The Validity of Pronunciation Measures:
Relative Impact of Assessment Methods on Learners' Pronunciation

Alyssa A. D. Kermad

Northern Arizona University

Abstract

In the pronunciation literature, there has been a wide use of diverse tasks to assess second language speech; however, there has been little research as to the effect of the task on listener ratings of comprehensibility, accentedness, or proficiency. Furthermore, few studies have explored the role of the task in the speech production of phonological features, including segmentals, fluency, and prosody. This paper compares the most commonly used tasks in the L2 pronunciation literature (i.e., read aloud, spontaneous speech, elicitation, and picture-description tasks) and examines their effect on listener ratings and phonological features. Fifteen speakers from three proficiency levels participated. Fifty-six untrained listeners rated speech stimuli across the four aforementioned tasks. One-way repeated measures ANOVAs were conducted to determine the effect of the task on the outcome variables. Results showed hierarchically parallel results across tasks: speakers were rated to be more comprehensible and proficient and have less of an accent on the elicitation task, followed by the read-aloud, spontaneous, and picture description tasks. Learners had more vowel and consonant deviations and a faster speech rate on the read-aloud task. The findings can be applied to assessment, research, and pedagogy as they lead to more valid measures of assessing L2 speech.

Keywords: pronunciation tasks, pronunciation assessment, tasks and listener ratings, tasks and phonological features

The Validity of Pronunciation Measures:
Relative Impact of Assessment Methods on Learners' Pronunciation

Background

In the field of second language (L2) pronunciation, a multiplicity of speaking tasks have been used in research and/or in the classroom based on popularity or convenience of the task. These tasks have been used by listeners to rate non-native speakers' (NNSs) speech production. The various tasks have been the mode used to channel speech production, and the task instructions generate the sort of language that is produced. Not all tasks yield the same language output; for example, a read-aloud or elicitation task is not generated by the speaker, so the speaker only needs to read or repeat the language provided. A spontaneous speech task or picture description task requires the speaker to generate their own language in order to convey ideas about the specific prompt, but both with different communicative goals. With such a variety of tasks available, a researcher/educator may need to recognize that a targeted language production can become specific to the chosen task.

The primary purpose of this study is to explore a variety of tasks used to stimulate speakers' language output and investigate the effect of these individual tasks on listeners' ratings of NNSs' comprehensibility, accentedness, and oral proficiency. Starting with the task as the source of speech stimuli, speech constructs can be more appropriately measured when the task is closely correlated with the targeted language output. This study starts at the source, with the task, and conducts analyses by using novice raters' judgments of comprehensibility, accentedness, and proficiency as a means of more appropriately justifying the use of a task for a specific construct. It further examines the influence of the task on the outcome of phonological features, including segmentals, fluency, and prosody. The findings of this study can contribute to informed choices of task usage in L2 pronunciation research and pedagogy.

Research Questions

The current study attempts to answer the following research questions: 1) How do different tasks used to assess non-native speech affect listeners' rating scores of overall pronunciation (comprehensibility, accentedness, and proficiency)? And 2) How do different tasks used to assess non-native speech affect phonological features including vowels, consonants, speech rate, and pitch range?

Methods

Participants

Speakers. Fifteen participants were recruited from the Program in Intensive English (PIE). Five participants were enrolled in Level 3, four participants were enrolled in Level 4, and six participants were enrolled in Level 5.

Listeners. Fifty-six untrained raters were recruited from four undergraduate classes at Northern Arizona University. These participants were enrolled in required, 4-credit composition classes at the university. Nine were NNSs of English and 47 were NSs.

Recorded listening stimuli. The recorded listening stimuli reflected the four most commonly used assessment methods (e.g. read-aloud tasks, spontaneous speech tasks, elicitation tasks, and picture description tasks) in the field of L2 pronunciation (Thomson & Derwing, 2014). Each participant produced a total of 4 recordings each. As a result, 60 speech files were included in the overall analysis: 15 read aloud samples; 15 spontaneous speech samples; 15 elicitation samples; and 15 picture-description samples.

Instrument

There were three outcome variables included in this study as a result of listener ratings: comprehensibility, accentedness, and oral proficiency. Comprehensibility was operationalized by a single item, 9-point scale, as in Munro and Derwing (1995a, 1995b): "The speaker to whom

I just listened was extremely easy/impossible to understand." Accentedness was operationalized by a single item, 9-point scale, as in Munro & Derwing (1995a, 1995b): "The speaker to whom I just listened had no/a very strong foreign accent." Oral proficiency was measured using an adapted version of Kang's (2012) eight item oral English proficiency scale (having .92 internal consistency of subscales). This adapted scale consisted of four semantic differential items measuring (1) pronunciation/accent, (2) grammar, (3) vocabulary, and (4) rate of speech.

Procedures

The sixty files were presented to listeners in a randomized order. Through the online survey program, Survey Gizmo, (https://www.surveygizmo.com/), the listeners rated each speech file on the constructs in the order presented above: comprehensibility, accent, proficiency (pronunciation/accent, grammar, vocabulary, and rate of speech).

Phonetic and Phonological Analysis

The native-speaker rating judgments were accompanied by quantitative phonological and phonetic analyses conducted through Praat. The quantitative analyses of linguistic variables in this study included measures of segmentals (i.e. vowel and consonant deviations), fluency (i.e. speech rate), and prosody (i.e. pitch range).

Data Analysis

Three one-way, repeated measures, analyses of variance (ANOVA) were conducted to answer the first research question concerning the effect of the task on the dependent variables (comprehensibility, accentedness, and oral proficiency). To answer the second research question concerning how the task type affects phonological variables (i.e. vowel deviations, consonant deviations, pitch range, and speech rate), four one-way repeated measures ANOVAs were conducted.

Results

Listener Ratings of Accentedness

There was a significant overall effect of the task and *accentedness* ratings: Greenhouse Geisser (used to correct for the violation of sphericity) = $F(2.94, 2464.43) = 37.42$, $p < .001$. There was a small-sized effect for the non-significant omnibus, *partial* $\eta^2 = .043$. Repeated-measures *t*-tests (using a Bonferroni adjustment, $\alpha = .05/6 = .008$) showed participants were rated as having less of an accent on the read aloud task than on the spontaneous speech task, $t(839) = -4.45$, $p < .001$, $ES = .21$, or on the picture-description task $t(839 = -5.56$, $p < .001$, $ES = .26$. They were rated as having less of an accent on the elicitation task than on the read-aloud task, $t(839) = 4.39$, $p < .001$, $ES = .20$, the spontaneous speech task, $t(839) = 7.92$, $p < .001$, $ES = .41$, and on the picture-description task, $t(839) = -8.91$, $p < .001$, $ES = .46$. For all pair-wise comparisons, there was a large effect size.

Listener Ratings of Comprehensibility

There was a significant effect of the task and *comprehensibility* ratings: Greenhouse Geisser, $= F(2.91, 2442.99) = 70.08$, $p < .001$. There was a medium-sized effect for significant omnibus results, *partial* $\eta^2 = .077$. Repeated-measures *t*-tests (using a Bonferroni adjustment, $\alpha = .05/6 = .008$) showed participants were rated as being more comprehensible on the read-aloud task than on the picture-description task, $t(839) = -9.03$, $p < .001$, $ES = .41$. They were rated as being more comprehensible on the elicitation task than on the read-aloud task, $t(839) = 6.45$, $p < .001$, $ES = .28$, the spontaneous speech task, $t(839) = 7.01$, $p < .001$, $ES = .35$, and on the picture-description task $t(839) = -13.59$, $p < .001$, $ES = .69$. Finally, speakers were rated as more comprehensible on the spontaneous speech task than on the picture description task $t(839) = -$

6.75, $p < .001$, *ES* = .34. Similar with the accentedness ratings, for all pair-wise comparisons of comprehensibility ratings, there was a large-sized effect.

Listener Ratings of Oral Proficiency

There was a significant effect of the task and *oral proficiency* ratings: Greenhouse Geisser = F (2.86, 2396.25) = 77.58, $p < .001$. There was a medium-sized effect for significant omnibus results, *partial* $\eta^2 = .085$. Repeated-measures *t*-tests (using a Bonferroni adjustment, $\alpha = .05/6 = .008$) showed participants were rated as being more proficient on the read-aloud task than on the picture-description task, $t(839) = -9.20$, $p < .001$, *ES* = .41. They were rated as being more proficient on the elicitation task than on the read-aloud task, $t(839) = 7.35$, $p < .001$, *ES* = .32, the spontaneous speech task, $t(839) = 7.78$, $p < .001$, *ES* = .41, or on the picture-description task $t(839) = -13.83$, $p < .001$, *ES* = .72. Finally, they were rated as more proficient on the spontaneous speech task than on the picture description task, $t(839) = -6.62$, $p < .001$, *ES* = .32. Similar to the ratings of accentendess and comprehensibility, all pair-wise comparisons for proficiency ratings had a large-sized effect.

Effect of the Task on Phonological Features

The effect of the task on *vowel deviations* was significant: Mauchly's W = F (3, 42) =3.90, $p < .05$. There was a large-sized effect for significant omnibus results, *partial* $\eta^2 = .218$. Post hoc comparisons using the paired-samples *t*-test procedure were used to determine which pairs differed significantly using the Bonferroni adjustment, $\alpha = .05/6 = .008$. Results indicated that participants had more vowel deviations on the read-aloud task than on the picture description task, $t=(14) = 4.04$, $p = .001$, *ES* = 1.56, having a large-sized effect.

There was also a significant effect of the task on *consonant deviations*, Greenhouse Geisser = F (2.10, 29.37) = 4.41, $p < .05$. There was a large-sized effect for significant omnibus

results, *partial* η^2 = .240. Post hoc comparisons using the paired-samples *t*-test procedure were used to determine which pairs differed significantly using the Bonferroni adjustment, α = .05/6=.008. Results indicated that participants had more consonant deviations on the read-aloud task than the picture-description task, *t*=(14) = 3.20, *p* =.006, *ES* = .85, having a large-effect size.

There was not a significant effect of the task for *pitch range*, Mauchly's M = *F* (3, 42) = 1.807, *p* >.05. There was, however, a medium-sized effect for non-significant omnibus results, *partial* η^2 = .114.

Finally, there was a significant effect of the task for *speech rate*, Greenhouse Geisser = *F* (1.812, 25.372) = 4.777, *p* <.05. There was a medium-sized effect for significant omnibus results, *partial* η^2 = .114. Post hoc comparisons using the paired-samples *t*-test procedure were used to determine which pairs differed significantly using the Bonferroni adjustment, α = .05/6=.008. The results indicated that participants had a faster speech rate on the read-aloud task than on the elicitation task, *t*=(14) = 5.46, *p* < .001, *ES* = .95, and a faster speech rate on the read-aloud task than on the picture description task, *t*=(14) = 4.46, *p* $\leq$ .001, *ES* = .87. All pair-wise comparisons had a large-sized effect.

Relevance to PIE and Second Language Learning

The findings indicate that the more controlled the task, the less of an accent and more proficient and comprehensible the speakers were rated. On the elicitation, speakers had less of an accent and were most comprehensible and proficient. The read-aloud task was similar to the elicitation task, in that the expected output was provided, but without the exemplary native-speaker model. This could explain why the speakers received better ratings on the read-aloud task than the spontaneous or elicitation task. The spontaneous speech task consistently followed the read-aloud task, as speakers had more control of their output in this task than with the

picture-description. The spontaneous speech task was relevant to the speakers, and they could also control for the language which was accessible to them. The picture-description task, however, not only required learners to speak of vocabulary and concepts with which they were not familiar, but it required learners to link the ideas in some sort of cohesive fashion. This task was the most cognitively demanding, and the speakers consistently received more negative ratings of accentendess, comprehensibility, and proficiency.

The nature of the task also affected certain phonological features. For segmentals, the speakers produced the most vowel and consonant deviations on the read-aloud task and the elicitation task. Again, these two tasks did not allow the opportunity for learners to avoid problematic segmentals. Finally, speakers had an increased speech rate on the read-aloud task.

The PIE can benefit from the results of this study, as task choice becomes important for both research and testing purposes. To the researcher's knowledge, the picture-description task if often employed at the PIE. The results of this study revealed that the picture-description task affected accentedness, comprehensibility, and proficiency ratings most severely. As the picture-narrative task involves "encoding new, visual information into linguistic form" (Foster & Skehan, 1996, p. 307), this task not only requires a greater cognitive effort, but more complex linguistic forms in order to synthesize and connect ideas.

For diagnostic purposes, instructors may turn to the read-aloud task to more appropriately diagnose segmental deviations of learners. This may be a more adequate representation of a learner's repertoire of sounds, rather than a spontaneous speech task, for example. Instructors may also have students engage in reading activities for fluency and speed purposes. Picture description tasks can be used in the classroom to help students prepare their speech under pressure, and elicitation tasks may be used to help students replicate native models.

References

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299-323.

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

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

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*, 326-344.