Research Article

Goodness and Accentedness Ratings of /hVt/ Tokens by Aware and Naive Listeners

Amber D. Franklin,^a Kara A. Oksanen,^a and Kaitlyn E. Gilfert^a

Purpose: This study compares goodness and accentedness ratings of speech tokens rated by listeners who are naive to and aware of speakers' native language backgrounds. Listener responses to open-ended questions regarding goodness and accentedness ratings are also examined.

Method: Twenty-eight monolingual speakers of U.S. English served as listeners. Listeners were presented with 5 blocks of /hVt/ tokens. Each block represented a different vowel target and comprised correct and incorrect productions from English, Spanish, Korean, and Japanese speakers. Listeners rated goodness and accentedness using a 9-point Likert scale and explained their decision-making criteria when judging goodness versus accentedness.

Results: There is a high positive correlation between goodness and accentedness. Both naive and aware listeners assigned poorer ratings when judging goodness compared with accentedness, but results varied on the basis of target accuracy. Aware listeners assigned better goodness and accentedness ratings compared with naive listeners. This difference was highly statistically significant. Both accentedness and goodness ratings are susceptible to listener bias.

Conclusions: Goodness and accentedness are highly correlated yet distinct measures. Goodness is more reflective of target accuracy than is accentedness. Native English tokens were affected by listener bias to a greater extent than nonnative English tokens.

Speech-language pathologists must always guard against linguistic, cultural, or ethnic bias when they assess and treat clients with communication disorders. Concern over the presence of bias in clinical practice has led to the development of several cultural-competence checklists by the American Speech-Language-Hearing Association (ASHA), including ones on policies and procedures (ASHA, 2010b), service delivery (ASHA, 2010c), and personal reflection (ASHA, 2010a). ASHA acknowledges that neither regional nor foreign accents represent speech and language disorders, but rather speech differences (ASHA, 1997). In this vein, the aforementioned checklists address bias concerning not only individuals with communication disorders but also individuals with speech and language differences.

For example, the Personal Reflection checklist (ASHA, 2010a) asks users if they recognize that the use of a foreign accent does not reflect reduced intellectual capacity. The Policies and Procedures checklist (ASHA, 2010b) asks users whether their agency's promotion of persons who speak with accents is based on job performance rather than their use of Mainstream American English. This last checklist

item acknowledges that Mainstream American English is only one of the many varieties of English spoken in the United States.

It is important for clinicians to improve their cultural competency regarding the decisions they make and their attitudes toward clients' linguistic backgrounds. In this vein, clinicians should understand whether particular assessment measures may lead to inadvertent bias toward different linguistic groups. Even within the practice of accent modification—an area in which clinicians should be aware of differential bias toward particular accents (Deprez-Sims & Morris, 2010; Kim, 2001)—clinicians may be at risk of using assessment measures that show a preference for some accents over others. To this end, the present study compares two measures of pronunciation: goodness and accentedness ratings. This study further examines how listeners who are naive to and aware of speakers' language backgrounds apply both measures when rating foreign-accented and native English speech tokens.

Measures of Pronunciation Proficiency

There are a number of ways to measure pronunciation proficiency among English language learners (ELLs). Examples of these measures include accentedness, intelligibility, comprehensibility, and goodness. Intelligibility is defined by Derwing and Munro (2005) as how well a listener understands

^aMiami University, Oxford, OH

Correspondence to Amber Franklin: franklad@miamioh.edu

Editor: Krista Wilkinson

Associate Editor: Jack Ryalls

Received July 21, 2015

Revision received April 2, 2016

Accepted April 19, 2016

DOI: 10.1044/2016_AJSLP-15-0106

Disclosure: The authors have declared that no competing interests existed at the time of publication.

a speaker's speech expression, or the number of words understood by the listener. Comprehensibility is a listener's perception of how challenging it is to comprehend an utterance, and accentedness is defined as a listener's perception of the ELL's speech as different from that of the listener's own language community (Derwing & Munro, 2005). Goodness, on the other hand, is defined as a listener's judgment of an utterance's accuracy compared with the listener's concept of an ideal prototype (Franklin & Stoel-Gammon, 2014). A number of studies have investigated how these measures of speech relate to one another (Behrman & Akhund, 2013; Derwing & Munro, 1997; Franklin & Stoel-Gammon, 2014; Munro & Derwing, 1995b; Trofimovich & Isaacs, 2012).

A study by Derwing and Munro (1997) investigated intelligibility, comprehensibility, and accentedness ratings of speakers from Cantonese, Polish, Japanese, and Spanish backgrounds. Ratings were assigned by 26 native speakers of English. Comprehensibility and accentedness were rated using a 9-point Likert scale and intelligibility was measured through an orthographic transcription task. The study's findings revealed that intelligibility, accentedness, and comprehensibility are related but not equivalent measures. Listeners assigned poorer ratings to accentedness than to comprehensibility, and comprehensibility was rated poorer than intelligibility. These results suggest that nonnative speakers can be intelligible and comprehensible even when they speak with a noticeable accent.

Behrman and Akhund (2013) investigated variables affecting accentedness, comprehensibility, and intelligibility ratings by having adult Spanish ELLs read true/false statements, semantically meaningful statements, and semantically unusual statements. Comprehensibility and intelligibility were measured by five English speakers using a 9-cm visual analog scale. Intelligibility was measured by comparing orthographic transcriptions of the information-bearing words with the statement targets. Findings indicated that as semantic context became more abstract, listeners rated accentedness and comprehensibility poorer, and intelligibility scores decreased. This study also found that intelligibility and comprehensibility ratings were correlated more strongly with each other than either rating was with accentedness.

A study by Trofimovich and Isaacs (2012) also found accentedness and comprehensibility ratings to be correlated. The researchers investigated which components of an utterance influenced ratings of comprehensibility and accentedness. Accentedness was correlated with phonological errors and syllable structure, whereas comprehensibility had stronger correlations with errors in grammar and vocabulary.

Hayes-Harb and Watzinger-Tharp (2012) explored the correlation of accentedness and intelligibility, finding pronunciation to be a key component of ratings. In this study, German listeners rated English speakers learning German. Findings indicated that half of the raters thought that pronunciation was the most important factor when rating intelligibility of the German language learners. Three of the 12 listeners noted that pronunciation was also important for a learner to be accepted by native German speakers as having near-native speech. When raters thought

a speaker sounded more native, correlating to the measure of accentedness, they also found the speaker to be more intelligible. However, listeners still allowed some degree of variance from native speech when judging if the speaker sounded near native.

In a final study, Franklin and Stoel-Gammon (2014) investigated the use of goodness and intelligibility ratings to measure changes in vowel production following accent modification training. Fifteen speakers—five each from Japanese, Korean, and Spanish language backgrounds—were recorded saying tokens in an /hVt/ context (e.g., /hut/) before and after training. Native English listeners then rated goodness of pre- and posttraining tokens on a scale from 1 (*very good*) to 5 (*very poor*). Intelligibility of the same tokens was measured through a forced-choice task in which listeners selected the word that they heard from among 10 options. The results indicated that many vowels improved after training and that goodness and intelligibility of the speakers' productions both showed improvement. However, the relationship between goodness and intelligibility differed across vowels. For example, listeners were able to identify the vowels /e/ and /i/ even when the goodness of the production was determined to be poor. By contrast, the vowels /æ/ and /ʌ/ demonstrated a stronger relationship between intelligibility and goodness ratings, such that intelligibility decreased precipitously as goodness became poorer.

Goodness Ratings

Goodness as a measure of pronunciation has been used to investigate dialectal differences among speakers. For example, Jacewicz, Fox, and Salmons (2006) asked listeners to rate vowel goodness according to three perceptual categories (good, fair, and poor) in a study investigating the effects of prosodic prominence on vowels. In a study by Nishi, Strange, Akahane-Yamada, Kubo, and Trent-Brown (2008), American English listeners rated the category goodness of Japanese vowels according to American English vowel categories. They rated goodness on a Likert scale from 1 (*foreign*) to 7 (*English*). The labels assigned to the anchors of the goodness scale conflate goodness with accentedness without considering a possible distinction between the measures; however, there is no research to either support or refute the idea that goodness and accentedness are parallel measures.

A study by Evans and Iverson (2004) investigated the effect of a listener's first-language accent on goodness ratings for vowels produced by speakers who did and did not share the listener's first-language accent. They investigated listener ratings of category goodness by groups of speakers with northern and southern British accents. They were specifically investigating whether listeners adjusted vowel best-exemplar prototypes to suit the accent they were listening to. The target words of interest were *bath*, *bud*, and *cud*, because the vowels in these words are produced differently in each dialect.

Two experiments were conducted. The first involved two groups of listeners who lived in London, United Kingdom,

at the time of the study. One group spoke with a northern British accent and the other with a southern British accent. The listeners were exposed to two listening conditions. The first consisted of a familiarization passage and carrier phrases spoken by a speaker with a northern accent. Inserted in the carrier phrases were synthesized vowels in monosyllabic target words. Listeners were required to rate the goodness of each vowel relative to the expected target on a scale that ranged from *close* to *far away*. The second listening condition consisted of the same familiarization and rating tasks as in the first condition, but with a southern accent.

The results of this first experiment indicated that the listeners normalized their vowel expectations to suit the accent in which the carrier phrase and familiarization passages were spoken. When listeners with northern British accents heard a southern British speaker, they chose best-exemplar vowels with formants that approximated (but were not identical to) the vowels consistent with the southern accent. Likewise, listeners with southern accents chose best-exemplar vowels that were more consistent with the northern accent when the carrier phrases were spoken with a northern accent. This effect was noted for the words *bud* and *cud* but not for the word *bath*. (For a more detailed explanation of this finding, see Evans & Iverson, 2004.)

The second experiment was identical to the first but involved young listeners (ages 16 and 17 years) who lived in the north (Ashby). The Ashby listeners did not live in a community where they were exposed to the variety of accents spoken in London; however, they were exposed to the more standard southern British accent through media. The findings of the second experiment revealed that these listeners did not normalize their vowel expectations according to the speaker's accent. The young listeners from Ashby chose best-exemplar vowels that were more consistent with their own northern British accent, regardless of the speaker's accent.

Evans and Iverson (2004) interpreted these findings to suggest that listeners can change their perceptual expectations to accommodate different vowels for the same target words within their language. However, this ability may not be a matter of perceptual accommodation alone: It may also involve sociolinguistic factors that affect the ways people try to fit in. For the young Ashby listeners, being exposed to southern British English through the media was not enough. These listeners never had the experience of trying to adjust their own vowels to approximate those of a wider community. As a result, their inflexibility in identifying best-exemplar tokens for other accents may reflect their inexperience in attempting to produce these best-exemplar tokens themselves.

Listener Bias

In their study on goodness and intelligibility, Franklin and Stoel-Gammon (2014) stated that they chose to rate vowel goodness rather than accentedness to avoid possible negative connotations associated with the word "accent." They also wanted to avoid possible listener bias that may be associated with preferences for some accents over others.

Listeners associate a variety of perceptions, biases, and stereotypes with accented speech. Some accents are perceived as more favorable or even sophisticated, whereas others are associated with decreased comprehensibility simply because of the presence of an accent (Derwing & Munro, 2009). In addition, the preconceptions and negative feelings a listener has toward accents may cause the listener to rate a successful interaction with a nonnative speaker as unsuccessful (Lindemann, 2002). Listeners are also more likely to rate nonnative speakers as having negative traits, as was investigated in another study by Lindemann (2003). Listeners were asked to rate traits (e.g., intelligent, successful, lazy, friendly, uneducated, unkind) of native English speakers and Korean ELLs after hearing a reading passage. The listeners were then asked to identify whether they thought the speaker was a native English speaker and what the speaker's ethnicity was, if not English. Although many listeners were unable to identify the ethnicity of the speakers, they rated traits of the nonnative speakers more negatively than those of the native English speakers, thus demonstrating the effect of accent bias.

In a study by Delamere (1996), listeners rated personal characteristics of ELLs (e.g., likability, self-confidence, good character, intelligence) on the basis of connected speech samples of five speakers from different language backgrounds (Arabic, Farsi, French, Malay, and Spanish). Each speaker produced a speech sample without grammatical errors and one with grammatical errors typical of intermediate-proficient speakers of English as a foreign language. When the speech sample contained errors, listeners rated personality traits similarly across each of the five language backgrounds, therefore seeing all nonnative speakers in a similar way. However, when no grammatical errors were present, personality trait ratings varied by language background. This outcome may suggest that listeners hold a stereotype or bias toward particular nonnative accents more than toward grammatical errors.

Derwing and Munro (1997) conducted a study in which they asked 26 listeners to identify the native language (NL) background of nonnative English speakers from four alternatives: Spanish, Polish, Japanese, and Cantonese. Before participating in the NL identification task, listeners reported the degree of familiarity they had with languages other than English and the amount of interaction they had with Spanish-, Polish-, Japanese-, and Cantonese-accented English. The listeners were presented with connected speech samples by speakers from the aforementioned NL backgrounds. The results revealed that listeners were unable to identify the correct NL background of the speakers in about 48% of opportunities and that no one language was more difficult to identify than another. However, listener familiarity with a speaker's accent had the most significant influence on NL background identification accuracy. In addition, the researchers found that intelligibility scores were higher when listeners were able to accurately identify the NL. Another study revealed that familiarity with a particular accent was found to cause listeners to rate accents more leniently on oral-proficiency interview tests (Carey, Mannell, & Dunn, 2011).

Purpose of the Present Study

The purpose of the present study is to compare goodness and accentedness ratings of speech tokens among naive and aware listeners. The comparison of goodness with accentedness will reveal whether listeners view these measures differently. This comparison will also extend the research of Franklin and Stoel-Gammon (2014), who previously evaluated the relationship between goodness and intelligibility. Investigating the relationship between accentedness and goodness can provide valuable information regarding appropriate outcome measures for accent modification training. As a final note, comparing the responses of naive listeners with the responses of listeners who are aware of the speakers' language backgrounds will further elucidate the effect of listener bias on ratings of accented speech. If the findings reveal that there is a larger effect of listener bias on goodness compared with accentedness (or vice versa), they will further inform clinically appropriate outcome measures for accent modification. The study addresses the following research questions:

1. Do listeners rate goodness of /hVt/ tokens differently than they rate accentedness?
2. Do listeners who are aware of the speaker's language background rate accentedness and goodness differently than listeners who are naive to the language background?

Method

The institutional review board at the University of Washington approved the speaker protocol prior to recruitment activities. The study protocol for listeners was approved by Miami University's institutional review board prior to recruitment activities.

Speakers

The speech stimuli presented in this study were previously recorded for a study by Franklin and Stoel-Gammon (2014). For that study, 20 speakers (15 adult ELLs and five adult native English speakers) were recruited from the University of Washington and surrounding areas. The ELLs were native speakers of Japanese, Korean, and Spanish who had begun learning English after the age of 12 years. The five adult native English speakers spoke with a dialect that was consistent with Pacific Northwest English. All 20 speakers self-reported typical speech, language, and hearing.

For the present listening study, recorded speech stimuli from three of the five speakers in each language group were selected for presentation. The authors determined that a total of 12 speakers would provide a sufficient number of speech tokens with which to conduct the present study. In determining which 12 speakers to select from among the original 20, the authors' primary concern was to include at least one male and one female speaker from each language group. Previous research has determined that obvious voice characteristics, such as nasality, loudness, or breathiness can lead to bias in listener judgments (Derwing & Munro,

2005; Esling & Wong, 1983). Therefore, one of the female Spanish speakers was excluded because she spoke with very noticeable glottal fry, and one of the male Spanish speakers was excluded because his voice was determined to be noticeably breathy. None of the speakers from the other language backgrounds were excluded because of noticeable voice-quality characteristics. Table 1 presents the gender, age, and age of learning of the speakers from the Franklin and Stoel-Gammon (2014) study whose speech tokens were used in the present study.

Recordings and Stimuli Preparation

Speech stimuli were recorded at the University of Washington's Speech and Hearing Clinic in a sound-attenuated room. Recordings were made onto a Sony (Tokyo, Japan) laptop using an Audio-Technica (Tokyo, Japan) ATM75 condenser head-mounted microphone placed approximately 3 in. from the speaker's mouth. Praat 4.1.27 (Boersma, 2002) was used to record the speakers at a sampling rate of 44.1 KHz. An Apogee Electronics (Santa Monica, CA) Mini-Me digital-to-analog converter was used to digitize the signal at a sampling rate of 44.1 KHz with a resolution of 16 bits. The Mini-Me soft-limiting setting was applied to prevent peak clipping. The sound files were saved in .wav format and normalized using Audacity 2.0.6 (<http://www.audacityteam.org>) to ensure a consistent presentation volume across all items.

The speakers were recorded as they produced /hVt/ tokens (Peterson & Barney, 1952) for each of 10 English vowels (/i/, /I/, /e/, /ɛ/, /æ/, /ʌ/, /u/, /ʊ/, /o/, /a/), before and after pronunciation training, in the carrier phrase "I can say _____ again _____." Using visual information from the acoustic waveform as well as auditory cues, the /hVt/ stimuli were segmented from the carrier phrase using Praat 4.5.21. Segmentation began at the onset of /h/ frication and ended at the release of the final /t/, which was marked by the point of aspiration on the acoustic waveform. If the final /t/ was not released, segmentation ended at the point of oral closure after the vowel, which was marked by the end of periodicity and lack of acoustic energy on the waveform and auditory silence. The speakers produced each /hVt/ token a total

Table 1. Speaker demographic information.

Language	Gender	Age (years)	Age (years) of learning
English	F	33	
English	M	24	
English	F	21	
Japanese	F	37	13
Japanese	F	37	13
Japanese	M	25	13
Korean	M	25	13
Korean	F	46	13
Korean	M	32	13
Spanish	M	42	40
Spanish	F	34	13
Spanish	F	33	23

of 20 times in the original Franklin and Stoel-Gammon (2014) study.

Selection of /hVt/ Targets to Be Rated by Listeners

Of the 10 /hVt/ tokens produced by the speakers, seven constituted real words and three were nonsense words. The seven real words were /hæt/ (*hat*), /hut/ (*hoot*), /hat/ (*hot*), /hit/ (*heat*), /hʌt/ (*hut*), /hit/ (*hit*), and /het/ (*hate*). The three nonsense words were /hʊt*/ (rhymes with *foot*), /het*/ (*het*), and /hot*/ (*hote*). The nonsense words were excluded as targets because listeners may rate nonsense target words differently than real words. Of the seven real words, the following five were used as targets to be judged by listeners: /hit/, /hit/, /het/, /hæt/, and /hat/. The targets /hut/ and /hʌt/, though real words, occur less frequently in spoken English than the other five words (Davies, n. d.).

Selection of /hVt/ Prototypes to Be Presented to Listeners

For each of the 12 speakers, the authors identified a single best-exemplar prototype for each of 10 English vowels produced in the /hVt/ context. Acoustic analysis of all /hVt/ productions for the sake of stimulus selection was deemed to be beyond the scope of the study. Rather, the second and third authors perceptually identified and selected the best-exemplar prototypes from among each speaker's 20 productions by considering three factors. First, the vowel was judged perceptually for accuracy to the intended target. Next, the /h/ productions were judged auditorily; productions that were inaudible or produced as a velar fricative were not preferred for the prototype. As the last factor, whenever possible, /hVt/ productions with a released /t/ were favored over those with unreleased codas. Vowel accuracy was the priority in identifying best-exemplar prototypes, followed by the quality of the /h/ and lastly the release of the /t/. Not every best-exemplar prototype had all three preferred elements. Of the 120 productions (12 speakers × 10 prototypes), 85 (70.83%) had all three favorable elements. The percentage of prototypes with a velar fricative instead of /h/ was 7.5%, and the percentage with an unreleased /t/ was 10%. The remaining 14 productions, or 11% of the /hVt/ prototypes, did not have the exact target vowel but were the best option from among their speakers' 20 productions. For example, a speaker's target /i/ may have been slightly lowered but not lowered enough to be considered an /ɪ/ perceptually.

Selection of "Incorrect" /hVt/ Substitutions

Three "incorrect" /hVt/ productions were selected as substitutions for each of the five target words. These incorrect substitutions were chosen from among each speaker's 10 best-exemplar prototypes. The three substitutions for each target were selected on the basis of common vowel-error patterns produced by ELLs from Spanish, Korean, and Japanese language backgrounds (Celce-Murcia, Brinton, & Goodwin, 2010; Compton, 1996). For example, many NL vowel inventories do not include the lax vowel /ɪ/ (Maddieson, 1984). This observation is true for Japanese, Korean, and Spanish, as well as many other languages. ELLs from such NL backgrounds often produce English words containing

/ɪ/ with the tense vowel /i/ (e.g., /hit/ → /hit/; Celce-Murcia et al., 2010; Compton, 1996). Therefore, /i/ was selected as one of the incorrect substituted vowels for the target /i/ in the present study.

The vowels /e/ and /i/ are common among most languages in the world (Maddieson, 1984), and are rarely substituted with English vowels other than, respectively, /ɛ/ and /ɪ/ (Compton, 1996). However, for the vowels /e/ and /i/ the authors added an additional two front unrounded vowels as substitutions, in order to complete the set of three incorrect productions per target. The same substitutions were chosen for all 12 speakers. This approach ensured that the listeners heard very similar incorrect vowels produced by each speaker, regardless of the speaker's language background. Table 2 presents the substitutions selected for each of the real target words.

Listeners and Listening Conditions

Twenty-eight listeners were recruited from Miami University to participate in the listening tasks. The average age of the listeners was 20.7 years. The 28 listeners comprised one man and 27 women. All students at Miami University must complete four semesters of a foreign-language requirement prior to graduation. Therefore, all of the listeners had taken at least one semester of a foreign language when they participated in the study. Listeners were not excluded from participation on the basis of the number of foreign-language semesters completed; however, they were excluded if they were fluent in Japanese, Korean, or Spanish. By self-report, all listeners were native speakers of English and none were fluent in Japanese, Korean, or Spanish.

All of the listeners spoke a dialect of English that was consistent with the North Midland region. Listeners were excluded from the study if they had taken, or were currently enrolled in, a phonetics course. This exclusion criterion was included to ensure participation by individuals who were untrained and unfamiliar with phonology and the study of accented speech patterns. Listeners were excluded if they had familiarity with the International Phonetic Alphabet.

After providing consent to participate in the study, each listener was randomly assigned to either an aware listening condition or a naive listening condition. Listeners in the aware condition were made aware of the NL background of a speaker before they rated each /hVt/ target. Listeners in the naive condition rated each target without knowing the NL background of the speaker who produced the targets. There were 14 listeners assigned to each listening condition.

Table 2. Substituted vowels selected for each /hVt/ target.

Target	Substituted vowels			
/i/ (heat)	/ɛ/ (het)	/e/ (hate)	/ɪ/ (hit)	
/ɪ/ (hit)	/ɛ/ (het)	/e/ (hate)	/i/ (heat)	
/e/ (hate)	/ɛ/ (het)	/i/ (heat)	/ɪ/ (hit)	
/æ/ (hat)	/a/ (hot)	/ɛ/ (het)	/ʌ/ (hut)	
/a/ (hot)	/ʌ/ (hut)	/æ/ (hat)	/o/ (hote)	

Listening Tasks

Listeners participated in two tasks: a goodness rating task and an accentedness rating task. Both tasks were prepared using Alvin 2 experimental-design software (Hillenbrand & Gayvert, 2005). Each task was completed on a different day, and the order of task assignment on the first or second day of participation was randomized across all listeners. Half of the listeners in each listening condition participated in the goodness rating task on the first day, and the other half participated in the accentedness rating task on the first day.

For both the goodness and accentedness tasks, each listener was presented with five listening blocks, with each block representing one of the five /hVt/ target words. Within each block, a total of 48 /hVt/ tokens was presented to the listener: 12 speakers (three from each language background) $\times$ 4 productions per speaker. The four productions per speaker comprised an accurate production of each target vowel as well as each of the three substituted vowels identified in Table 2. For both listening tasks and across all listeners, block presentation was randomized, as was the order of /hVt/-token presentation within each block.

Accentedness Rating Task

The accentedness rating task required listeners to rate the accentedness of each vowel on a scale from 1 (*no accent*) to 9 (*very strong accent*). Previous research determined that a 9-point Likert scale effectively captured listener responses to accented speech (Munro & Derwing, 1994, 1995a, 1999b). Listeners were seated in front of a desktop computer and presented with the visual display for the first listening block. They were instructed to listen to /hVt/ words and rate the accentedness of each word. They were also told that tokens within each block comprised speakers attempting to say a particular /hVt/ word.

The visual display presented the instructions and an orthographic representation of the target word for that block—e.g., “The target word is *hat*. Rate the accentedness of each production.” Below the target word and instructions, the Likert scale was displayed with the descriptors *no accent* and *very strong accent* presented below the 1 and 9 anchors, respectively. The target word was displayed throughout the entirety of the block so listeners could remain focused on the target. In order to capture first impressions, listeners were not permitted to replay tokens. Once a token was rated, the next token was automatically presented. When one block was completed, the next block was automatically presented. A new block was signaled by a new target label and a change in background color on the visual display.

The listeners in the aware condition completed an accentedness task that was similar to the task in the naive condition. However, in addition to seeing the target word for a particular block, aware listeners also saw a text box that alerted them to the NL background of each token’s speaker—e.g., “This speaker’s native language is _____.” Before the token was played, aware listeners were required to click a button labeled “Go” that was displayed next to the NL text box. Requiring listeners to click “Go” provided

extra assurance that they noted the speaker’s NL prior to hearing each token. The NL label remained on the screen until the speaker rated the token, after which the next trial’s NL label was presented. Figure 1 represents the visual display presented to listeners in the aware group. Although a replay button is shown, listeners in both conditions were instructed not to select this button at any time during the listening tasks. The study’s results data file confirmed that listeners did not replay tokens.

Goodness Rating Task

The goodness rating task was structured identically to the accentedness task for the naive and aware listening groups. The listening instructions were the same as for the accentedness rating task; however, listeners were now instructed to rate how good each production was. Goodness was rated on a scale from 1 (*very good example*) to 9 (*very poor example*).

Open-Ended Questions

At the end of the study, listeners answered the following two open-ended questions:

1. What were you listening for when you were rating the accentedness of each word?”
2. What were you listening for when you were rating the goodness of each word?”

The verbal responses were written down by research assistants. The authors were concerned that posing an open-ended question after the listener’s first session would bias listener ratings during the second session; therefore, both questions were presented to the listeners after the second listening session was complete.

Results

Data Set and Reliability

Listeners were presented with five different blocks of /hVt/ tokens. Each block represented a different target vowel

Figure 1. Visual display presented to listeners in the aware listening group.

The target word is: Hit

This speaker's native language is: Spanish

Rate the accent of each production

Go

1 2 3 4 5 6 7 8 9

no accent very strong accent

Pause Replay Back Up

and comprised 48 tokens (12 speakers $\times$ 4 tokens). As a result, each listener rated a total of 240 /hVt/ tokens (48 tokens $\times$ 5 blocks). There were 14 listeners in each listening condition (naive and aware). Therefore, the goodness and accentedness group statistics should be based on a total of $N = 3,360$ observations for each listening condition (240 /hVt/ tokens $\times$ 14 listeners). In the aware data set, 16 accentedness ratings were missing from the results file (giving $N = 3,344$), as were eight goodness ratings ($N = 3,352$). The number of missing observations from the aware data set was small compared with the total sample size. No observations were missing from the naive data set.

Intraclass correlation coefficients (ICCs) for consistency were determined for both goodness and accentedness as a function of listening condition. ICCs were determined for consistency rather than absolute agreement because goodness and accentedness ratings are subjective. Following are the average ICC results on the basis of a mixed-effects model: aware accentedness: $ICC = .954, p < .001$; aware goodness: $ICC = .969, p < .001$; naive accentedness: $ICC = .952, p < .001$; naive goodness: $ICC = .963, p < .001$.

Group Statistics

For the first research question (goodness vs. accentedness ratings), the results will be presented across language groups. This will be followed by the results for the second research question (aware vs. naive listeners). In the final subsection, goodness and accentedness ratings from naive and aware listeners will be presented as a function of the speakers' NL backgrounds.

Results for Question 1: Goodness Versus Accentedness Ratings

Table 3 displays the mean ratings of accentedness and goodness for both aware and naive listeners by target vowel. These mean values are based on listener ratings of the correct target vowel as well as the three incorrect substituted vowels. As a result, the mean values in Table 3 tend to hover around the middle of the rating scales. There is a significant positive correlation between ratings of

Table 3. Mean (standard deviation) ratings of accentedness and goodness by target vowel.

Target vowel	Naive listeners ^a		Aware listeners ^b	
	Accentedness	Goodness	Accentedness	Goodness
/ɪ/	5.23 (2.89)	5.26 (2.93)	4.92 (2.61)	4.98 (2.89)
/i/	5.71 (2.66)	5.63 (2.85)	5.12 (2.58)	5.39 (2.83)
/e/	5.32 (2.76)	5.91 (2.92)	5.34 (2.73)	5.42 (2.81)
/æ/	5.06 (2.73)	5.09 (2.78)	4.98 (2.50)	4.72 (2.61)
/ɑ/	5.01 (2.67)	5.17 (2.94)	4.81 (2.60)	4.81 (2.78)
Mean	5.27 (2.75)	5.41 (2.90)	5.03 (2.60)	5.06 (2.80)

^a $r^2 = .714, p < .001$. ^b $r^2 = .693, p < .001$.

accentedness and goodness. As goodness ratings improve, accentedness is also rated closer to native. For naive listeners, the correlation between accentedness and goodness was $r^2 = .714, p < .001$. For aware listeners, the correlation between the two ratings was $r^2 = .693, p < .001$.

There was an overall trend for listeners to rate goodness poorer than accentedness. Within the aware group, the mean accentedness rating across all vowels was 5.03 ($SD = 2.60$) and the mean goodness rating was 5.06 ($SD = 2.80$). A two-tailed t test revealed that the difference between the means for aware listeners failed to reach statistical significance ($p = .66$). However, within the naive listening group, the difference between mean accentedness (5.27, $SD = 2.75$) and mean goodness (5.41, $SD = 2.90$) did reach statistical significance ($p = .036$).

Table 4 demonstrates the mean accentedness and goodness ratings by target vowel (columns) and presented vowel (rows) for aware listeners. Scores for accurate target productions are indicated in bold along the diagonal of the table. Among the five accurate target productions, the general trend among aware listeners was to rate accentedness poorer than goodness. However, only the accurate production of /æ/ demonstrated a statistically significant difference ($p < .05$) between accentedness and goodness among aware listeners. Three of the incorrect /hVt/ substitutions

Table 4. Confusion matrix, for aware listeners, of mean accentedness (A) and goodness (G) ratings by vowel.

Presented vowel	Target vowel				
	/i/	/ɪ/	/e/	/æ/	/ɑ/
/i/					
A	1.79	5.38	7.14		
G	1.63	6.18	7.01		
/ɪ/					
A	6.22 ^a	1.83	6.40		
G	6.38 ^a	1.56	6.61		
/e/					
A	6.07 ^a	7.06	1.49		
G	6.85 ^a	7.32	1.46		
/ɛ/					
A	6.44	4.95	6.31	5.12 ^a	
G	6.67	4.87	6.62	4.57 ^a	
/æ/					
A				2.30^a	6.53
G				1.97^a	6.40
/ɑ/					
A				6.38	2.19
G				6.36	1.97
/ʌ/					
A				6.11	4.66
G				5.99	4.93
/o/					
A					5.86
G					5.92

Note. Accentedness is on a scale from 1 (No accent) to 9 (Very strong accent). Goodness is on a scale from 1 (Very good example) to 9 (Very poor example).

^aStatistically significant difference between accentedness and goodness ($p < .05$).

demonstrated a significant difference ($p < .05$) between accentedness and goodness among aware listeners: /hit/ → /hit/, /hit/ → /het/, and /hæt/ → /het/. In each case, goodness was rated poorer than accentedness. Among the 15 incorrect /hVt/ substitutions rated by aware listeners, there was no overall trend toward either rating goodness poorer than accentedness or vice versa.

Table 5 depicts the mean accentedness and goodness ratings by target vowel (columns) and presented vowel (rows) for naive listeners. Scores for accurate target productions are indicated in bold along the diagonal of the table. When the target was accurate, the trend among naive listeners was to rate accentedness poorer than goodness. This pattern was seen previously with the aware listeners. However, of the five accurate targets, only the accurate productions of /i/ and /e/ demonstrated statistically significant differences ($p < .05$) between accentedness and goodness ratings.

Among the naive listeners, there was an overall trend to rate goodness poorer than accentedness among incorrect /hVt/ substitutions. Of the 15 incorrect substitutions, 13 were assigned poorer goodness ratings than accentedness ratings. The following four incorrect /hVt/ substitutions demonstrated a significant difference ($p < .01$) between accentedness and goodness: /het/ → /hit/, /het/ → /hit/, /het/ → /hæt/, and /hat/ → /hæt/.

Table 5. Confusion matrix, for naive listeners, of mean accentedness (A) and goodness (G) ratings by vowel.

Presented vowel	Target vowel				
	/i/	/ɪ/	/e/	/æ/	/a/
/i/					
A	2.42^a	6.38	6.92 ^a		
G	1.70^a	6.21	7.88 ^a		
/ɪ/					
A	6.78	1.93	6.15 ^a		
G	7.18	1.76	7.10 ^a		
/e/					
A	6.53	7.52	1.87^a		
G	6.42	7.80	1.61^a		
/ɛ/					
A	7.13	5.07	6.35 ^a	4.65	
G	7.22	5.26	7.06 ^a	4.88	
/æ/					
A				2.50	6.39 ^a
G				2.15	7.04 ^a
/a/					
A				6.84	2.33
G				7.04	2.02
/ʌ/					
A				6.25	5.08
G				6.30	5.35
/o/					
A					6.26
G					6.25

Note. Accentedness is on a scale from 1 (*No accent*) to 9 (*Very strong accent*). Goodness is on a scale from 1 (*Very good example*) to 9 (*Very poor example*).

^aStatistically significant difference between accentedness and goodness ($p < .05$).

Listener Responses to Open-Ended Questions

After the second listening session, listeners were asked to answer the following open-ended questions:

1. What were you listening for when you were rating the accentedness of each word?
2. What were you listening for when you were rating the goodness of each word?

After listeners provided an answer, the researchers followed up by asking, “Anything else?” Additional comments, if any, were noted. The following five variables emerged in the listener responses: vowels, consonants, “My own pronunciation,” “How well I could understand,” and “Does it match the target?” Table 6 summarizes the percentage of listeners who identified these variables as influencing their goodness and accentedness ratings. There was no distinction in the types of verbal responses given by listeners in the naive group compared with listeners in the aware group. Therefore, Table 6 presents group data from 24 listeners.

The variable with the highest percentage of listener responses was “My own pronunciation” of a target. Roughly half of the listeners identified this variable when rating both accentedness and goodness. An equal percentage of listeners for each rating (20.8%) identified vowel accuracy as influencing goodness and accentedness ratings. However, it is important to note that individual listeners identified vowels as influencing either goodness ratings or accentedness ratings but not both. The variable that demonstrated the largest discrepancy with respect to influence on goodness versus accentedness ratings was “How well I could understand.” Only 8.3% of listeners stated that understanding the target influenced goodness ratings, whereas 25.0% stated that it influenced accentedness ratings. Another variable that emerged from listener responses was whether the production matched the target. “Does it match the target?” was reported as a factor in rating goodness and accentedness, respectively, by 29.2% and 16.7% of listeners.

Considering the variables “My own pronunciation,” “How well I could understand,” and “Does it match the target?” over 90% of listeners mentioned only one of the three in their verbal responses. This was true for both goodness and accentedness ratings. Therefore, these three variables may reflect different phrases the listeners used to express the same thing, namely overall accuracy of the intended /hVt/ token. Recall that listeners saw an orthographic representation of the target word during the rating tasks but were not presented with an audio example as a gold standard. Therefore, the target referenced in the statement “Does it match the target?” is likely based on the speaker’s own pronunciation.

Results for Question 2: Naive Versus Aware Listeners

Analysis by Vowel

Figures 2 and 3 depict, respectively, the differences in mean ratings between naive and aware listeners, by vowel,

Table 6. Percentage of listeners identifying particular variables influencing /hVt/ ratings.

	Accentedness	Goodness
Variables identified from verbal responses		
Vowel	20.8%	20.8%
Consonants	8.3%	12.5%
"My own pronunciation"	50.0%	54.2%
"How well I could understand"	25.0%	8.3%
"Does it match the target?"	16.7%	29.2%

for accentedness and goodness. The numerical values for each target and presented vowel are presented in Tables 4 and 5.

A comparison of naive and aware listeners' accentedness ratings (see Figure 2) reveals that naive listeners' ratings were poorer than those of aware listeners for most /hVt/ tokens presented. The mean accentedness rating across all tokens was 5.03 ($SD = 2.60$) for aware listeners and 5.27 ($SD = 2.75$) for naive listeners (see Table 3). A two-tailed t test revealed that this difference is statistically significant ($p < .001$). Following are the /hVt/ tokens for which there was a statistically significant difference ($p < .05$) in accentedness ratings between aware and naive listeners: the accurate production of the target /hit/, /hit/ → /hæt/, /hit/ → /het/, /hit/ → /hɪt/, /hit/ → /hɛt/, /hit/ → /hɪt/, the accurate production of /het/, and /hæt/ → /hat/.

A comparison of naive and aware listeners' goodness ratings (see Figure 3) reveals that naive listeners assigned poorer ratings than aware listeners for almost all /hVt/ tokens except /hit/ → /het/. The mean goodness rating across all tokens was 5.06 ($SD = 2.80$) for aware listeners and 5.41 ($SD = 2.90$) for naive listeners (see Table 3). A two-tailed t test revealed that this difference is statistically significant

($p < .001$). Following are the /hVt/ tokens that demonstrated statistically significant differences ($p < .05$) in goodness ratings between aware and naive listeners: /hit/ → /hæt/, /hit/ → /hɪt/, /hit/ → /het/, /het/ → /hæt/, /het/ → /hɪt/, /het/ → /het/, /hæt/ → /hat/, and /hat/ → /hæt/.

Analysis by Speaker NL Background

Table 7 presents mean goodness and accentedness ratings by speaker NL for aware and naive listeners. Two-tailed t tests were used to determine statistical significance.

For tokens produced by native English speakers, there was a significant difference between aware and naive listeners on accentedness ratings but not goodness ratings. Aware listeners assigned better accentedness ratings to tokens produced by native English speakers than did naive listeners. For tokens produced by native Spanish, Korean, and Japanese speakers, there was a significant difference between aware and naive listeners' goodness ratings. The only nonnative group for which aware and naive listeners demonstrated a significant difference on accentedness ratings was Japanese speakers. In all of the aforementioned cases, aware listeners assigned better ratings than naive

Figure 2. Mean accentedness ratings by target vowel and substituted vowel for aware and naive listeners. *Statistically significant differences between aware and naive listeners' ratings ($p < .05$). 1 = No accent, 9 = Very strong accent.

Figure 3. Mean goodness ratings by target vowel and substituted vowel for aware and naive listeners. *Statistically significant differences between aware and naive listeners' ratings ($p < .05$). 1 = *Very good example*, 9 = *Very poor example*.

listeners. Table 7 also illustrates that the mean goodness rating among aware listeners was 5.30 for English tokens. This rating was poorer than the aware listeners' mean goodness ratings for Spanish, Korean, and Japanese tokens (5.04, 4.99, and 4.92 respectively). As a final matter, when differences between goodness and accentedness ratings were analyzed by NL group, the only instance where the difference was statistically significant was when aware listeners judged native English tokens.

Discussion

This study investigated goodness and accentedness ratings of /hVt/ tokens between two groups of listeners:

those who were aware of the speakers' NL backgrounds and those who were naive to them. The results of this study lend insight into the differences between goodness and accentedness and the effect of bias on listeners' ratings.

Distinguishing Goodness From Accentedness

A high positive correlation exists between goodness and accentedness. Tokens that were rated as being less accented were also rated as being good examples of the target.

Although the positive correlation between goodness and accentedness is high and statistically significant for

Table 7. Aware and naive listeners' mean (standard deviation) goodness and accentedness ratings by speaker native language.

Native language		Goodness	Accentedness	Significant difference between goodness and accentedness?
English	Aware	5.30 (2.97)	4.99 (2.70)	Y*
	Naive	5.41 (2.90)	5.20 (2.80)	N
	Difference	$p = .43$	$p = .027^*$	
Spanish	Aware	5.04 (2.79)	5.12 (2.63)	N
	Naive	5.34 (2.86)	5.27 (2.72)	N
	Difference	$p = .0028^{**}$	$p = .25$	
Korean	Aware	4.99 (2.67)	5.07 (2.47)	N
	Naive	5.48 (2.93)	5.28 (2.72)	N
	Difference	$p = .0003^{**}$	$p = .09$	
Japanese	Aware	4.92 (2.74)	4.95 (2.63)	N
	Naive	5.41 (2.91)	5.22 (2.77)	N
	Difference	$p = .0004^{**}$	$p = .042^*$	

Note. Bold denotes differences between aware and naive listeners that reached statistical significance on the basis of paired two-tailed t tests.

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

both naive and aware listeners, goodness ratings are not perfect predictors of accentedness ratings. Just as previous research has demonstrated that intelligibility, comprehensibility, and accentedness are related but not equivalent measures (Behrman & Akhund, 2013; Derwing & Munro, 1997; Munro & Derwing, 1995b; Trofimovich & Isaacs, 2012), the same is true for the relationship between goodness and accentedness. The results of the present study call into question the rating scale used by Nishi et al. (2008), in which the anchors of the 7-point “goodness” scale were *foreign* and *English*. The use of these two descriptors on a goodness scale conflates goodness with accentedness. The present study’s findings emphasize the importance of treating accentedness and goodness as distinct measures.

The difference between mean goodness and accentedness ratings across all tokens was not statistically significant among aware listeners but did reach statistical significance among naive listeners. It is possible that because aware listeners assigned better ratings overall, the margin between goodness and accentedness was too narrow for statistical significance to be reached. The results revealed a tendency for listeners to assign poorer goodness than accentedness ratings in most cases where the presented token was clearly not the intended target. However, when listeners were presented with the correct production of a target, they assigned goodness ratings that were better than their accentedness ratings. This last point was true across all accurate productions for both naive and aware listeners. Together, these observations suggest that goodness is more reflective of vowel accuracy than is accentedness.

In Franklin and Stoel-Gammon’s work (2014) examining the relationship between intelligibility and goodness of English vowels, they chose to have participants rate goodness rather than accentedness to avoid possible negative connotations associated with the term “accent.” Further, they stated, “Goodness provides another way to rate vowel accuracy without explicitly using the term *accent*” (p. 4). The results of the present study support this proposition. However, it must be noted that listeners in Franklin and Stoel-Gammon’s study were tasked with rating the goodness of vowels in the /hVt/ context. By contrast, listeners in the present study were asked to rate the goodness of the entire /hVt/ token.

When listeners were asked open-ended questions about what influenced their ratings, there was overlap in the variables associated with goodness and accentedness ratings. Most of the responses reflected the importance of /hVt/-token accuracy in rating both goodness and accentedness. However, the concept of understanding the speaker was invoked more often for accentedness ratings, whereas the concept of matching the target was invoked more often for goodness ratings. These different concepts might simply be alternative ways for average, untrained listeners to refer to overall target accuracy. However, it is also possible that “accent,” a speech concept familiar to the average listener, is readily associated with difficulty understanding an adult speaker. By contrast, “goodness,” a speech concept that is relatively unfamiliar to average listeners, may be more

readily associated with the less familiar and less naturalistic task of matching a production to a target.

On a final note, 20% or fewer of the listeners referenced consonants or vowels as variables that affected their ratings. We cannot, however, conclude from this result that listeners who did not identify segments as factors were not influenced by them when rating accentedness and goodness; rather, these results suggest that the majority of the listeners did not report a conscious awareness of basing their ratings on the accuracy of consonants and/or vowels.

Differences Between Naive and Aware Listeners

The results of the present study demonstrate a general trend for aware listeners to assign better accentedness and goodness ratings than naive listeners. One possible interpretation of these findings is that aware listeners were making a conscious effort not to appear judgmental toward accented speech. However, this interpretation does not explain why aware listeners assigned better ratings to English tokens than did naive listeners. An alternative interpretation is that aware listeners assumed more of the communicative burden than naive listeners, leading to more favorable ratings. This second interpretation is supported by a study conducted by Lindemann (2002). In that study, listeners were more likely to rate communicative interactions with non-native speakers as successful when they shared the responsibility of the interaction and when they had a positive attitude toward nonnative speakers. However, because listeners in the present study were not asked about their attitudes toward nonnative speakers, we cannot draw any conclusions about their language attitudes and the subsequent effects on ratings. Nevertheless, it is possible that simply knowing the NL background of each speaker allowed the aware listeners to rate each production more favorably.

Listeners are perceptive at detecting when someone’s speech differs from that of their own language community. A study by Munro, Derwing, and Burgess (2010) demonstrated that listeners were able to detect accented speech that was presented backward and in unknown contexts. Lindemann (2003) found that listeners assigned negative stereotypes to nonnative speakers even when they could not identify the speakers’ ethnicity. On the basis of these findings, we might expect that naive listeners in the present study would have been able to distinguish native English tokens from accented tokens. If naive listeners were indeed able to clearly identify tokens produced by native English speakers, then we would predict no difference between naive and aware listeners’ ratings of the native English tokens. This prediction was substantiated for goodness ratings but not accentedness ratings. Recall that for each of the five /hVt/ targets, one correct and three best-exemplar “incorrect” tokens were presented to listeners. This design ensured that listeners heard similar incorrect tokens produced by each speaker, regardless of the speaker’s language background. This design, and the use of short /hVt/ tokens rather than

connected speech samples, may have made it difficult for naive listeners to clearly recognize tokens produced by native English speakers.

When naive and aware ratings were analyzed by speaker NL background, several interesting observations emerged. First, when results were compared across NL backgrounds, aware listeners rated the goodness of tokens produced by native English speakers most severely. The same was not true for aware listeners' accentedness ratings of English tokens compared with nonnative tokens. In addition, aware listeners rated the accentedness of English speakers better than did naive listeners. This finding was statistically significant. At the same time, the difference in goodness ratings between aware and naive listeners failed to reach statistical significance for native English tokens. The only NL condition that demonstrated a significant difference between goodness and accentedness was aware listeners' ratings of English tokens.

Three questions of interest emerge from these data: Why did the aware listeners assign poorer goodness ratings to tokens produced by native English speakers compared with tokens produced by nonnative speakers? Why was this pattern not reflected in the aware listeners' accentedness ratings of native English tokens? And why was English the only NL condition that demonstrated a significant difference between aware listeners' goodness and accentedness ratings? The answers to these questions can be understood if we incorporate the present study's earlier conclusion, that goodness reflects target accuracy whereas accentedness reflects a quality of pronunciation that resides above the level of accuracy. In this regard, it may be possible for a production to be a good example of a target but still be accented. On the converse, it should also be possible for a production to be a poor example of a target without sounding as if the speaker has a very strong accent.

The aware listeners rated the goodness of native English tokens more poorly than the goodness of nonnative tokens because the listeners' expectations of accuracy were likely higher for native English speakers. Therefore, an incorrect production by a native English speaker would have been treated more critically than a similar incorrect production by a nonnative speaker. At the same time, when aware listeners rated accentedness, they were likely biased by the expectation that native English speakers should not have strong accents. As a result, they were less likely to rate incorrect English tokens as more heavily accented than similar incorrect nonnative tokens. By contrast, the naive listeners were free to assign accentedness and goodness ratings primarily, if not solely, on the basis of the perceived differences between the intended target and presented tokens. The aware listeners' bias regarding native English speakers ultimately pulled their goodness and accentedness ratings of English tokens in opposing directions. The poorer goodness ratings of English tokens, combined with better accentedness ratings, yielded a significant difference between goodness and accentedness among aware listeners that was not observed in any of the other combinations of NL and listening group.

Clinical Implications

This study reveals that goodness and accentedness ratings are vulnerable to bias on the basis of listener expectations for different NL groups. Clinicians and researchers should be mindful of this when they design outcome measures assessing either of these constructs. Bias can be minimized in conditions where listeners are naive to the NL of the speaker. Listeners may recognize that a token is a good exemplar even when a foreign accent is noticeable. Although accentedness is still an important and relevant outcome measure, goodness may be more appropriate for assessing the accuracy of speech targets.

Limitations and Future Directions

Unlike global measures of pronunciation proficiency, such as intelligibility, accentedness, and comprehensibility, goodness may not have much utility beyond individual segments or words. As a result, goodness ratings may prove to be less ecologically valid than measures that can be applied to larger linguistic units such as phrases and sentences.

The listeners answered open-ended questions about goodness and accentedness at the end of their second listening session rather than after each session. The study was designed this way to avoid biasing the listeners' ratings during their second session. However, this strategy likely introduced another unintended artifact. Because listeners were asked about goodness and accentedness in the same session, they may have framed their answers to create contrast between the two measures. It is possible that if one open-ended question had been posed at the end of each session, the listeners might have provided answers that demonstrated more overlap between goodness and accentedness ratings.

Listeners in this study rated /hVt/ words that were produced by speakers who had participated in a previous study. The authors were therefore limited with regard to the corpus from which speech stimuli were selected. A small percentage of the /hVt/ stimuli had less-than-favorable elements associated with them. For example, the /t/ may have been unreleased or the /h/ may have been slightly velarized. Fourteen of the 120 tokens had vowels that were approximations of the exact target. These differences likely affected listener ratings of tokens compared with tokens in which all three phonemes were ideally produced. Acoustic analysis of tokens and listener intelligibility data would strengthen the findings of the present article and should be considered for future research of this type.

The listeners recruited for this study were university students who spoke a dialect of English that was consistent with the North Midland region. By contrast, the native English speakers who contributed speech stimuli were from the Pacific Northwest. U.S. English comprises several dialects that vary according to geographical region. Regional varieties of U.S. English demonstrate variation in vowel production (Jacewicz et al., 2006). The difference in dialect between the English speakers and listeners in the present

study may have influenced the ratings assigned to native English tokens. Although listeners in this study were not fluent in Japanese, Korean, or Spanish, no attempt was made to control for the number of foreign-language classes taken by the listeners. It is possible that students who had completed more foreign-language courses may have responded differently to the speech stimuli.

This last point is related to the study by Evans and Iverson (2004). In that study, Ashby listeners with northern British accents were not adept at selecting good vowel tokens consistent with the southern British English accent. The Ashby listeners may have been less adept at this task because they had little to no experience attempting to produce southern British vowels. With respect to the present study, these findings suggest that listeners with more foreign-language course experience may have responded differently than listeners who had taken fewer foreign-language classes. Those with greater experience attempting to produce words in a different language may have been more flexible in their rating of incorrect tokens.

Although the present study was advertised widely across Miami University's campus, all but one of the participating listeners were women. More male listeners should be included in future studies to enhance the generalizability of the findings. Future studies should include listeners from a variety of backgrounds, regional dialects, ages, familiarity with accents, and language study, to allow for greater generalizability of results. Including ELLs as listeners in the study, as well as listeners who are familiar with accents, would be valuable in future research on listener bias and listener perceptions of accentedness and goodness. As a final matter, the authors are designing a follow-up study to investigate listener preference for some incorrect productions (e.g., /l/ → /ɛ/) over others (e.g., /l/ → /i/). This study will include a wider array of target and substituted vowels than was included in the present study.

Acknowledgments

This research was supported by National Institute of Child Health and Human Development National Institutes of Health; Predoctoral Fellowship F31 HD046412-05. Support was also provided by the ASHA 2014 Students Preparing for Academic Research Careers (SPARC) award. Special thanks to members of Miami University's English Language Learning Pronunciation Lab for assistance with participant recruitment and data collection and to Mitch Bebee from Miami University's Statistical Help Desk.

References

- American Speech-Language-Hearing Association. (1997). *Accent modification*. Retrieved from <http://www.asha.org/public/speech/development/accent-modification/>
- American Speech-Language-Hearing Association. (2010a). *Cultural competence checklist: Personal reflection*. Retrieved from <http://www.asha.org/uploadedFiles/Cultural-Competence-Checklist-Personal-Reflection.pdf>
- American Speech-Language-Hearing Association. (2010b). *Cultural competence checklist: Policies & procedures*. Retrieved from <http://www.asha.org/uploadedFiles/Cultural-Competence-Checklist-Policies-Procedures.pdf>
- American Speech-Language-Hearing Association. (2010c). *Cultural competence checklist: Service delivery*. Retrieved from <http://www.asha.org/uploadedFiles/Cultural-Competence-Checklist-Service-Delivery.pdf>
- Behrman, A., & Akhund, A. (2013). The influence of semantic context on the perception of Spanish-accented American English. *Journal of Speech, Language, and Hearing Research*, 56, 1567–1578.
- Boersma, P. (2002). Praat: A system for doing phonetics by computer. *Glott International*, 5, 341–345.
- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28, 201–219.
- Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (With Griner, B.). (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.
- Compton, A. J. (1996). *Foreign accent norms of American English for 40 of the world's languages*. San Francisco, CA: Carousel House.
- Davies, M. (n.d.). *Word frequency data: Corpus of contemporary American English*. Retrieved from <http://www.wordfrequency.info/>
- Delamere, T. (1996). The importance of interlanguage errors with respect to stereotyping by native speakers in their judgments of second language learners' performance. *System*, 24, 279–297.
- Deprez-Sims, A.-S., & Morris, S. B. (2010). Accents in the workplace: Their effects during a job interview. *International Journal of Psychology*, 45, 417–426.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16.
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39, 379–397.
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42, 476–490.
- Esling, J. H., & Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly*, 17, 89–95.
- Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America*, 115, 352–361.
- Franklin, A. D., & Stoel-Gammon, C. (2014). Using multiple measures to document change in English vowels produced by Japanese, Korean, and Spanish speakers: The case for goodness and intelligibility. *American Journal of Speech-Language Pathology*, 23, 625–640.
- Hayes-Harb, R., & Watzinger-Tharp, J. (2012). Accent, intelligibility, and the role of the listener: Perceptions of English-accented German by native German speakers. *Foreign Language Annals*, 45, 260–282.
- Hillenbrand, J. M., & Gayvert, R. T. (2005). Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research*, 48, 45–60.
- Jacewicz, E., Fox, R. A., & Salmons, J. (2006). Prosodic prominence effects on vowels in chain shifts. *Language Variation and Change*, 18, 285–316.

-
- Kim, M.** (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18, 89–114.
- Lindemann, S.** (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, 31, 419–441.
- Lindemann, S.** (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, 7, 348–364.
- Maddieson, I.** (1984). *Patterns of sounds*. Cambridge, United Kingdom: Cambridge University Press.
- Munro, M. J., & Derwing, T. M.** (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, 11, 253–266.
- Munro, M. J., & Derwing, T. M.** (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 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, 289–306.
- Munro, M. J., Derwing, T. M., & Burgess, C. S.** (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, 52, 626–637.
- Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., & Trent-Brown, S. A.** (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America*, 124, 576–588.
- Peterson, G. E., & Barney, H. L.** (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24, 175–184.
- Trofimovich, P., & Isaacs, T.** (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15, 905–916.