Language Testing in Asia

RESEARCH

Open Access

Rater attitude towards emerging varieties of English: a new rater effect?

Tammy Huei-Lien Hsu

Correspondence: 089975@mail.fju.edu.tw
English language and literature, Fu Jen Catholic University, New Taipei City, Taiwan

Abstract

Background: A strong interest in researching World Englishes (WE) in relation to language assessment has become an emerging theme in language assessment studies over the past two decades. While research on WE has highlighted the status, function, and legitimacy of varieties of English language, it remains unclear how raters respond to the results of the global spread of English. Also unclear is whether their attitudes towards the varieties of English constitute a biasing factor in the scores they award in English speaking tests. As such, this study investigates the relationship between rater attitudes towards Indian English as an example of WE, as measured by the "rater attitude instrument" (RAI), and scores that raters awarded to IELTS speech samples produced by Indian examinees.

Methods: A total of 96 teacher raters rated six IELTS speech samples and then completed the RAI online. Correlation analysis, MANOVA, and Tukey contrasts were performed to test the extent to which rater attitudes towards Indian English as an example of WE affect rater scoring decisions on IELTS speech samples.

Results: Moderate to strong correlations were observed between the RAI and IELTS speech sample scores. The MANOVA results suggest significantly different ratings, with the positive attitude group consistently awarding higher scores to IELTS speech samples in comparison to the negative attitude group on all of the four analytic rating criteria. Furthermore, the RAI appears to be a significant predictor of IELTS speech sample scores.

Conclusion: A link between rater attitude towards Indian English, as an example of WE, and scoring tendency for Indian examinees may exist in a language assessment context. Thus, as raters reoriented their views, broadened their grasp of WE, and as awareness of WE increased in the language testing community in recent decades, the findings here show that testing agencies must add an understanding of potential rater bias towards WE to the current relevant literature.

Keywords: Rater attitude, World Englishes, Constructs of English language speaking

Background

Practices and discussions in the field of second and foreign language assessment have evolved over the past decades in response to the re-orientation of the English language that sociolinguistic scholars advocate. World Englishes (WE) research has documented the status, function, and linguistic creativity of many different varieties of English, a development which strongly indicates that we can no longer view it as a homogenous entity (Jenkins 2006; Kachru 1992; Smith 1992). The WE paradigm is also an ideology that includes issues such as power, politics, creative writing, pedagogy, and legitimacy through the descriptions and codification of “New Englishes” (Kachru et al. 2006). Given these re-conceptualizations of the English language, an increasing number of language testing (LT) research agendas have emerged on conceptual and empirical levels to contribute to the understanding of the interface between WE and LT. Some scholars are concerned about whether it may be unacceptable or unethical to ignore the diverse varieties of English in use; according to Davidson (2006), this ignorance is an indicator of a new form of imperialism operating in large-scale English tests. LT researchers are being urged by numerous sources to re-examine issues related to target language use (TLU), particularly in regard to international tests (Abeywickrama 2013). Some have expressed concerns over norm selection, test fairness, and authenticity (Canagarajah 2006; Davidson 1993; Lowenberg 2002; Elder and Davies 2006; Elder and Harding 2008; Leung and Lewkowicz 2006), and the input and output languages to be used (Taylor 2002). Questions are also surfacing about English language test designs for WE (Canagarajah 2006; Elder and Davies 2006) and English curriculum design (Brown 2014). In recent years, empirical studies have continued to examine the attitudes of examinees towards the relevance of WE in high-stake tests (Hamid 2014), and the impact on test scores in the listening section (Harding 2008). In particular, studies of rater behavior on English speaking tests have drawn attention to the vital question of a new rater biasing factor in speaking tests. Questions about bias on the part of raters towards the varieties of English in the world today are arising in the relevant discussions (Davies et al. 2003). Studies have investigated the impact of rater nationality on speaking test scores (Chalhoub-Deville and Wigglesworth 2005; Hamp-Lyons and Zhang 2001), differences in scores due to rater attitudes towards Korean English (Kim 2005), and the recent development of a “rater attitude instrument” that measures raters’ attitudes towards WE (Hsu 2016). The emerging agenda on rater psychological traits and attitude-behavior relationship includes broad concerns about the impact of WE on English speaking test scores, score validity (Davies et al. 2003), fairness (Kunnan 2004), and unexpected consequences of test use (Davidson 2006). This situation is a reflection of the post-Messick validity paradigm Messick (1989), which incorporates social dimension into validity inquiry.

The current study, informed by psychology and language attitude research, aims to address the extent to which rater attitudes towards Indian English as an example of WE are associated with scoring decisions in communicative-oriented English speaking tests. Although the standards the tests adopt depends on their purpose, an increasing number of large-scale English proficiency tests for study abroad, immigration, or work include the use of more than one variety of English in the test input to reflect authentic communicative situations, particularly in the inner circle countries (Kachru 1992) including Australia, Canada, New Zealand, the USA, and the UK, where multiple varieties of English are common on many fronts. The standards for the evaluation of speaking output vary. TOEFL, for example, uses norms for native speakers except when judging advanced examinees. In the latter case, “the highest performance levels described in the writing or speaking scoring rubrics emphasize overall effectiveness of the written or spoken performance rather than native-like performance” (Xi and Mollau 2011, p. 1223). On the other hand, in IELTS, “all standard varieties of English are accepted in examinees’ written and spoken responses” (IELTS information for candidate).

Raters now face challenges in handling varieties of speech in communicative-oriented speaking tests. Changes are needed in the assessment practices given the growing recognition of varieties of Englishes; as such, the extent to which rater's psychological traits influence their scoring decisions requires fresh investigation.

Literature review

The field of psychology and language attitude

A possible definition of "attitude" in the context of our discussion is: "consciously held ways or beliefs about a specific language or an orientation (positive or negative) towards a specific language that influences the individual's evaluation of that language and its speaker" (Cluver 2000, p. 315). Scholars have commonly assumed that attitudes about language are the result of a process stretching over many years, and possibly are even unchangeable after the passing of generations. Complex factors may affect the formation of language attitude, such as a listener's experience and education in "a virtually endless, recursive fashion" (Cargile et al. 1994, p. 215). The "social process model of language attitude" (Cargile et al. 1994) further explains this formation. The model includes factors such as the characteristics of the speaker and listener, and contextual factors such as how a listener perceives a speaker's culture, social situation, and interpersonal history. The model also traces the interactions between the speaker's linguistic features and the listener's characteristics (e.g., expertise and social identity), particularly when listeners are "actively involved in selecting and attending to those language behaviors around which they construct their attitudes and evaluation" (Cargile et al. 1994, p.218).

A negative attitude towards persons may lead to negative evaluations of their behavior. Psychologists are well aware of attitude-behavior relationships (Ajzen and Timko 1986; Albarracin et al. 2005; Fazio et al. 1989; Hrubes et al. 2001). One respected theory supporting links between attitude and behavior is the "Theory of Reasoned Action" (Ajzen and Fishbein 1980). This theory suggests that attitudes about a certain behavior, common perceptions related to that behavior, and an individual's perceived behavioral control about that behavior all influence how evaluators make judgments. The current literature on language attitude shows that a negative attitude towards non-standard speakers draws unfavorable evaluations of the speakers' competency. Note that terms, "standard" and "non-standard" varieties/speakers used in the attitude studies most likely refer to the inner circle varieties/speakers and non-inner circle varieties/speakers, respectively, in relation to WE.

Rubin (1992) found that when an American English speech on tape was playing, while listeners were facing a photograph with an Asian instructor's face, they said that they were listening to a non-standard speech. Identification of the instructor as Asian appeared to undermine listeners' comprehension. Listeners' perception of a speaker's/instructor's accent as foreign undermined their evaluation (Rubin 1992). Moreover, Lindemann (2002) observed that the attitudes of native speakers towards Koreans determine whether their actual and perceived interaction with Koreans is successful. Choices of communication strategies by native speakers of English, including "avoidance" and "problematizing strategies" (p. 433), seem to influence these interactions. The avoidance strategy meant refusal to offer feedback to partners when difficulties arose in understanding, while the problematizing strategy meant not acknowledging Korean partners' contributions to the communication. Furthermore, in a broad review of issues, Giles and Billings (2004) surmise that speakers of standard variety are typically considered superior in regard to confidence, intelligence, and ambition. In contrast, speakers of non-standard varieties are generally perceived favorably on qualities such as friendliness and honesty, particularly when the listeners and speakers share the same non-standard variety. Similarly, empirical studies suggest that in the contexts of education, law, and health, speakers of non-standard variety may be judged as less educated or competent than speakers of standard varieties (Garrett 2010).

Attitude-behavior relationship in language testing

An increasing number of studies place stakeholders' perceptions of WE as a core research concern. While IELTS examinees (Hamid 2014) and various TOEFL stakeholders (Gu and So 2014) indicate support of WE, their attitude may be context- and issue-dependent. The majority of participants in both studies have expressed reservations about the inclusion of WE in the testing situation, particularly in regard to the inclusion of accents and written conventions. Hamid (2014) concludes that this mixed attitude towards WE is a result of a linguistic hierarchy and misunderstanding of WE as an unstable language form, probably "informed by social and linguistic prejudices" (p. 273).

Empirical studies that look into stakeholders' attitudes towards WE and their subsequent behavior tendencies, such as raters' scoring and examinees' performance, present mixed findings (Harding 2008; Kim 2005; Kang 2008; Hsu THL: The impact of World Englishes on language assessment: perception, rating behavior, and challenges, Unpublished). The research does not indicate, however, that views about WE necessarily influence test scores. The contexts of the studies and perhaps inconsistencies in the investigation of rater or examinee attitude may have affected the results of speaking and listening test. Kim's study (Kim 2005) focuses on how rater attitudes towards WE relate to ratings for the speech performances of Korean students, using holistic and analytic scales. The study categorized groups of teacher raters with different language backgrounds into different attitude groups: positive, neutral, or negative, according to the findings of a questionnaire that raters completed. The result showed that raters, despite different language backgrounds, had similar rating performances on their holistic ratings. However, attitudes that raters held towards WE significantly affected their analytical scoring on criteria such as grammar, rate of speech, and task fulfillment. As noted, raters labeled as "positive" raters were more lenient in their ratings.

Focusing on examinees, Harding (2008) looked into the use of the accents of WE speakers (i.e., Chinese, Japanese, and Australian) in an academic listening test and the extent to which it affected examinees' test results. The findings showed that examinees generally displayed positive attitudes towards accented WE speakers. Nevertheless, the attitude-behavior relationship, as claimed by the psychologists, was not an issue in this study. Examinees' attitudes towards WE speakers did not associate with examinees' performance in a listening test in which the WE speakers' voices were used.

A relevant attitude study was carried out by Kang (2008), although attitude towards ethnic group was being investigated instead of the variety of English the group of people speaks. Following Rubin (1992), Kang (2008) focused on college student rater attitudes towards two ethnicities, Asian and Caucasian. While there was no significant effect on rater attitudes towards ethnicity, NNS student raters consistently appeared to give international teaching assistants (ITAs) lower ratings than NS on all rating categories, including communication skills, expression of ideas, and overall proficiency. To improve student raters' attitude towards ITAs and evaluation of ITAs, Kang intervened in an attempt to create interaction between student raters and ITAs, leading to a change in student raters' attitudes and their subsequent evaluation of ITAs. She concluded: "informal and pleasant contact with interpersonal intimacy and equality can bring a positive change in undergraduate attitudes toward ITAs and consequently influence undergraduates' perceptions of ITA speech performances.. ." (p. 200).

The review of relevant attitude studies on language testing shows a firm conclusion that a listener's attitude and evaluation of a speaker's variety, or competency, cannot be drawn due to different listeners (e.g., rater, student, examinee) used as participants, and a lack of consistent measurement tool to evaluate a listener's attitude towards a speaker's variety. Although the review shows that examinees' attitudes towards WE are generally positive (Hamid 2014; Harding 2008; Gu and So 2014), and their test performance is not associated with their attitudes, raters seem to differ from this preliminary finding (Kim 2005). As the rater is a decisive factor in speaking test scores, the review indicates a pressing need to undertake more empirical studies to look into raters' psychological traits and evaluate the extent to which these traits constitute a potential rater biasing factor.

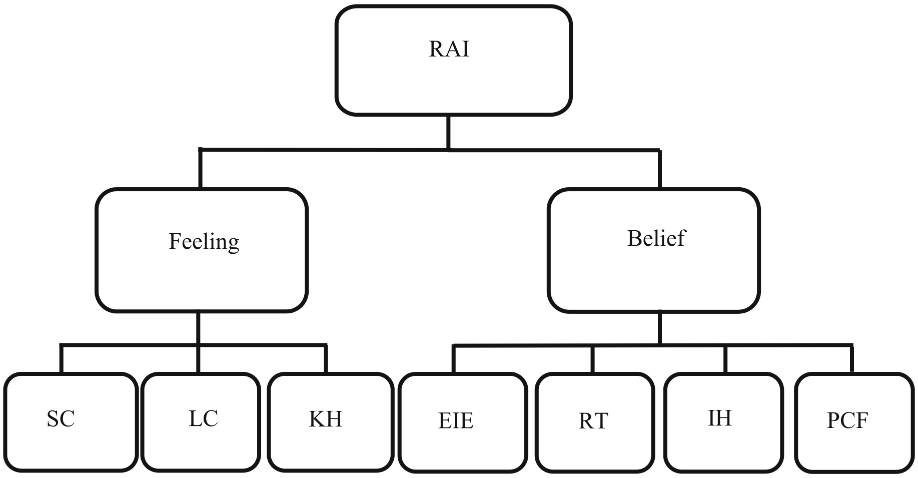
Overall, the emerging new line of inquiry into stakeholders' attitudes towards WE takes a step further to bring not only psychometric inquiry but also a wider social context into LT research agenda (McNamara and Roever, 2006) and touches on issues about the intended and unintended consequences of the use of WE in English speaking tests, thereby placing the social dimension in the core of post-Messick (1989) validity inquiry. Placing the perspective on the attitudes of raters or examinees towards WE as part of the agenda for research differs from the tradition of exploring examinee group differences, which makes the inquiry more socially responsive.

Rater attitude instrument in language testing

The rater attitude instrument (RAI) has been discussed in detail in Hsu (2016). This section will provide a brief summary of the RAI and highlight the aspects that are relevant to the current study. The development of the RAI was informed by the three types of attitude constructs, identified by psychologists: affective, cognitive, and behavioral (Ajzen and Fishbein 1980; Albarracin et al. 2005; Cargile et al. 1994). "Affect" refers to feelings. "Cognizance" refers to an individual's belief structure. "Behavior" is the result of certain tendencies in one's personality. Confirmatory factor analysis was performed in the RAI development stages and supported a two-factor internal structure of the RAI with acceptable model fit indices ($\chi 2 = 20.052$, p = .094, RMSEA = 0.076, CFI = 0.954, TLI = 0.926). The two factors are labeled as "rater feeling" and "rater belief." Rater feeing includes three components: speech competency, intelligence, and kind-heartedness, which are measured in 11 7-point semantic differential scale items. Rater belief covers four elements: (1) expectation of Indian English, (2) perceived cultural factor, (3) interpersonal history, and (4) rating tendency; these are measured in the 31 5-point Likert scale items. The author also included a third section composed of rater background items, such as native language and Indian vs. non-Indian status. Figure 1 shows the RAI measurement model.

Sample items in the RAI are as follows. An item from "perceived cultural factor" in rater belief asks if "standard English (e.g., American English) should be used to judge examinees' performance in the test setting." Raters respond to each statement by selecting one of five responses: strongly agree, generally agree, neutral, generally disagree, and strongly disagree. Furthermore, in the rater feeling, contrasting adjectives, or adjective phrases are placed at each end of the scale. An example from the kind-heartedness category is "The speaker sounds.. .." Raters indicated their position on the scale between "considerate" and "inconsiderate." The RAI composite score was calculated and the severity of rater attitude was examined in FACETS, which classified raters into three attitude groups: positive, neutral, and negative.

The development of the RAI, also a validation process, makes it different from the measurement tools used in attitude studies, as reviewed in the previous section (Harding 2008; Kang 2008; Kim 2005); the latter mainly used questionnaire and the speech evaluation instrument (SEI) (Zahn and Hopper 1985) as the major tools to explore participants' attitudes. Unlike the RAI that is built upon theoretical rationale in the fields of LT, WE, and social psychology, the development of the questionnaire (Kim 2005) and adaption/use of the SEI (Harding 2008; Kang 2008) seem to lack elaboration on how the measurement tool is situated within the context of LT. On the other hand, the RAI includes two different measurement tools: semantic differential scale and Likert scale to elicit raters' immediate feeling towards Indian speakers and relatively steady belief in WE, including their experience in interacting with WE speakers and rating tendency when assessing WE speakers in speaking tests. Furthermore, the RAI used teachers as participants in all the development stages to increase the validity of the findings, as opposed to student participants commonly used in many attitude studies.



SC=speech competency, LC=level of confidence, KH=kind-heartedness,
EI= expectation of Indian English, RT=rating tendency, IH=interpersonal history,
PCF=perceived cultural factor

Fig. 1 Measurement structure of the RAI

As such, the validated RAI is a comprehensive measurement tool that not only tailors the LT situation, but also links different disciplines (i.e., LT, WE, and social psychology) to unfold the complexity of rater attitude towards WE.

Research questions

The main objective of the current study is to examine the extent to which rater attitudes towards Indian English as an example of WE, according to RAI measurement, are associated with the scoring of the IELTS speech samples of the Indian examinees. The present study will address the following questions:

1. To what extent is rater attitudes towards Indian English associated with IELTS speech sample score?
2. To what extent can the RAI scores help predict scoring on the IELTS speech samples?

Methodology

Participants

To reach out to potential participants, the study author contacted members of TESOL organizations and directors of ESL programs affiliated with universities in India and metropolitan cities in the USA, including New York and San Francisco.

The rationale for contacting this group of potential participants was their possibly greater exposure to multiple varieties of English, and higher sensitivity to, and easier understanding of, second language speech (Saito and Shintani, 2016). Email invitations, approved by the Institutional Review Board of the author's university by the time the study was conducted, were sent out to 150 potential participants. The email explained that the purpose of the study was to elicit opinions about a variety of English as heard during rating situations but, to avoid socially desirable responses (Steenkamp et al. 2010), without explicitly saying it was an attitude study. A total of 96 teachers participated in this study: 13 Indians and 83 US-based non-Indians, of whom 90% were Caucasian, 8% Asian, and 2% African-American. Nearly one third of the teachers ($N = 23$) reported having rating experience in operational large-scale English proficiency tests. The majority (75%) of the teachers held a Master's degree in TESOL. All the teachers had experience with Indian English in non-test situations, with 63.9% of the teachers indicating they had no problem in understanding Indian English speakers. Each received $25 for participation.

Speech stimulus and procedure

IELTS

IELTS was chosen because of its explicit statement about encouraging varieties of English to responds to test tasks, including the speaking section. Its exam handbook states:

> IELTS is internationally focused in its content.. . a range of native-speaker accents (North American, Australian. .. etc.) are used in the listening test; and all standard varieties of English are accepted in test takers' written and spoken

> responses. (IELTS Information for Candidate, *https://www.britishcouncil.hu/sites/default/files/ielts_information_for_candidates_0.pdf*).

As the IELTS research notes, the purpose of varieties of English in the speaking and writing tests is to "enable candidates to function in the widest range of international contexts" (Taylor 2002, p. 20). This appears to be a reflection of greater authenticity of English use in the international context.

Cambridge English Language Assessment granted access to IELTS speech samples. To control for extraneous variables, the author extracted six IELTS descriptive tasks (i.e., part two of IELTS speaking section) that Indian examinees completed from actual IELTS speaking data. The descriptive task requires examinees to provide descriptions on particular topics. The task topics in this study included "describe an elderly person you know," "describe something useful that you have recently learned," "describe a sports event you watched at a party," etc. The six IELTS descriptive tasks covered a range of IELTS band scores, including band four to nine. Each descriptive task was approximately 90 s long.

Indian English is the sole variety serving as a stimulus in this study because of its wide use in WE research (Bhatt 2001; Kachru 1992, 2001; McArthur 2003); it continues to receive attention in recent empirical studies where it is the stimulus for Indian raters' scoring performance (2016; Xi and Mollau 2011).

Scoring procedure

The raters received an URL address to access the study materials. The materials included the RAI, six IELTS speech samples, instructions for completing the RAI, rating categories for the IELTS samples, and a consent form. The raters first listened to an IELTS speech sample and scored it according to four IELTS rating criteria: fluency, pronunciation, sentence structure, and vocabulary. They then repeated this step for the remaining five IELTS samples. Upon completion of scoring, the raters proceeded to the RAI.

Rater attitude instrument

This study uses Hsu's (2016) RAI to elicit rater attitude towards Indian English as an example of WE. After scoring the IELTS samples, the raters completed part two of the RAI, rater belief, the five-Likert scale items. Then the raters listened to each IELTS speech sample again to indicate their feelings about each speaker in the first part of the RAI, rater feeling on a seven-point semantic differential scale. Additionally, the raters completed part three of the RAI, rater background information. Such a scoring procedure to separate scoring of the English proficiency level of Indian speakers and the raters' feelings towards speakers was informed by the pilot study results. The participants indicated that scoring two tasks simultaneously would affect their judgment in either task. As such, the two scoring tasks separated the second part of the RAI, rater belief in WE from the first part.

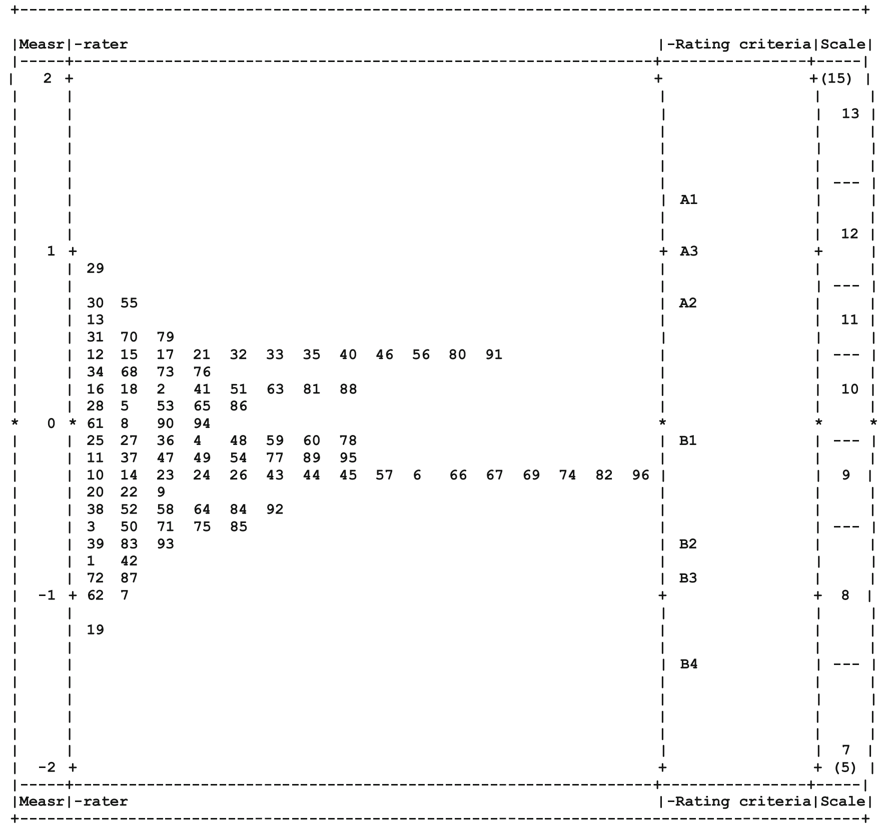
Finally, due to two different measurements, the author adjusted the scores to generate proportional scores for comparison on a like basis. The RAI composite score was a calculation of the sum total of scores from the two parts.

Results

Research question 1: To what extent is rater attitude towards Indian English associated with IELTS speech sample scores?

Rater attitude group

A one-factor multivariate analysis of variance (MANOVA) determined how rater attitude towards Indian English relates to the four analytical rating scores. Prior to the MANOVA analysis, the raters needed to be in different attitude groups to serve as independent variables. Unlike the attitude studies that typify group participants according to the score ranking of attitude measurements, this study used FACETS analysis (Linacre 1989) to examine rater attitude's severity levels, placing them in different attitude groups. The study employed a two-faceted design, modeling raters, and difficulty of RAI components. The latter includes the seven subscales of the RAI: three factors representing the rater feeling (i.e., speaking competency, kind-heartedness, and level of confidence) and the four elements for rater belief (i.e., perceived cultural factor, expectations of Indian English, rating tendency, and interpersonal history). The examinee speaking proficiency was the controlled variable, and not a factor in the measurement model. The computer program FACETS handled the analyses (Linacre 1989).



```
+---------------------------------------------------------------------------------------------+

|Measr|-rater                                                          |-Rating criteria|Scale|
|-----+----------------------------------------------------------------+----------------+-----|
|   2 +                                                                +                +(15) |
|     |                                                                |                |     |
|     |                                                                |                | 13  |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                | --- |
|     |                                                                | A1             |     |
|     |                                                                |                |     |
|     |                                                                |                | 12  |
|   1 +                                                                + A3             +     |
|     | 29                                                             |                |     |
|     |                                                                |                | --- |
|     | 30  55                                                         | A2             |     |
|     | 13                                                             |                | 11  |
|     | 31  70  79                                                     |                |     |
|     | 12  15  17  21  32  33  35  40  46  56  80  91                 |                | --- |
|     | 34  68  73  76                                                 |                |     |
|     | 16  18  2   41  51  63  81  88                                 |                | 10  |
|     | 28  5   53  65  86                                             |                |     |
*   0 * 61  8   90  94                                                 *                *     *
|     | 25  27  36  4   48  59  60  78                                 | B1             | --- |
|     | 11  37  47  49  54  77  89  95                                 |                |     |
|     | 10  14  23  24  26  43  44  45  57  6   66  67  69  74  82  96 |                |  9  |
|     | 20  22  9                                                      |                |     |
|     | 38  52  58  64  84  92                                         |                |     |
|     | 3   50  71  75  85                                             |                | --- |
|     | 39  83  93                                                     | B2             |     |
|     | 1   42                                                         |                |     |
|     | 72  87                                                         | B3             |     |
|  -1 + 62  7                                                          +                +  8  |
|     |                                                                |                |     |
|     | 19                                                             |                |     |
|     |                                                                |                |     |
|     |                                                                | B4             | --- |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |  7  |
|  -2 +                                                                +                + (5) |
|-----+----------------------------------------------------------------+----------------+-----|
|Measr|-rater                                                          |-Rating criteria|Scale|
+---------------------------------------------------------------------------------------------+
```

A1= speaking competency, A2= kind-heartedness, A3= level of confidence, B1= expectation of Indian English, B2= rating tendency, B3= interpersonal history, B4= perceived cultural factor

Fig. 2 FACETS summary
]

Figure 2 shows the relative *severity* of the raters and difficulty of the RAI subscales. The first column is the logit scale, which is the unit of measurement in Rasch analysis. The far right column shows the scale of the scoring used. The second column shows the severity variation among raters. A measure of 0 represents the average severity for rater performance. A rater scoring most severe, which may indicate a negative attitude towards Indian English as an example of WE, is at the top of the logit scale, and most lenient suggests a positive attitude, at the bottom. The raters' logit values extend from + .89 (rater 29) to − 1.22 (rater 19), a range of 2.11 logit. We can determine the extent to which the 2.11 logit is meaningful by checking the fixed (all same) chi-square which FACETS analysis provides. The fixed chi-square tests the null hypothesis that all the elements of the facet are equal. For the current data set, the chi-square of 197.4 with 95 df is significant at $p = 0$, indicating a rejection of the hypothesis; thus, the raters were not equally severe. The placement of the relative attitude standing of raters into groups according to their logit values is possible. Raters who had positive logits were classified into the group "negative attitude" ($N = 36$, 37.5%), negative logits to "positive attitude" ($N = 56$, 58.3%), and zero logit to "neutral attitude" ($N = 4$, 4.2%).

IELTS speech sample scores

The author calculated the inter-rater reliability of IELTS speech sample scorings. As reported in Table 1, Cronbach's alpha displayed an acceptable to high level of internal consistency for rater performance (i.e., above 0.6) (Nunnally 1978), except for one case. Alpha for pronunciation and fluency in the neutral group ($N = 4$) was low (i.e., .526 and − .017). The low Cronbach's alpha for the current data set may derive from several causes: (1) the small sample size in the neutral group, and (2) the fact that raters' scoring decisions on categories of pronunciation and fluency differed considerably, which caused the variability of the individual raters to exceed their shared variance (Henson 2001). Nevertheless, the author kept the grouping results given that alpha is not a perfect tool for measuring as it may underestimate the reliability of multidimensional scales (Shrout and Yager 1989). Furthermore, scholars argue that an evaluation of internal reliability also needs a consideration of the overall results of the analysis (Streiner and Norman 2000).

The author administered *MANOVA* to achieve a greater understanding of how rater attitude groups affect variability in their ratings of the IELTS speech samples. The attitude groups of raters are the independent variables, and the rating criteria are the dependent variables. Table 2 presents the mean scores and standard deviation of the four dependent variables for the three levels of the independent variables. The scores range from 0 to 9, as per the IELTS band scale. Note that Indian and US-based raters are all distributed in the three attitude groups (i.e., positive attitude group: 7 Indians

Table 1 Inter-rater reliability

	Positive attitude group	Neutral attitude group	Negative attitude group
Fluency	.825	.526	.930
Pronunciation	.674	− .017	.863
Sentence structure	.810	.733	.885
Vocabulary	.829	.892	.886

Table 2 Mean and standard deviation for proficiency variables by three attitude groups of raters

		Positive			Neutral			Negative	
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
	(6 × 56)			(6 × 4)			(6 × 36)		
Fluency	336	7.26	.16	24	5.92	.71	216	6.10	.31
Pronunciation	336	7.26	.16	24	5.96	.69	216	5.54	.31
Sentence structure	336	6.81	.16	24	6.67	.53	216	5.89	.32
Vocabulary	336	6.88	.17	24	6.50	.51	216	5.81	.32

N is the number of ratings completed by the raters (i.e., six speech samples × the number of raters in each attitude group)

and 49 US-based raters; neutral attitude group: 2 Indians and 2 US-based raters; and negative attitude group: 4 Indians and 32 US-based raters). An examination of these means revealed that the positive attitude group consistently rated the IELTS speech samples the highest among the three attitude groups, followed by neutral and negative attitude group on all rating categories. The negative attitude group consistently gave the lowest ratings across all categories, except for fluency, which the neutral attitude group rated lowest.

The rating for each category by the three attitude groups was further examined to evaluate the rater's rating tendency. In addition, the rating by the 96 raters in this study was compared with each speech sample's official band score to check the extent to which the rating in the current study differs from the official band score. Note that the author of this study was only able to obtain each speech sample's holistic band score, an average score of the three speaking tasks, from Cambridge Assessment English. Each speech sample's analytical scores were not available. Table 3 summarizes the mean and standard deviation for each speech sample.

All the ratings awarded to the six speech samples were generally consistent with the rank order of IELTS official scores. Except for the speech sample with a band score of 7, the higher the sample's official score, the higher the ratings given by the three groups of raters. This result reveals that raters in this study, although untrained, demonstrate a rating tendency similar to trained IELTS raters, which also provides confidence in the

Table 3 Mean and standard deviation for all rating categories

IELTS speech sample band score	Fluency			Pronunciation			Sentence structure			Vocabulary		
	P	Neu	Neg	P	Neu	Neg	P	Neu	Neg	P	Neu	Neg
4	6.00 (.173)	3.25 (.854)	4.81 (.298)	6.00 (.173)	3.75 (1.03)	4.08 (.325)	5.64 (.175)	5.25 (.479)	4.58 (.348)	5.64 (.175)	5.00 (.408)	4.06 (.331)
5	6.43 (.219)	4.00 (1.10)	5.08 (.327)	6.43 (.219)	3.25 (1.03)	4.11 (.340)	5.57 (.208)	4.50 (.500)	4.50 (.317)	5.57 (.208)	5.00 (.00)	5.06 (.333)
6	7.48 (.189)	6.75 (.479)	6.36 (.326)	7.48 (.189)	7.75 (.479)	5.81 (.313)	7.27 (.177)	7.75 (.479)	6.19 (.335)	7.27 (.177)	7.25 (.750)	6.22 (.382)
7	6.56 (.199)	5.25 (.946)	5.03 (.353)	6.56 (.199)	4.74 (.629)	4.36 (.336)	5.85 (.197)	5.75 (.750)	4.44 (.373)	6.11 (.226)	5.75 (.479)	4.33 (.361)
8	8.45 (.102)	8.25 (.479)	7.03 (.289)	8.45 (.102)	7.75 (.479)	6.81 (.281)	8.05 (.128)	8.25 (.479)	7.22 (.290)	8.08 (.135)	8.00 (.707)	7.00 (.285)
9	8.65 (.093)	8.00 (.408)	7.61 (.274)	8.65 (.093)	8.50 (.500)	7.47 (.294)	8.45 (.098)	8.50 (.500)	7.78 (.236)	8.59 (.091)	8.00 (.707)	7.53 (.216)

P stands for positive attitude group, *Neu* for neutral attitude group, and *Neg* for negative attitude group credibility of the current data. For the sample with a band score of 7, the score discrepancy between rating awarded by the raters in this study and the IELTS official score is probably because this examinee did not do well in the descriptive task, but did better in other two tasks, which led to a higher holistic band score.

Furthermore, raters in the positive attitude group consistently gave higher mean scores than the negative attitude group did in all four rating categories. Compared to the negative attitude group, the neutral attitude group in most cases rated higher, except for speech samples with lower band score (e.g. 4, 5, and 6) in which the neutral attitude group gave lower mean scores than the negative attitude group did in all rating categories.

The MANOVA revealed that the main effect for the group variable was significant (lambda = .866). It showed that speaking test scores of the Indian examinees in this study depended very much upon the group their rater belonged to. The tests of between-subject effects showed that rater attitude towards Indian English as an example of WE had a significant effect on all four dependent variables: fluency (F (2, 573) =29.194; $p < .0005$; partial eta squared = .092), pronunciation (F (2, 573) = 8.268; $p < .0005$; partial eta squared = .028), sentence structure (F (2, 573) = 16.327; $p < .0005$; partial eta squared = .054), and vocabulary (F (2, 573) = 30.918; $p < .0005$; partial eta squared = .097).

Post hoc analysis of means by using Tukey contrasts tested for mean differences among the positive, neutral, and negative attitude groups of raters (see Table 4 for the results of the Tukey tests). Differences emerged for all variables between the positive and negative attitude groups. The attitude that the positive group held towards Indian English obviously resulted in significant mean differences in fluency, pronunciation, sentence structure, and vocabulary from the negative attitude group. Raters with a positive attitude towards Indian English provided higher mean scores than the negative attitude groups did. Mean scores on sentence structure and vocabulary were statistically different between the neutral and negative attitude groups. Raters in the neutral attitude group gave higher mean scores than the negative attitude group did.

To sum up, the findings on research question one reveal an important and potential rater biasing factor: the raters' attitude towards Indian English as an example of WE appears to relate to their scoring decision on IELTS speech samples scores. MANOVA further confirmed that the main group effect was significant, indicating that the IELTS speech sample scores depend on which attitude group of raters rated the sample. The current data show significant differences between positive and negative attitude groups. The mean scores of all of the four rating categories in the positive attitude group are

Table 4 Tukey multiple comparisons of four analytic scores awarded by different attitude groups

Rating criteria	Attitude	group	Mean difference	Std. error	Sig
Fluency	Positive	Negative	1.26*	.165	.000
Pronunciation	Positive	Negative	.75*	.193	.000
Sentence structure	Positive	Negative	.97*	.316	.000
	Neutral	Negative	.93*	.316	.010
Vocabulary	Positive	Negative	1.37*	.176	.000
	Neutral	Negative	1.07*	.319	.002

*The mean difference is significant at the.05 level consistently higher than those of the negative attitude group. Moreover, significant differences in mean ratings were found between the neutral and negative attitude group on categories of sentence structure and vocabulary. The neutral group gives higher mean score ratings on these two categories.

Research question 2: To what extent can the RAI scores help predict scoring on IELTS speech samples?

Correlation analysis and multiple regression analysis were combined to address this question. Table 5 reports the correlation between scores for IELTS speech samples (i.e., total and four subscores) as dependent variables, and RAI scores (total and part scores) as well as rater background characteristics as criteria variables. Note that the point-biserial correlation was used for rater background variables due to their dichotomous nature (e.g., gender). The IELTS speech sample total scores and four subscores show a significant connection with the RAI total score and part score one (i.e., rater feeling), ranging from .418 to .560 ($p < .01$) and .272 to .556 ($p < .01$). Following Plonsky and Oswald (2014), the strength of association in score relations appears medium, except for the correlation of pronunciation and RAI part score one ($r = .272$), which was weak. The RAI part score two (i.e., rater belief) was significantly linked only with the IELTS speech sample total scores ($r = .225$, $p < .05$) and pronunciation ($r = .317$, $p < .01$), featuring very weak to weak associations. As for rater background variables, only the Indian/non-Indian variable was significantly related to the IELTS speech sample total score ($r = -.252$, $p < .05$), sentence structure ($r = -.329$), and vocabulary ($r = -.303$), featuring weak associations. The negative correlation suggests that low scores are associated with high group membership (i.e., Indian was coded 1 and non-Indian, 0). The coding is based on the hypothesis that Indian raters rated higher to the Indian speech samples. Thus, as group membership increases, the IELTS speech sample scores decrease. In other words, Indian raters in the current study rated lower on the IELTS speech samples than those of non-Indian raters. Other rater background variables were non-significant.

It was important to examine how much of the variance of IELTS speech sample scores is due to RAI scores and rater background variables. The author performed regression analyses using stepwise methods. Prior to the analysis, Box's test was performed to check whether the assumption of homogeneity of covariance across the two

Table 5 Correlations among IELTS tasks scores, attitude scores, and background variables

Predictors	IELTS sample total scores	FLU	PRON	SS	VOC
RAI total score	.560**	.534**	.418**	.470**	.569**
RAI part score one	.498**	.508**	.272**	.422**	.556**
RAI part score two	.225*	.168	.317**	.177	.159
RAI rating tendency	.206	.125	.233	.236	.177
Indian/non-Indian	– .252*	– .192	– .063	– .329*	– .303*
Native language	.133	.128	.061	.164	.121
Gender	– .073	– .018	– .116	– .041	– .089
Teaching experience	– .128	– .137	.000	– .123	– .180
Education level	.002	– .056	– .021	.089	– .003

RAI part score one rater feeling, *RAI part score two* rater belief, *FLU* fluency, *PRON* pronunciation, *SS* sentence structure, *VOC* vocabulary

*$p < .05$, **$p < .01$

groups (i.e., Indian and non-Indian raters) was met due to a large discrepancy between the number of Indian and non-Indian raters (i.e., 13 vs. 83). Box's *M* value of 14.700 was associated with a *p* value of .341, which was interpreted as non-significant based on Huberty and Petoskey's (2000) guideline (i.e., $p < . 005$). Thus, the covariance matrices between the Indian and non-Indian rater groups were assumed to be equal. Furthermore, each regression analysis used one IELTS speech sample score, a total or sub, as a dependent variable. The author performed five regression analyses (see Table 6 below). Insignificant variables do not appear in the IELTS speech sample scores.

The findings show that the RAI total score was the strongest predictor of IELTS speech sample total score, accounting for 31.3% of its variance. The Indian/non-Indian variable was also a significant predictor, despite its small contribution (3.2%).

Breaking IELTS speech sample total scores into four subscores shows that the strongest predictor for all the subscores was the RAI total score. Its variance ranged from 17.5% for pronunciation, 22.1% for sentence structure, and 28.5% for fluency, to 32.4% for vocabulary. The second predictor for the IELTS speech sample subscores varied. For fluency and pronunciation, no second predictor appeared significant at the.05 alpha level. For sentence structure and vocabulary, the second predictor was the Indian/non--Indian variable, which contributed significantly to the 7.2% and 5.2% of the total variance, respectively, with relatively small contributions.

To sum up, the findings on research question 2 provide some evidence for the importance of the power of RAI in predicting the scoring tendency of the IELTS speech samples. The RAI total score was found to be a significant predictor of the IELTS speech sample total scores and all four of the IELTS subscores. Furthermore, the Indian and non-Indian variable also contributed significantly to the variance in IELTS total scores and some of the IELTS subscores, though its contributions are relatively small.

Table 6 Summary results of multiple regressions for rater attitude towards Indian English and background variables predicting ratings of IELTS speech samples

	R	R2	R2 change	Standardized beta	*F* change
IELTS speech sample total score					
RAI total score	.560	.313	.313	.536	42.883
Indian/non-Indian	.587	.345	.032	− .180	4.511
IELTS speech sample subscores					
Fluency					
RAI total score	.534	.285	.285	.534	37.469
Pronunciation					
RAI total score	.418	.175	.175	.418	19.946
Sentence structure					
RAI total score	.470	.221	.221	.433	29.596
Indian/non-Indian	.582	.293	.072	− .271	9.475
Vocabulary					
RAI total score	.569	.324	.324	.538	45.087
Indian/non-Indian	.613	.376	.052	− .230	7.773

Discussion and conclusion

This study investigated the impact of rater attitude towards Indian English, as an example of WE, on scoring for IELTS descriptive tasks. Our findings suggest raters with a positive attitude towards Indian English, according to the RAI, give higher mean scores than the negative attitude groups do in all rating criteria. A comparison of the neutral and negative attitude groups shows differences in the mean score. The category of sentence structure and vocabulary, with the neutral attitude group giving higher mean scores than the negative attitude group did, suggests that raters' attitude towards Indian English generally relates to their scoring tendency. However, there is less or no effect on scoring fluency and pronunciation. Future research could explore underlying factors in the way raters score fluency and pronunciation through, for example, a verbal protocol report. The overall findings support claims put forward in psychology and language attitude research. A link between rater attitude towards Indian English and scoring tendency for Indian examinees may exist in a language assessment context. Thus, as raters reoriented their views and broadened their grasp of WE, and as awareness of WE increased in the language testing community in recent decades, the findings here show that testing agencies must add an understanding of potential rater bias towards WE to the current relevant literature. We cannot afford to overlook, or be blind to, the possibility of prejudiced or otherwise unfair scoring of speaking tests.

The RAI total score was the strongest predictor of IELTS total and subscores. The total variance of the RAI total score ranged from 17.5% in the pronunciation score to 32.4% for vocabulary. This included quite a surprisingly high proportion of variance (31.3%) in the IELTS descriptive task total scores. Therefore, the RAI appears to be an important tool for monitoring and predicting rater attitude tendency prior to actual rating. Ideally, rater attitude should be neutral to avoid giving scores that are either too lenient, as by raters with positive attitude, or too harsh, as by raters with negative attitude. Rater training models may factor in this potential rater bias variable to intervene to reduce prejudice among raters, and to encourage overall neutrality in attitude. Two approaches suggest themselves: first, introducing raters to state-of-art English pedagogical development driven by WE to increase their understanding of the concept of WE; secondly, bringing WE speakers into rater training sessions to increase raters' familiarity with the varieties of English among examinees. As the literature suggests, long-lasting stereotypes about non-native varieties are negative. Similarly, as regards English language teaching, learning, and assessment, Elder and Harding (2008) claim a widely held perception among stakeholders that standard English alone is the appropriate norm and more prestigious for teaching and testing. Nevertheless, a growing number of proposals and initiatives have called for WE-informed English curricula in recent years in high school (Lee 2012), undergraduate, graduate (D' Angelo 2012), and ESL programs (Kubota 2001; Villarreal et al. 2014), to academic WE-driven English programs at Osaka University, Japan (as cited in Matsuda and Friedrich 2011) and Chukyo University, Japan (Sharfian and Marlina 2012). In these cases, the English curriculum emphasizes preparing students to handle interactions in international contexts where English is used to communicate with either native or non-native speakers from other L1s. Towards this aim, the curriculum attempts to foster students' awareness of English language changes and contacts in a worldwide context through coursework, international exchange programs, selection of teaching materials, and topics that reflect relevant social contexts in which students use English. The materials and topics include L2-L2 interactions instead of the traditional predominant materials to interact with, and mimic, native speakers only, as well as international staff recruitment and training workshops for teachers (McKay 2012). Moreover, students' L1, local and multi-cultures are valued, as opposed to treating their L1 as interference in reaching the ultimate English learning goal, to increase students' confidence in introducing their culture to others. In addition, other scholars have suggested autonomous approaches in which students' diverse needs in using English in international contexts should be self-analyzed and tailored in the curriculum (Galloway and Rose 2015). Scholars have also urged the use of the Common European Framework of Reference (CEFR) (Council of Europe 2001) in an Asian context as a learning model because native speakers are not used as a benchmark (Kirkpatrick 2012).

Empirical studies show that WE-driven curricula have generated a positive impact on students' perception of WE. Students provide positive feedback on the new curriculum, and their perspectives about WE have also changed (Bayyurt and Altinmakas 2012; Lee 2012). Regarding a broader impact on policy decision, the newly proposed curriculum changed the program policy in a Turkish university, where graduates of the English undergraduate program need to demonstrate WE-relevant knowledge (Bayyurt and Altinmakas 2012). It is hoped that introducing raters to state-of-the-art pedagogical proposals and changes will increase their awareness of English curriculum initiatives and encourage a more neutral attitude towards WE.

Furthermore, actual interaction with WE speakers in training modules may facilitate a neutral attitude towards WE among raters. Drawing on the theory of uncertainty reduction (Berger and Bradac 1982), increasing interactions between strangers would reduce uncertainty about "the other." Studies show that direct, as opposed to indirect, experience or second-hand information strengthens the stability of attitudes (Eagly and Chaiken 1993; Fazio and Towles-Schwen 1999). Direct cross-cultural contact facilitates a listener's comprehension of speech (Derwing and Munro 1997; Field 2003; Kang 2008; Powers et al. 1999). WE speakers that raters are most likely to assess may actually participate in training sessions to increase the exposure of raters to speakers via designed activities, and to increase experience with WE, in contrast to more passive methods of training (Fulcher 2003; Luoma 2004; Taylor 2002).

In the examination of the rater background variables to predict scoring tendency, the results of correlation and multiple regression analyses show that the Indian/non-Indian variable accounted for only 3.2% of the total variance in the IELTS sample total scores. We should be cautious about interpreting the current data set. The Indian/non-Indian variable, though a statistically strong predictor at the .05 level, was not necessarily of greater practical significance (Krueger 2001). The small contribution of the Indian/non-Indian variable to the IELTS sample total scores indicates it plays a small role in the measurement of the IELTS sample total scores, or a spurious occurrence may have resulted. A more complete interpretation of the impact of the Indian/non-Indian variable on ratings of IELTS sample requires further investigation (Xi and Mollau 2011).

Furthermore, a more pressing need is for language assessment to agree upon what new norms should be used to assess the vast number of non-inner circle World Englishes examinees in English speaking tests. This is particularly important when the tests are no longer using the native speaker model to judge performance. This brings up the question of what the construct of L2 speaking is when it is communicative- oriented in international contexts, and how language assessment can benefit from WE and exert influences on English language teaching. In placing communicative efficiency as a core feature of international communication, we cannot ignore English as Lingua Franca (ELF) research. Although both WE and ELF explore the spread of English beyond its original contexts, and treat newly developed English expressions as innovation and identity representative, WE is concerned with relatively “linguistically identifiable, geographical definable” varieties of Englishes in inter- and intra-national communication (Kachru 1992, p. 67). ELF is more concerned with “fluid and flexible kinds of English use that transcend geographical boundaries” (Jenkins 2015, p. 42) and treats English as a contact language among speakers of different L1 s for communication in the international context, especially within relevant expanding circles. It is apparent that both lines of research complicate the construct of L2 speaking performance in the context of international communication, but help language assessment define L2 speaking with more accuracy in response to the evolving nature of English language. Scholars argue that L2 speaking should evaluate examinees’ use of their own varieties and their linguistic resources to satisfy communication needs in broader contexts (Canagarajah 2006). An examinee’s strategic competence, the ability to make effective use of communicative strategies, should therefore take precedence over linguistic accuracy (Elder and Davies 2006). Nevertheless, what constitutes ratable strategic competence in response to current English language development needs more operationalized definition before raters can assess such competence. In addition, a body of L2 intelligibility studies on segmental and supra-segmental features (Munro and Derwing 2006; Saito et al. 2017) that cause intelligibility problems provides compelling evidence to build up a part of the construct definition of L2 speaking performance. Recent empirical studies on the analysis of segmental categories in Hong Kong English (Sewell 2013), expanding circle varieties elicited in the Cambridge ESOL exam (Kang and Moran 2014) and in Indian English, all demonstrate that examinees’ segmental errors with low-functional load do not impede intelligibility, indicating that L2 examinees should not be penalized if their intended communicative goal is achieved, even though some of their speech features differ from standard English form. Additionally, Sewell’s (2013) analysis of segmental categories that focus on the intelligibility of Hong Kong English is aligned to the Lingua Franca Core (LFC) (Jenkins 2000). As such, across regions and LFC, more research should look into salient and overlapping linguistic and non-linguistic features, including phonology, lexis, grammar, and style. Scholars should also explore resemblances to help generalize the construct of L2 speaking in the global context.

As for rating, Elder and Davies (2006) propose that a WE (although they called it “ELF test”)-based test should include the reciprocal nature of communication. The point was to allow “group scores,” and “peer assessment” to elicit better judgment about a speaker’s competence to communicate successfully with different L1 speakers. Additionally, raters should be proficient users of English, regardless of their L1 background (Taylor 2005; Elder and Davies 2006), although empirical data show us that rater familiarity with examinees’ L1 is associated with the extent to which they perceive intelligibility (Browne and Fulcher 2017). Whether such claims actually are L2 speaking construct relevant and feasible in testing situations and help establish L2 speaking construct, it is apparent that this will be a complex task. It will require continual effort in language assessment to unfold what it really means when we speak of L2 speaking competency in international communication.

Finally, this study is guided by Messick's validity paradigm that re-conceptualizes validity as a unitary concept, looks at broader issues of the social dimension of the test, and seeks value and consequences of score interpretation and use. The current study looked into the wider context of global English use and found a potential rater biasing factor due to emerging WE. Nevertheless, the power of the RAI becomes crucial to predict rater's rating tendency in order to increase the fairness of the test results, linking validity and consequences of test results to address Messick's influential aspects of test validity. Similar to Kunnan's (2004) fairness framework put forward after Messick's validation framework, the current study treats fairness as important validity evidence prior to a testing event, rather than just an after-test quality. In the same vein, ETS standards for quality and fairness (Educational Testing Service 2014) highlight the role of rater and examinees' non-native status as essential data to address fairness, which places fairness investigation at the forefront as part of validation evidence. Finally, a thoughtful remark by McNamara (1996) re-states the crucial need to place the psychological traits of raters as an important research agenda to maximize the credibility of score use and interpretation:

> We must remain skeptical about the meaning of our test scores, and do everything we can to improve our understanding of what they mean, in the interests primarily of fairness to the test candidates, but also of the informativeness of our reports on candidates to test users (p. 246).

Language testing professionals and agencies should re-consider what it really means when we say English language speaking performance. We need new understandings to keep up with the change of English sociolinguistic landscape; professionals must keep in mind the potential threat to validity because of raters' attitudes towards WE and various intended and unintended consequences in English language learning and teaching.

Limitations

It is important to cite the limitations of this research. First, the current study did not involve rigorous rater training procedures; such training might have resulted in different findings. Follow-up studies that replicate the current study are therefore necessary in order to explore whether the rater bias observed in this study still has an effect even after a properly supervised rater training program. If rater bias still exists, the extent to which the 30% of the variance in test scores related to rater attitude towards WE will occur requires further examination. On the other hand, if rater bias is eliminated after a training program, intervention during rater training is essential, and could be a model for other L2 speaking tests. Second, Indian English speech is the only stimulus in this study, so the current findings may not apply to raters' attitudes towards examinees of other varieties. Extending the current study using alternative stimuli, such as other outer or expanding circle varieties, would offer insights into the generalizability of these findings. Third, this study used descriptive tasks only as elicitation stimulus of rater attitude and rating performance. Given that task types may affect test scores

(Chalhoub-Deville 1995; Wigglesworth 2001), the findings may have been different if speech tasks or a combination of task types were used. Fourth, the current study relies primarily on quantitative data. To further examine factors or reasons for scoring decisions by different attitude groups, a qualitative approach could be used in future studies, such as verbal protocol analysis. Last but not the least, raters completed the RAI and rating IELTS samples according to the suggested order (i.e., IELTS samples, part two of the RAI, and finally part one of the RAI). Nevertheless, given that all the tasks were done online, it is unknown whether raters followed this order. The deviation of the suggested order may result in an order effect, which affects the scoring results of any aforementioned task.

Abbreviations

ELF: English as lingua franca; RAI: Rater attitude instrument; WE: World Englishes

Acknowledgements

I would like to thank the two anonymous reviewers for Language Testing in Asia who made insightful and useful comments. I would also like to express my deepest gratitude to IELTS validation group in providing resources integral to the study and all ESL teachers and raters for participating in every stage of the study.

Funding

This work was funded by TOEFL small grants for doctoral research in second or foreign language assessment, Educational Testing Service, in 2010.

Availability of data and materials

Data and materials are available.

Author's contributions

The author confirms being the sole contributor of this work and approved it for publication.

Author's information

Dr. Tammy Huei-Lien Hsu is an assistant professor of English Literature and Language at Fu-Jen Catholic University, Taiwan.

Competing interests

The author declares that she has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 22 November 2018 Accepted: 28 February 2019
Published online: 20 March 2019

References

Abeywickrama, P. (2013). Why not non-native varieties of English as listening comprehension test input? *RELC Journal, 44*, 59–74.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs: Prentice Hall.

Ajzen, I., & Timko, C. (1986). Correspondence between health attitudes and behavior. *Basic and Applied Social Psychology, 7*(4), 259–276.

Albarracin, D., Johnson, B. T., & Zanna, M. P. (Eds.). (2005). *Handbook of attitudes*. Mahwah: Erlbaum.

Bayyurt, Y., & Altinmakas, D. (2012). A WE-based English communication skills course at a Turkish university. In A. Matsuda (Ed.), *Principles and practices of teaching English as an international language* (pp. 169–182). Bristol: Multilingual Matters.

Berger, C. R., & Bradac, J. J. (1982). *Language and social knowledge: Uncertainty in interpersonal relations*. London: Arnold.

Bhatt, R. M. (2001). World Englishes [Electronic version]. *Annual Review of Anthropology, 30*, 527–550.

Brown, J. D. (2014). The future of World Englishes in language testing. *Language Assessment Quarterly, 11*, 5–26.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 37–53). Bristol: Multilingual Matters.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242.

Cargile, A. C., Giles, H., Ryan, E. B., & Y Bradac, J. J. (1994). Language attitudes as a social process: A conceptual model and new directions. *Language & Communication, 14*, 211–236.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383–391.

Cluver, A. D. (2000). Changing language attitudes: The stigmatization of Khoekhoegowap in Namibia. *Language Problems and Language Planning, 24*(10), 77–100.

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

D' Angelo, J. (2012). WE-informed EIL curriculum at Chuko: Towards a functional, educated, multilingual outcome. In A. Matsuda (Ed.), *Principles and practices of teaching English as an international language* (pp. 121–136). Bristol: Multilingual Matters.

Davidson, F. (1993). Testing English across cultures: summary and comments. *World Englishes, 13*(1), 113–115.

Davidson, F. (2006). World Englishes and test construction. In B. Kachru, Y. Kachru & C. Nelson (Eds.), The handbook of world Englishes (pp. 709–717). Malden: Blackwell Publishes Ltd.

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22*(4), 571–584.

Derwing, T., & Munro, M. (1997). Accent, intelligibility and comprehensibility: evidence from four L1s. *Studies in Second Language Acquisition, 19*, 1–16.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitude*. Fort Worth: Harcourt Brace Jovanovich.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton: Educational Testing Service.

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics, 26*, 282–301.

Elder, C., & Harding, L. (2008). Language testing and English as an international language: constraints and contributions. In Sharifian, F. and M. Clyne (eds.): Australian Review of Applied Linguistics (Special forum issue), 31*(3),* 34.1–34.11.

Fazio, R. H., Powell, M. C., & Williams, C. J. (1989). The role of attitude accessibility in the attitude-to-behavior process. *Journal of Consumer Research, 16*(3), 280–288.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97–116). New York: Guilford.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal, 57*(4), 325–334.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman.

Galloway, N., & Rose, H. (2015). *Introducing Global Englishes*. Abingdon: Routledge.

Garrett, P. (2010). *Attitudes to language. key topics in sociolinguistics*. Cambridge: Cambridge University Press.

Giles, H., & Billings, A. C. (2004). Assessing language attitudes: Speaker evaluation studies. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 187–209). Malden: Blackwell.

Gu, L., & So, Y. (2014). Voices from stakeholders: What makes an academic English test "international"? *Journal of English for Academic Purposes, 18*, 9–24.

Hamid, M. O. (2014). World Englishes in international proficiency tests. *World Englishes, 33*(2), 263–277.

Hamp-Lyons, L., & Zhang, B. W. (2001). World Englishes: issues in and from academic writing assessment. In L. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 101–116). Cambridge: Cambridge University Press.

Harding, L. (2008). *The use of speakers with L2 accents in academic English listening assessment: A validation study*. Unpublished doctoral dissertation, The University of Melbourne, Melbourne.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Hrubes, D., Ajzen, I., & Daigle, J. (2001). Predicting hunting intentions and behavior: An application of the theory of planned behavior. *Leisure Sciences, 23*(3), 165–178.

Hsu, T. H. L. (2016). Removing bias towards World Englishes: The development of a Rater Attitude Instrument using Indian English as a stimulus. *Language Testing*, 33 (3), 367–389.

Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 183–208). New York: Academic Press.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals*. Oxford: Oxford University Press.

Jenkins, J. (2006). Current perspectives on teaching world Englishness and English as a lingua franca. *TESOL Quarterly, 40*(1), 157–181.

Jenkins, J. (2015). *Global Englishes: a resource book for students*. Abingdon: Routledge.

Kachru, B. (1992). *The other tongue* (2nd ed.). Urbana: University of Illinois press.

Kachru, B. (2001). World Englishes. In R. Mesthrie (Ed.), *Concise encyclopedia of sociolinguistics* (pp. 519–524). New York: Elsevier.

Kachru, B., Kachru, Y., & Nelson, C. (2006). *The handbook of world Englishes*. Oxford: Blackwell.

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. (2014). Functional loads of pronunciation features in non-native speakers' oral assessment. *TESOL Quarterly, 48*(1), 176–187.

Kim, H. J. (2005). *World Englishes and language testing: the influence of rater variability in the assessment process of English oral proficiency*. Unpublished doctoral dissertation. Iowa city: University of Iowa.

Kirkpatrick, A. (2012). English as an Asian lingua franca: the " Lingua Franca Approach" and implications for language education policy. *Journal of English as a Lingua Franca, 1*(1), 121–139.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *The American Psychologist, 56*, 16–26.

Kubota, R. (2001). Teaching world Englishes to native speakers of English in the USA. *World Englishes, 20*(1), 47–64.

Kunnan, A. J. (2004). Test fairness. In M. Milanvoic, C. Weir, & S. Bloton (Eds.), *European year of language conference papers, Barcelona* (pp. 27–48). Cambridge, U.K.: CUP.

Lee, H. (2012). World Englishes in a high school English case: A case from Japan. In A. Matsuda (Ed.), *Principles and practices of teaching English as an international language* (pp. 154–168). Bristol: Multilingual Matters.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: language testing and assessment. *TESOL Quarterly, 40*(1), 211–234.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States of America. *Language in Society, 31*, 419–441.

Lowenberg, P. H. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*(3), 431–435.

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

Matsuda, A., & Friedrich, P. (2011). English as an international language: a curriculum blueprint. *World Englishes, 30*(3), 332–344.

McArthur, T. (2003). *The Oxford guide to world Englishes*. Oxford: Oxford University Press.

McKay, S. L. (2012). Teaching materials for English as an international language. In A. Matsuda (Ed.), *Principles and practices of teaching English as an international language* (pp. 70–83). Bristol: Multilingual Matters.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520–531.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect size in L2 research. *Language Learning, 64*, 878–912.

Powers, D. E., Scheldi, M. A., Leung, S. W., & Butler, F. A. (1999). Validating the revised test of spoken English against a criterion of communicative success. *TOEFL Research Report, 99*(5), 63.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*, 511–531.

Saito, K. & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially percieve comprehensibility in second language speech? *TESOL Quarterly, 50*(2), 421–446.

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: the role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 141–156). Bristol: Multilingual Matters.

Sewell, A. (2013). Language testing and international intelligibility: a Hong Kong case study. *Language Assessment Quarterly, 10*, 423–443.

Sharfian, F., & Marlina, R. (2012). English as an international language (EIL): An innovative academic program. In A. Matsuda (Ed.), *Principles and practices of teaching English as an international language* (pp. 140–153). Bristol: Multilingual Matters.

Shrout, P. E., & Yager, T. J. (1989). Reliability and validity of screening scales: Effective of reducing scale length. *Journal of Clinical Epidemiology, 42*, 69–78.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 75–90). Urbana: University of Illinois Press.

Steenkamp, J. B. E., De Jong, M., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research, 47*, 199–214.

Streiner, D. L., & Norman, G. R. (2000). *Health measurement scales. A practical guide to their development and use*. New York: Oxford University Press.

Taylor, L. (2002). *Assessing learners' English: but whose/which English (es)? Research Notes 10*. Cambridge: University of Cambridge ESOL Examinations.

Taylor, L. (2005). *Linguistic diversity: Language varieties and their implications for testing and assessment*. Berlin: Paper presented at the Association of Language Testers in Europe (ALTE) Conference.

Villarreal, D., Loring, A., & Evans, K. (2014). *Teaching World Englishes to undergraduates: tensions and pedagogical insights*. Portland: Paper presented at American Association for Applied Linguistics Roundtable. March 22–25.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 186–209). Harlow: Longman.

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

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: the speech evaluation instrument. *Journal of Language and Social Psychology, l4*(2), 113–123.