This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Listening on April 4, 2016, available online: http://www.tandf.com/10.1080/10904018.2015.1056877.

Effects of strength of accent on an L2 interactive lecture listening comprehension test

This paper reports on a study which aimed to determine the effect of strength of accent on listening comprehension of interactive lectures. Test takers (N = 21,726) listened to an interactive lecture given by one of nine speakers and responded to six comprehension items. The test taker responses were analyzed with the Rasch computer program WINSTEPS (Linacre, 2012) to investigate the relative difficulty of the items associated with the nine versions of the interactive lectures. Results indicated that comprehension of interactive lectures was diminished with quite light accents, as has been found with monologic lectures.

An increasing number of researchers and practitioners advocate the use of a variety of accents for assessing second language listening ability (Abeywickrama, 2013; Smith & Bisazza, 1982; However, research indicates that listening comprehension can be impacted by a speaker's accent (Adank, et al., 2009; Adank & Janse, 2010; Derwing & Munro, 1997; Gass & Varonis, 1984; Ockey & French, in press), and as a result researchers have cautioned against randomly selecting speakers for listening inputs from the target language use domain (Elder & Harding, 2008; Ockey & French, 2014). This growing body of research has dealt primarily with monologic discourse, i.e. a single speaker giving a talk or presentation without two-way interaction with the audience. However, many listening contexts include interaction between the speaker and the listener. Even most lectures commonly involve interaction between the speaker and the audience. For instance, in many classrooms, some students ask questions when they do not understand

information presented by the speaker, and this interaction leads to different discourse than that found in a monologic lecture (Kostin, 2004). The ability to comprehend an interactive lecture, in which a test taker needs to listen to both the lecturer and other students interact with the instructor, for example, by asking questions, are therefore an important type of listening task in an academic setting. We underscore that this type of listening is not the same as the test taker interacting with the lecturer by asking questions or clarifying points of misunderstanding.

Research suggests that interactive discourse tends to be easier to comprehend than monologic discourse (Fox Tree, 1999), suggesting that it might be the case that test takers will not be affected by a speaker's accent in interactive discourse to the same extent as they are when listening to monologic lectures. Thus, there is a need to investigate whether the effect of accent on performance found in studies that have used monologic discourse holds for interactive lectures. Given this context, this paper seeks to determine the effect of accent on listening comprehension of interactive lectures. To accomplish this purpose, we used the same speakers and accent measure that was used in Ockey and French's (2014) study on monologic discourse, to investigate the possible effects of accent on interactive lectures.

Literature Review

Strength of accent

We adopt the definition of accent used by Ockey and French (2014): "The degree to which an individual's speech patterns are perceived to be different from the local variety, and how much this difference is perceived to impact comprehension of listeners who are familiar with the local variety." Thus, in our study, we assume that the variety of English that the test takers share familiarity with and expect to encounter on the assessment is the local variety. Varieties judged

by the local community of speakers to be different from this variety have stronger accents based on how far away from the local variety they are judged to be. This definition makes it clear that strength of accent is based on the speech variety of the speaker as well as that of the listener.

Research on effect of accent on L2 listening comprehension

The studies that we encountered in the literature on the effects of accent on listening comprehension are based on monologic speech. The majority of these studies found that accents do impact listening comprehension. For instance, Smith and Bisazza (1982) presented identical inputs spoken by Indian, Japanese, and American speakers. Listeners were L2 speakers of English from Hong Kong, India, the Philippines, Japan, Taiwan, and Thailand, and L1 speakers from the United States. The study found that the United States speaker was significantly easier to comprehend than the Japanese speaker, who in turn was significantly easier to comprehend than the Indian speaker. Test takers were also asked to judge the difficulty of comprehending each speaker, and their perceptions were similar to the results of the test scores. Eisenstein and Berkowtiz (1981) had similar findings in their study of adult ESL learners, who had higher comprehension for "Standard" United States English than "foreign accented English."

A notable exception to the finding that accent matters to listening comprehension of L2 learners is the study of Abeywickrama (2013). In her study, Chinese, Korean, Sri Lankan, and United States speakers delivered monologic lectures, and Brazilian, Korean, and Sri Lankan English learners were assessed on their comprehension of the lecture. No significant differences were found between the scores of the test takers who listened to the United States, Chinese, Korean, or Sri Lankan speakers.

The results of each of these studies must be treated with caution, however, since none of them include a measure of the accents of the speakers in the studies. In each study, the researcher used only country of origin of the speaker as an indicator of accent. Speakers could have had accents that were very similar to or very different than the variety of English familiar to and expected by the listeners. A very plausible explanation for the studies that found a difference due to accent is that the speakers had strong accents, whereas the results of Abeywickrama's may be explained by the use of speakers who had accents that were very weak compared to the speech variety that the test takers were used to and expected to encounter on the assessment.

One study that did attempt to control for strength of accents was conducted by Anderson-Hsieh and Kohler (1988). However, the study was based on the listening comprehension of L1 speakers of English, and therefore the results may not be completely indicative of what might be expected of L2 speakers. In the study, United States university students listened to four speakers, three who were Chinese, and one who was from the United States. The researchers used Test of Spoken English scores and judgments of the Chinese speaker's "pronunciation" to measure accent. The results of the study indicated that "heavier accents" were more difficult to comprehend and were more affected by increased spoken pace. That is, the United States speaker, who spoke the local speech variety, was the easiest to comprehend, and the Chinese speaker with the "heaviest accent" was the hardest to comprehend. The Chinese speaker with the variety of speech most similar to the United States variety was the easiest of the Chinese speakers to comprehend.

Ockey and French (in press) also considered strength of accent in their investigation of the impact of accent on listening comprehension. They used the judgments of 100 listeners to determine the strength of 20 speakers' accents, and then included nine of these speakers in their

study. All nine speakers were recorded giving the same monologic lecture, and then test takers were assigned to listen to one of them. The results indicated that the stronger the accent the lower the comprehension. The effect for decreased comprehension occurred for what Ockey and French judged to be light accents: notably different than the local dialect but not requiring more effort to comprehend than to comprehend the local variety of English.

Difficulty of interactive lectures and monologic speech

Research that compares the difficulty of comprehension of interactive and monologic speech seems to suggest that interactive speech is easier to comprehend. A possible reason is that comprehension might be facilitated by the repetition that takes place when people converse. A number of studies provide support for this notion (Freedle & Kostin, 1996; Jensen *et al.*, 1997; Buck & Tatsuoka, 1998). Other research has more generally concluded that dialogic discourse in which meaning is negotiated among the speakers increases comprehension (Ross & Langille, 1997, Rost & Ross, 1991).

A few studies have appeared in the language assessment research that either provide support or challenge the notion that interactive lectures are easier to comprehend than monologic lectures. For instance, Shohamy and Inbar (1991) found that a "consultive dialog", in which two people interacted about a topic by asking for clarification, explanation, etc., was easier to comprehend than a monologic lecture. It should be noted, however, that the monologic lecture content may have been more difficult than the consultative dialog content. Read (2002), on the other hand, found that interactive lectures that included three speakers were harder to understand than a monologue given by one person for second language adult learners. The content for the two conditions was designed to be the same in both lectures. However, Read based his

interactive lecture on the monologic lecture, which may have biased the results in the study toward making the monologic lecture easier. Brindley and Slatyer (2002) also aimed to determine if a monologic lecture would be harder to comprehend than an interactive lecture in their study of the listening comprehension of adult ESL learners. They failed to find a significant difference in the difficulty of these two lecture types. Papageorgiou, Stevens and Goodwin (2012) analyzed the performance of test takers on items developed to accompany three pairs of stimuli on the same topic, as part of a routine administration of the Michigan English Test. Each pair of stimuli consisted of a monologue and a dialogue with identical content and vocabulary and identical test items. Items associated with dialogic input were in general easier for learners than the same items associated with identical monologic input.

Summary of Relevant Research and Research Questions

The research indicates that accents judged to be increasingly different from the one familiar to the listener do decrease listening comprehension for monologic lectures. While much less clear, the research also suggests that interactive lectures can be easier to comprehend than monologic lectures. Based on these findings, it is not clear to what degree accent might impact listening comprehension of interactive lectures. Thus, it was our aim to determine the extent to which strength of accent impacts listening comprehension of interactive lectures. To achieve this goal, we addressed the following research questions:

1. How does the mean difficulty of an interactive lecture item set vary based on the speaker's strength of accent?
2. How does the difficulty of the individual test items of an interactive lecture set vary based on the speaker's strength of accent?

Method

Participants

The three types of participants of this study, university students and instructors who judged the strength of accent of speakers, speakers with different accent strengths, and all test takers who took operational TOEFL iBT on two consecutive weekends. For further details see Ockey and French (2014).

One-hundred judges were used to rate the accents of the nine speakers in the study. These judges were instructors, graduate students, and undergraduate students who were at one of three colleges or universities in the United States. Approximately one-third were advanced second language speakers and the others were first language English speakers. They were studying in a variety of disciplines.

The speakers were adult males and females from Australia, the United States, and the United Kingdom. These speakers were selected from a larger group of speakers based on the aim of including a range of accent strengths, both males and females, and speakers from each of these three countries. The average ratings of the judges for each of the nine speakers are provided in Table 1. Gender and country of origin of each of the speakers is also included.

Table 1 Speakers selected for the study

Country of origin	Gender	Mean rating on Strength of Accent Scale
United States	Female	1.1
Australian	Male	1.7
United Kingdom	Male	1.8

United Kingdom	Female	1.9
Australian	Male	2.0
Australian	Female	2.1
United Kingdom	Male	2.2
United Kingdom	Female	2.6
Australian	Female	2.7

The US female speaker was judged by the 100 judges to have the accent of the local variety; almost all judges rated her as a "1" on the Strength of accent scale, which indicates that she was not noticeably different from the local dialect. The other nine speakers had accents strengths that ranged from 1.7, on the five-point Strength of Accent Scale, which meant they were "a bit stronger than half way between not noticeable and noticeable and 2.7, which would indicate well above noticeable and nearing required concentrated listening to comprehend. Importantly, the judges' average scores suggested that none of the speakers had accents that limited comprehensibility. A further description of the Strength of Accent scale is provided in the Materials section.

All test takers who took TOEFL on two consecutive weekends (N = 21,726 from 148 countries) participated in the study. The TOEFL iBT test takers were the participants whose listening comprehension was compared across the speakers with the varying accent strengths. The TOEFL iBT test taker population includes adults L2 English learners from many parts of the world and many first languages (see http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf for details about the TOEFL test taker population).

Materials

Listening to Interactive lectures

An interactive lecture, 756 words in length, was the stimulus to be comprehended by the test takers. The lecture was on an archaeology topic, and described and considered the discovery and significance of an ancient Egyptian city. There were two speakers: the professor delivering the lectures and a student asking three clarification questions and making two comments. A total of 703 words were delivered by the professor and 53 words by the student. As test takers listened to the stimulus, several context photographs of the speakers appeared on the computer screen. Additionally, written on a single blackboard were the names of three ancient cities discussed in the lecture. Test takers were able to take notes as they listened to the lecture, and to use their notes when they answered the questions. The lecture was followed by six test items: a general idea question asking about the topic of the lecture, an inference question about a point made in the lecture, two detail questions asking about important points made in the lecture, a pragmatic understanding question asking about an opinion the professor expressed, and a connecting information question asking about the relationship between two pieces of information.

The interactive lecture that was used in the study was selected based on the following criteria: First, it had been developed and pretested to be used as a TOEFL listening interactive lecture. Second, it had content that was judged to be accessible to the large majority of TOEFL test takers. And third, it was judged to have no technical terms or vocabulary items that would be inappropriate for United States, British or Australian speaker's speech variety (further information about the TOEFL test can be found at: http://www.ets.org/toefl/ibt/about/content).

TOEFL listening test

The TOEFL listening test is presented to test takers in blocks of three sets of items. A single block consists of a conversation between two speakers, a monologic lecture, and an interactive

lecture. There are five questions based on the conversation and six questions for each lecture. Thus, for a single block of three sets of items, a test taker must answer seventeen questions. For this research, in two of these three blocks of items (k = 34), all speakers had accents judged to be of the local United States variety. As is discussed in the Procedures section below, these 34 items were used as common items across the nine tests form.

The third block of items included nine different forms of an interactive lecture accompanied by six comprehension questions. The interactive lecture and six comprehension questions were identical across the nine forms, but the accent of the speaker, who gave the lecture varied for each form. That is, each of the nine forms was delivered by one of the nine speakers selected to be in the study. Country of origin and gender of the nine speakers are shown in Table 1. It should be noted that the accent of the speaker varied for the professor across the nine test forms. The student, who asked the questions during the lecture and the narrator, who read the items after the lecture had accents that were deemed to be representative of the United States speech variety were consistent across all nine forms.

Strength of Accent Scale

To assess strength of accent, we used Ockey and French's (2014) Strength of Accent Scale. The instrument was based on a five-point scale, in which "1" indicated that the accent could not be distinguished from the local variety, "2", the accent was different from the local variety, but no extra effort was needed for comprehension; "3" indicated that the accent was notably different and extra effort was needed for comprehension. Speakers with accents stronger than "3" were not included in the study, so we do not describe levels "4" and "5" of the Strength of Accent Scale here. Ockey and French reported that their instrument was fairly reliable. To determine its

reliability, the accent strength of each speaker was judged twice by the 100 judges described in the participants' section. That is, judges listened to two independent speech samples of each speaker and provided ratings for each sample. Cronbach's Alpha indicated a reliability of .69 across the two ratings. Further analysis indicated that the average score given by the 100 judges for the two judged samples was within .2 points for eighteen of the speakers and within .3 points for one of the other two speakers on the five-point scale.

Procedures

Before they were recorded, speakers were provided with time to read through the scripts. They were given time to practice and received guidance from professional assessment developers to ensure that such factors known to affect listening comprehension such as pace were not factors in the study.

Following earlier studies investigating the comparative difficulty of items for which a specific condition has been controlled (e.g. monologic and interactive lectures input for the same items in Papageorgiou et al., 2012) we subsequently analyzed the data using the Rasch model (Rasch, 1980) operationalized by the computer program WINSTEPS (Linacre, 2012). The rationale behind this analysis was to explore the comparative difficulty of the items across the nine conditions (i.e. the nine versions of the interactive lectures that contained the different accents) on a common difficulty scale. The Rasch model produces linear measures of item difficulty and person ability on a common interval scale of 'log odds' units (McNamara, 1996, p. 165). This scale is centered on 0, and is called the 'logit' scale. Positive values indicate more difficult items, while negative values indicate easier items. The Rasch model analyzes the differences between observed and expected responses and it calculates fit statistics that indicate

the degree to which items fit the underlying construct. Of the various fit statistics calculated by WINSTEPS for each item, we inspected the infit mean square statistic because of its reliance on responses of test takers whose ability is well-matched with item difficulty on the logit scale (Bond & Fox, 2007, p. 240). The range of infit mean square statistic values across all 88 items were acceptable (0.84 to 1.27) without any items demonstrating significant underfit or overfit (see Bond & Fox, 2007, p. 240, Linacre & Wright, 1994, McNamara, 1996, p. 175); thus unidimensionality, which is an essential measurement condition for Rasch analysis, was shown to be tenable. Estimates of item difficulty and fit are further discussed in the Results section.

The robustness of the Rasch model to missing data (Bond & Fox, 2007, p. 312), as implemented in WINSTEPS, was critical for our analysis, since each test taker listened to only one of the nine speakers, who delivered the majority of the interactive lecture. Although the six items of interest, which were based on the interactive lecture given by the nine different speakers, were administered to nine different groups of examinees, a comparison of the difficulty of all items was possible using WINSTEPS. This was accomplished by using the 34 common items as linking items. . The responses to all 88 items: the34 common items and 54 unique items (six for each of the nine conditions) – were analyzed with the dichotomous Rasch model, in which 1 denotes a correct answer and 0 a wrong answer.

Results

Table 2 presents the mean logit values of the six items for each of the nine conditions. The first three columns provide the speakers' country of origin, gender, and strength of accent as measured by the Strength of Accent Scale (Ockey & French). The mean logit value of the six items of each condition, that is, the average listening comprehension of the test takers for each of

the nine speakers, is presented in the third column (see Appendix for individual item statistics across all nine speakers). We used a 95% confidence interval of the logit value as a measure of the reliability of logit estimate and as an indication of how substantively item difficulty varied across conditions. The model error calculated by WINSTEPS was 0.04 and it was employed in order to calculate the 95% confidence interval. The results in Table 2 suggest that the overall item difficulty is in general comparable across the nine conditions. However, item difficulty between the six weakest accents is lower than for the three strongest accents (2.2 or higher on the Strength of Accent Scale). This can be seen by comparing the mean logit values and the confidence intervals. The mean logits of the strongest three accents do not fit within the 95% confidence intervals of those of the weakest six accents, with the exception of the third speaker, whose rating on the Strength of Accent Scale was 1.8. In other words, the passages delivered by the speakers with the three accents that were judged to differ most from the local dialect, which was expected on the test by the test takers, were more difficult than the others.

Table 2 Mean logit value by condition

Speaker's country of origin	Speaker's gender	Speaker's rating on Strength of Accent Scale	Mean Logit value of listener's comprehension	Confidence interval[1] of logit value for listener's comprehension
United States	Female	1.1	0.19	(0.11, 0.27)
Australian	Male	1.7	0.18	(0.10, 0.26)
United Kingdom	Male	1.8	0.27	(0.19, 0.35)
United Kingdom	Female	1.9	0.20	(0.12, 0.28)
Australian	Male	2.0	0.13	(0.05, 0.21)
Australian	Female	2.1	0.09	(0.01, 0.17)
United Kingdom	Male	2.2	0.35	(0.27, 0.43)
United Kingdom	Female	2.6	0.31	(0.23, 0.39)
Australian	Female	2.7	0.40	(0.32, 0.48)

[1] Mean logit value +/- two times the model error. The model error is this analysis was 0.04.

Given that some accents were more difficult for test takers to comprehend than others at the test level, that is, across the whole six-item set, we desired to determine if any particular item could be shown to display differential difficulty for the different accents. To achieve this purpose, we examined the range of logit values for each individual item across the nine conditions. In other words, we identified for each individual item the lowest and the highest logit values obtained across the nine accents. Table 3 presents the minimum and maximum values observed for each of the six items across the 9 conditions along with the 95% confidence interval of the logit value as a measure of the reliability of logit estimate and as an indication of how substantively item difficulty varied across conditions. For example, the lowest value for Item 1 was -0.16, whereas the highest logit value was 0.10 (see Appendix for individual item statistics across all nine speakers). The difference in item difficulty ranged from 0.21 logits (Item 3) to 0.79 logits (Item 2). To interpret the above differences in the difficulty of individual test items across the nine conditions, as opposed to differences in difficulty for each six-item set examined previously, we turn to Linacre's (2012) guidelines for examining Differential Item Functioning (DIF) using WINSTEPS (see also Banerjee & Papageorgiou, this issue). Linacre categorizes differences between 0.40 and 0.60 logits as "slight to moderate" and differences larger than 0.60 logits as "moderate to large". Based on these criteria, one item demonstrated substantive difference in item difficulty across the nine conditions, while the item difficulty of the others was comparable, irrespective of the speaker's strength of accent. As can be seen in Table 3, Item 2, which required making an inference about what the professor implied, was the item shown to vary substantively, 0.70 to 1.49, in terms of difficulty across the various accents (0.79 logits difference between the easiest and most difficult accent). As can be seen in the Appendix, the

strongest three accents were associated with the more difficult conditions. It should be noted that this item was consistently the most difficult item in the six-item set across all nine speakers (see Appendix).

Table 3 Range of difficulty of items

Item Number	Minimum difficulty estimate (Logits)			Maximum difficulty estimate (Logits)		
	Logit Value	Error	Confidence interval[1]	Maximum	Error	Confidence interval[1]
Item 1	-0.16	0.05	(-0.26, -0.06)	0.10	0.05	(0.00, 0.20)
Item 2	0.70	0.05	(0.60, 0.80)	1.49	0.05	(1.39, 1.59)
Item 3	-0.02	0.05	(-0.12, 0.08)	0.19	0.05	(0.09, 0.29)
Item 4	0.04	0.05	(-0.06, 0.14)	0.51	0.05	(0.41, 0.61)
Item 5	-0.27	0.05	(-0.37, -0.17)	0.13	0.05	(0.03, 0.23)
Item 6	-0.02	0.05	(-0.12, 0.08)	0.28	0.05	(0.18, 0.38)

1 Mean logit value +/- two errors

Because the difficulty of Item 2 was found to vary substantively across the various accents, we conducted a content analysis of this item, with the aim of determining the extent to which certain aspects of strength of accent might be identified that could explain this difficulty difference. An inspection of the turn where the key to the item is found in the interactive lecture revealed that test takers may need to comprehend a word whose pronunciation varies between the standard United States accent that was used as the local variety of English to which other varieties were compared in our study and United Kingdom or Australian English. For instance, “artifact”, (UK: /ˈɑːtɪfækt/; US:/ˈɑɹtɪfækt/) varies on the use of the liquid “r”. Artifact appears once in the relevant sentences and twice before these and it it may be plausible that this could contribute to lack of comprehension of the information necessary to respond correctly to this item.

Discussion

To address the first research question, which was "how does the mean difficulty of an interactive lecture item set vary based on the speaker's strength of accent?", Rasch analysis was used to compare overall item difficulty for the six items across the nine conditions, i.e. the nine speakers with different ratings on Ockey and French's (in press) Strength of Accent Scale. In general, the scores for the test takers who encountered the six speakers with the weakest accents did not differ on average. However, an increase in item difficulty was found with the three stronger accents, whose ratings on the Strength of Accent Scale were 2.2 or higher. This finding is not surprising given that it is in line with of much of the other research that has been conducted on the effects of strength of accent on monologic listening comprehension (Anderson-Hsieh and Kohler, 1988; Eisenstein and Berkowtiz, 1981; Ockey & French, in press; Smith and Bisazza, 1982).

Rasch analysis conducted in order to address the second research question, which was "How does the difficulty of the individual test items of an interactive lecture set vary based on the speaker's strength of accent?", revealed that in general difficulty of individual items was similar, irrespective of the speaker's strength of accent. However, one item (Item 2) demonstrated substantive differences in item difficulty across the nine speakers, with the highest logit values consistently noted with speakers judged to have stronger accents. Subsequent content analysis revealed that to respond correctly to this item, test takers might have needed to comprehend words in the stimulus whose pronunciation varied depending on the speaker's accent. We hypothesize that these differences in pronunciation of a key word important for understanding may have led to this decreased comprehension. We note that we cannot make any strong claims based on this limited content analysis. We suggest, however, that further research be conducted to determine if certain item types could be affected by accent strength more than

others. It is also noteworthy that this item was the most difficult across all speakers, which may suggest that item difficulty and strength of accent could interact with each other. That is, difficult items may be more impacted by unfamiliar speech varieties than ones that are easier. Given that there was only one item that was shown to be differentially difficult for the different accent strengths, it was not possible to provide support for these hypotheses by investigating other items. Thus, we do not make any strong claims about these possible explanations. Rather we suggest that they might provide avenues for further research on the effects of accent on comprehension.

An implication of this study is that different strengths of accents, even very light ones, like those used in this study (judged to be completely comprehensible by a panel of 100 jduges), can affect comprehension of interactive lectures. Thus, it is important that careful attention to strength of accent be considered when designing listening comprehension assessments.

It could be argued that the results of this study suggest that accents that are slightly stronger than noticeable on the Strength of Accent Scale could be used for L2 interactive lecture listening comprehension inputs without unduly affecting test scores. This finding for the effects of accent on L2 listening comprehension of interactive lectures is similar to the finding of that of Ockey and French (2014), who found a similar but slightly lower threshold for the effects of accent on monologic discourse. Interactive lectures might be easier to comprehend given what others, who have conducted research on monologic versus dialogic discourse, have found from unaccented speakers. However, this conclusion should be considered in relation to the numerous, well-documented variables that affect listening comprehension (Brindley, 1998; Brindley & Slatyer, 2002; Buck & Tatsuoka, 1998; Freedle & Kostin, 1999; Kostin, 2004; Nissan et al., 1996). Although we controlled for such variables by administering identical items and stimulus

content, and by randomly sampling test takers for each of the nine conditions, the extent to which other variables might have affected item difficulty cannot be completely discounted.

While this study contained a large number of test takers and care was taken to control all variables that might have affected the results other than the variable of interest, which was accent of speakers, the generalizability of the results is limited by a number of design features. First, the test takers were all drawn from the TOEFL iBT test taker population. The effects of accent for these listeners may not be the same as for others, who may have more or less exposure to various accents. Second, we assumed that by random sampling techniques, familiarity, as well as many other factors known to affect listening comprehension would be controlled for in the study. However, we do not discount the possibility that these factors had some impact on our findings. Our results are based on the assumption that all test takers were familiar with and expected to encounter United States English on the test. While TOEFL iBT has only used United States accents, this assumption may not be completely defensible, given the diverse group of TOEFL test takers. Again, we count on random sampling techniques to control for possible effects of some test takers not being familiar with the United States variety of English. Third, our findings are based on six items. This means we were unable to make generalizations about particular item types or about the possible effects of a large number of items that could have an additive effect on differential difficulty across accents. Fourth, our research design is limited in that we were only able to use one passage. This means the findings may not be generalizable across passage types since a passage effect may have occurred. However, when our findings are judged in conjunction with other similar findings, it makes it possible to suggest that these findings may be generalizable across passages. We also note that attention span may play a role in our results, given that test takers may have lost focus while listening to the interactive lectures. However, we

believe that this would have limited effect given that the lectures were quite short, and test takers were allowed to take notes while listening.

Conclusion

Our results indicate that comprehension of interactive lectures is affected by rather light accents. Comprehension of speech that is more than noticeably different than the variety of speech familiar to and expected by the test takers is more difficult for them to comprehend. We also found that particular item types might be affected more by a speaker's accent than others. Based on our findings, we hypothesized that a possible explanation is that items with (repeated) words that are pronounced differently than the local dialect may affect comprehension. We conclude by stating that we strongly believe that when more than one variety of speech is commonly encountered in the target language use domain, then listening comprehension tests need to be multidialectal. However, a defensible approach--one that does not unfairly impact any of the test takers--for how and to what extent these varieties are used, must be employed.

References

Abeywickrama, P. (2013). Why not non-native varieties of English as listening comprehension test input? *RELC Journal, 44(1),* 59-74.

Adank, P., Evans, B., Stuart-Smith, J., & Scott, S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology, 35(2),* 520-529.

Adank, P., & Janse, E. (2010). Comprehension of a Novel Accent by Young and Older Listeners. *Psychology and Aging, 25(3),* 736–740

Anderson-Hsieh, J. & Kohler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38 (4), 561-613.*

Bilbow, G. T. (1989). Towards an understanding of overseas students' difficulties in lectures: A phenomenographic approach. *Journal of Further and Higher Education, 13,* 85–89.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18,* 171-191.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394.

Buck, G. (2001). *Assessing listening.* Cambridge, U.K.: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech.

Cognition, 106(2), 707-729.

Department of Institutional Research (2007) Faculty by ethnic group [On-line]. http://www.irs.ttu.edu/NEWFACTBOOK/ Faculty/2007/F07ETHNIC.htm

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20,* 1-16.

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

Eisentein, M. R., & Berkowitz, D. (1981). The effect of phonological variation on adult learner comprehension. *Studies in Second Language Acquisition, 4*, 75–80.

Ekong, P. (1982). On the use of an indigenous model for teaching English in Nigeria. *World Language English, 1*, 87–92.

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

Field, J. 2004. 'Pronunciation acquisition and the individual learner'. Presentation at the IATEFL Joint Pronunciation and Learner Independence Special Interest Groups Event, University of Reading, 26 June 2004.

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes, 27,* 35–53.

Freedle, R. and Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: implications for construct validity*. TOEFL Research Report 56. Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32.

Harding, L. (2011). *Accent and listening assessment: A validation of the use of speakers with L2 accents on an academic English listening test*. Frankfurt: Peter Lang.

Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. (Vol. TOEFL Research Reports): Educational Testing Service.

Jensen, C., Hansen, C., Green, S. and Akey, T. (1997). An investigation of item dif. culty incorporating the structure of listening tests: a hierarchical linear modeling analysis. In Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S., editors, *Current developments and alternatives in language assessment*. Jyva¨skyla¨: University of Jyva¨skyla¨, 151–64.

Kostin, I. (2004). *Exploring Item characteristics that are related to the difficulty of TOEFL dialogue items* (No. RR–79). Princeton, NJ: Educational Testing Service.

Linacre, J. M. (2012). WINSTEPS Rasch measurement computer program version 3.74.0. Chicago: Winsteps.com.

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

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

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*, 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38(3)*, 289-306.

Major, R., Fitzmaurice, S., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36(2),* 173-190.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension *TOEFL Research Report No. RR-51* (Vol. 51). Princeton, NJ: Educational Testing Service.

Ockey, G., & French, R. (in press). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*.

Ortmeyer, C., & Boyle, J. (1985). The effect of accent differences on comprehension. *RELC Journal 16(2),* 48–53.

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly, 9*(4), 375-397.

Ross, S. and Langille, J. (1997). Negotiated discourse and interlanguage accent effects on a second language listening test. In Brindley, G. and Wigglesworth, G., editors, *Access: issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 87–116.

Rost, M. and Ross, S. (1991). Learner use of strategies in interaction: typology and teachability. *Language Learning* 41, 235–73.

Schmid, P., & Yeni-Komshian, G. (1999). The effects of speaker accent and target predictability on perception of mispronunciation. *Journal of Speech, Language, and Hearing Research, 42,* 56-64.

Smith, L., & Bisazza, J. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning, 32(2),* 259-269.

Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal*, 28, 54-71.

Appendix: Items Statistics for All Conditions

Speaker's country of origin	Speaker's Gender	Speaker's strength of accent	Item	Number of test takers	Proportion correct	Logit	Standard Error	Infit mean square
United States	Female	1.1	1	4342	0.57	0.10	0.04	0.97
			2	4340	0.40	0.98	0.04	1.02
			3	4338	0.57	0.07	0.04	1.19
			4	4332	0.58	0.04	0.04	1.06
			5	4320	0.63	-0.22	0.04	1.11
			6	4264	0.56	0.16	0.04	1.13
Australia	Male	1.7	1	2212	0.59	0.01	0.05	0.98
			2	2211	0.43	0.87	0.05	1.06
			3	2211	0.58	0.04	0.05	1.18
			4	2210	0.55	0.20	0.05	1.13
			5	2201	0.63	-0.25	0.05	1.07
			6	2180	0.55	0.23	0.05	1.14
United Kingdom	Male	1.8	1	2184	0.60	-0.10	0.05	0.95
			2	2182	0.35	1.28	0.05	1.03
			3	2180	0.55	0.18	0.05	1.23
			4	2174	0.54	0.23	0.05	1.11
			5	2165	0.61	-0.15	0.05	1.11
			6	2149	0.54	0.20	0.05	1.12
United Kingdom	Female	1.9	1	2162	0.62	-0.14	0.05	0.92
			2	2162	0.37	1.21	0.05	1.11
			3	2162	0.57	0.11	0.05	1.21
			4	2159	0.54	0.25	0.05	1.12
			5	2145	0.63	-0.22	0.05	1.08
			6	2119	0.60	-0.02	0.05	1.10
Australia	Male	2.0	1	2155	0.59	-0.06	0.05	0.95
			2	2155	0.41	0.94	0.05	1.09
			3	2153	0.59	-0.02	0.05	1.21
			4	2147	0.57	0.04	0.05	1.16
			5	2137	0.63	-0.27	0.05	1.13
			6	2104	0.56	0.14	0.05	1.14
Australia	Female	2.1	1	2164	0.61	-0.16	0.05	1.01
			2	2164	0.45	0.70	0.05	1.05
			3	2161	0.56	0.10	0.05	1.22
			4	2156	0.55	0.15	0.05	1.11
			5	2146	0.63	-0.26	0.05	1.13
			6	2128	0.58	0.01	0.05	1.13
United Kingdom	Male	2.2	1	2147	0.57	0.08	0.05	1.00
			2	2147	0.34	1.31	0.05	1.08
			3	2147	0.56	0.18	0.05	1.20
			4	2142	0.55	0.20	0.05	1.09
			5	2136	0.59	0.02	0.05	1.15
			6	2101	0.54	0.28	0.05	1.11
United Kingdom	Female	2.6	1	2171	0.60	-0.04	0.05	0.97
			2	2171	0.32	1.49	0.05	1.06
			3	2168	0.58	0.08	0.05	1.17
			4	2167	0.55	0.25	0.05	1.13
			5	2162	0.59	0.04	0.05	1.12
			6	2134	0.59	0.02	0.05	1.11
Australia	Female	2.7	1	2130	0.59	-0.01	0.05	1.02
			2	2130	0.33	1.39	0.05	1.11
			3	2128	0.55	0.19	0.05	1.24
			4	2126	0.49	0.51	0.05	1.09
			5	2119	0.56	0.13	0.05	1.18
			6	2087	0.55	0.20	0.05	1.18