as a small range of age, similarities in education levels (i.e., doctoral students), English proficiency levels, and bilingual backgrounds (i.e., English-Chinese bilinguals). However, when the author calculated the IRR based on a single rater's ratings (i.e., the single-measures ICCs), the reliability estimates for accentedness and overall speaking proficiency were only .40 and .36, respectively. Thus, the large number of raters involved in the analysis contributed to the high IRR estimates (i.e., the estimation of the IRR for a situation where 32 ratings were available for a given speech sample boosted the IRR) instead of their homogeneous backgrounds. Accordingly, the high rating consistency across the 32 raters does not necessarily mean that the individual raters would yield reliable enough ratings in typical assessments where each learner's response is scored only by one or two raters.

Contrary to Xi and Mollaun's (2011) finding that the well-trained Indian raters with high English proficiency scored Indian TOEFL examinees consistently (i.e., assigning similar scores to one examinee's performance), these raters in this study still exhibited different rating patterns from each other. As these raters had similar backgrounds, raters' attitudes toward accents may explain their differences in rating patterns. Wei and Llosa (2015) suggested that the homogeneity of raters' backgrounds did not ensure the same language attitude, and rater attitude might play a role in rating. However, this study did not find any relationship between ratings and raters' attitudes toward accentedness. It may be due to the inconsistent rating patterns among these raters and the small sample size.

Additionally, the inexperienced raters in this study could identify the non-native accents in the speech samples but perceived them as not so heavily accented and assigned relatively high scores on overall speaking proficiency to these test-takers. It may be due to their familiarity with the test-takers' Chinese accents. Previous studies (e.g., Gass & Varonis, 1984; Saito, Tran, Suzukida, Sun, Magne, & Ilkan, 2019) have suggested that listeners' familiarity with interlocutors' non-native speeches and language backgrounds may improve their comprehensibility of the discourse. Conversely, Kang, Moran, Ahn, and Park (2020) reported that familiarity with the speech accents did not affect comprehensibility significantly, though the reason for Kang et al.'s divergent finding may be that the listeners in their study were homogeneously not familiar with the accents. Despite the contradictory conclusions regarding the effect of accent familiarity on comprehensibility, the finding of this study seems to concur with the former. In addition, from the perspective of language attitudes, as mentioned above, listeners generally evaluate speakers from the same ethnic group highly on "solidarity" related traits. There is a probability proposed by Sasayama (2013) that the aspects of traits considered by listeners as characteristics of a given ethnic group affect their language attitudes. Therefore, the solidarity may explain the lenient ratings.

In addition to the raters' familiarity with speakers' L1s and L2s, Saito et al. (2019) also found that another two factors (i.e., metacognition and use of English at work) significantly contributed to listeners' comprehensibility. The first factor (i.e., metacognition) is related to the awareness of the importance of comprehensible English compared with a native-like accent for effective communication. As Item 7 in Questionnaire 2 of this study compared the importance of content and accent in communication, it is relevant to this factor. For Item 7, nearly 72% of the respondents agreed that they paid more attention to content than accent, suggesting a high level of metacognition (i.e., high awareness that comprehensible English is more important than a native-like accent for communication). Furthermore, as for the second factor (i.e., use of English at work) concerning the quantity of exposure to the target language, the participants in this study reported that English was their main language instead of other languages in their campus life, suggesting frequent use of English in social life. Thus, given raters' solidarity, metacognition level, familiarity with Chinese, and use of English in campus life, it is reasonable that the Chinese-accented speeches employed in this study were highly comprehensible to these raters, contributing to lenient ratings assigned by them.

The result of the significant and negative relationship between accentedness and overall speaking proficiency is consistent with previous findings that the competence represented in the heavily accented speech was rated negatively (e.g., Cargile & Giles, 1998; McKenzie, 2008; Plakans, 1997). The impact of accentedness on speaking proficiency was reflected in the Questionnaire 2 as well. Most of the raters believed that the degree of accentedness indicated one's English speaking proficiency, although they did not show their attitudes explicitly as to whether accent should be employed as an assessment criterion of speaking proficiency. A plausible reason may be that the raters have experienced many English tests and know that most English speaking tests do not include accentedness explicitly as one criterion. Additionally, according to Questionnaire 1, all the raters reported that they generally learned American or British English in primary, middle, and high schools. Approximately 91% of the raters had been exposed to British, American, and Japanese English. Moreover, among the 32 raters, four raters reported that they had lived in a foreign country except for Japan for more than three months. Thus, although they have more exposure to accented-Englishes than typical students in mainland China, the exposure to this degree still did not cause any change to their entrenched notion of achieving a native-like accent.

Moreover, as these students are now living abroad, their studying and living environment provide many opportunities to use English in their daily life. Chinese English is often called "Yaba English" (dumb English) (Du & Guan, 2016; He, 2019). When communicating with students from other countries, mainly Japan, this label may cause their recall of negative stereotypes of Chinese-accented English, especially when they have communicative difficulties with their interlocutors. Therefore, they might have thought that the accent was related to proficiency and wanted to eliminate their Chinese accents. The negative impressions of stereotypes for own non-native English are pervasive among non-native English-speaking countries, especially the expanding circle countries such as Japan and China (e.g., He & Li, 2009; Sasayama, 2013). However, the present author did not interview their actual English use or communication difficulties in their campus life. Hence, raters' actual feelings and beliefs of Chinese accents in international communication are unknown. It needs further investigation.

Regarding TA, this study did not identify any relationship with ratings on the speech samples. This finding might be due to the low internal consistency of the TAS (.20), while previous studies reported acceptable internal reliability estimates (e.g., .64 in Dewaele & Wei, 2013; .73 in Herman. et al., 2010). A lack of living-abroad and job experience and cultural differences on the part of the raters in this study may explain the observed low internal consistency of the TAS. For instance, it might have been difficult to get reliable responses for Items 3 and 7 in the TAS related to visiting a foreign country, Item 9 regarding job experience, and Item 12 related to a party situation. This is because most of the raters reported that Japan was the only foreign country they had lived in, and 91% of the raters reported that they had no previous working experience. In addition, listeners' homogenous language backgrounds such as living-abroad and the number of learned languages made it difficult to examine its relationship with TA. The limited scope of raters' backgrounds may also explain the insignificant observed correlation coefficients between TA and the two rating scales. In Table 5, the clustering of the ratings around "Neutral" and "Agree" suggests the lack of variability in the raters' responses as well.

The results obtained in this study may have implications for understanding raters' language attitudes and rating behaviors, suggesting that these raters should be exposed to more accented Englishes, which is in line with Xi and Mollaun's (2011) conclusion. As most of the raters agreed that a native-like accent indicated a high English proficiency level, raters' attitudes toward accentedness should be taken up in rater training and monitoring. However, this study suffers from many limitations, notably those related to the small sample size and the homogeneity of raters' language backgrounds. The small sample size makes obtained correlation coefficients unstable. Moreover, the homogeneity of the raters' linguistic backgrounds might have lowered the obtained correlations among them as well. Thus, in future studies, a greater number of raters with more variations in backgrounds should be invited to participate. Furthermore, this work is also limited by the quantitative methods of data analysis. Therefore, future work should adopt qualitative methods such as interviewing raters' real communication in real life and their attitudes toward accented English. By doing so, we could compare different ethnic groups' attitudes and further our understanding of the effect of raters' language attitudes on rating behaviors.

References

Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality*, *30*(1), 29–50.

Behresi, M., Moulaei, A., & Motlag, H. S. (2016). Investigating the relationship between tolerance of ambiguity, individual characteristics and listening comprehension ability among Iranian EFL learners. *International Journal of English Linguistics*, *6*(7).

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

Cargile, A. C., & Giles, H. (1998). Language attitudes toward varieties of English: An American-Japanese context. *Journal of Applied Communication Research*, *26*(3), 338–356.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. Language *Assessment Quarterly*, *12*(3), 262–282.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews?. *Language Testing*, *28*(2), 201–219.

Dewaele, J. M., & Wei, L. (2013). Is multilingualism linked to a higher tolerance of ambiguity?. *Bilingualism: Language and Cognition*, *16*(1), 231–240.

Dewaele, J. M., & McCloskey, J. (2015). Attitudes towards foreign accents among adult multilingual language users. *Journal of Multilingual and Multicultural Development*, *36*(3), 221–238.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language teaching*, *42*(4), 476–490.

Du, H., & Guan, H. (2016). Hindrances to the new teaching goals of College English in China: Being contextually blind and linguistically groundless, current tertiary ELT policy needs to be redefined. *English Today*, *32*(1), 12–17.

Ely, C. M. (1989). Tolerance of ambiguity and use of second language strategies. *Foreign Language Annals*, *22*(5), 437–445.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*(2), 186–203.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of non-native speech. *Language Learning*, *34*(1), 65–87.

He, Y. (2019). Research on the application of cooperative learning in college English teaching. *Theory and Practice in Language Studies*, *9*(10), 1362–1367.

He, D., & Li, D. C. (2009). Language attitudes and linguistic features in the 'China English' debate 1. *World Englishes*, *28*(1), 70–89.

Herman, J. L., Stevens, M. J., Bird, A., Mendenhall, M., & Oddou, G. (2010). The tolerance for ambiguity scale: Towards a more refined measure for international management research. *International Journal of Intercultural Relations*, *34*(1), 58–65.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 181–205.

Kang, O., Moran, M., Ahn, H., & Park, S. (2020). Proficiency as a mediating variable of intelligibility for different variables of accents. *Studies in Second Language Acquisition*, *42*(2), 471–487.

Lin, D. T. A., Choo, L. B., Kasuma, S. A. A., & Ganapathy, M. (2018). Like that lah: Malaysian undergraduates' attitudes towards localised English. *GEMA Online® Journal of Language Studies*, *18*(2).

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of non-native speaker status from content-masked speech. *Speech Communication*, *52*(7–8), 626–637.

McLain, D. L. (1993). The MSTAT-I: A new measure of an individual's tolerance for ambiguity. *Educational and Psychological Measurement*, *53*(1), 183–189.

McKenzie, R. M. (2008). Social factors and non - native attitudes towards varieties of spoken English: a Japanese case study. *International Journal of Applied Linguistics*, *18*(1), 63–88.

Monoson, P. K., & Thomas, C. F. (1993). Oral English proficiency policies for faculty in US higher education. *The Review of Higher Education*, *16*(2), 127–140.

Minkel, C. W. (1987). A graduate dean's perspective. *Institutional responses and responsibilities in the employment and education of teaching assistants, Columbus, OH: The Ohio State University, Center for Teaching Excellence*.

Matsuda, A. (2003). The ownership of English in Japanese secondary schools. *World Englishes*, *22*(4), 483–496.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. *1*(1), 30–46.

Nunnally, J. C. (1978). *Psychometric Theory: 2d Ed*. McGraw-Hill.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly*, *31*(1), 95–119.

Sasayama, S. (2013). Japanese college students' attitudes towards Japan English and American English. *Journal of Multilingual and Multicultural Development*, *34*(3), 264–278.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech?: Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

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.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *Tesol Quarterly*, *47*(4), 762–789.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501–527.

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.

ABSTRACT

Exploring the Relationships among Raters' Attitudes toward Accentedness, Tolerance of Ambiguity, and Rating Behaviors in Speaking Assessment

Yanping Deng

This study reports on the results of an investigation into the relationship among raters' attitudes toward accentedness, tolerance of ambiguity (TA) as one indicator of raters' variables, and their TOEFL speaking test response rating results. Thirty-two inexperienced raters with high English proficiency rated 10 audio recordings of responses to one independent speaking task in terms of accentedness and overall speaking proficiency. Two online questionnaires were administered to investigate the raters' language backgrounds, TA, and attitudes toward accentedness. The results of intra-class correlations (ICCs) showed high inter-rater consistency in rank-ordering the speech samples. Moreover, a Pearson correlation analysis showed a significantly negative relationship between accentedness and overall speaking proficiency ratings and a significantly positive relationship between TA and attitudes toward self-accent. Lastly, the questionnaire results revealed that most of the raters held negative attitudes toward their Chinese accents and positive attitudes toward standard and native accents. Based on these results, the role of contemporary Chinese college students' accent attitudes in rater effect research is discussed.