

# Nonnative English Speaker Performance on the Basic English Lexicon (BEL) Sentences

Stacey Rimikis,<sup>a</sup> Rajka Smiljanic,<sup>b</sup> and Lauren Calandruccio<sup>a</sup>

**Purpose:** The purpose of this study was to examine sentence-recognition performance for a large, diverse group of nonnative speakers of English on the recently developed Basic English Lexicon (BEL) sentence materials and to determine whether BEL sentence lists are equated in difficulty for this population.

**Method:** The BEL sentences were presented to 102 nonnative speakers of English with normal hearing and varied linguistic and cultural backgrounds. Five hundred sentences were presented mixed with noise spectrally matched to the target sentences. Subjects completed an online questionnaire providing detailed demographic and linguistic information. Listeners' spoken English proficiency was also assessed using the Versant English Test (Pearson Education, 2010).

**Results:** Nonnative listeners showed equal word-recognition performance for 3 groups of BEL sentence lists, each group containing sentence lists that had equivalent difficulty. In addition, spoken language proficiency and several demographic and linguistic factors were significantly correlated with BEL performance.

**Conclusion:** The BEL sentence materials have been tested on a large cohort of nonnative speakers of English and have been found to be appropriate for use in speech-perception testing with this population.

**Key Words:** speech perception, audiology, cultural and linguistic diversity

The number of nonnative speakers of English residing in the United States has increased steadily over the past three decades (U.S. Census Bureau, 2007). The American Speech-Language-Hearing Association has acknowledged this increase in cultural and linguistic diversity within the population, and the subsequent need for evidence-based practice to address this population's unique needs (American Speech-Language-Hearing Association [ASHA], 2011). With this increase in linguistic diversity, audiologists and speech-language pathologists are faced with the task of finding effective tools to serve individuals from culturally and linguistically diverse backgrounds. Despite the growing need for test materials that would be appropriate for use with nonnative speakers, materials that take into account the specific linguistic needs of this population remain scarce. With the goal of addressing this gap in availability of appropriate test materials for diverse groups of nonnative (as well as native) speakers of English, Calandruccio and Smiljanic (2012) recently developed new sentence-recognition

materials, the *Basic English Lexicon (BEL) sentences*. Listening-in-noise results for native-English-speaking listeners with normal hearing indicated similar performance across a large set of BEL sentence lists. The main goal of this study was to examine speech-in-noise performance using the BEL sentences for a large, diverse cohort of nonnative English listeners. In addition, we explored how BEL sentence-in-noise performance correlates with listener demographic and linguistic backgrounds. In doing this, we were able to show that BEL sentences are appropriate for use with speakers of a wide range of linguistic backgrounds and levels of proficiency.

Although the comprehension of speech in noise is a difficult task in general, past studies have shown that this task becomes even more challenging when the target speech is in a language not native to the listener (e.g., Bradlow & Pisoni, 1999; Takata & Nabelek, 1990). Less accurate speech recognition in noise has been demonstrated by nonnative English speakers in comparison to native speakers, even when the listener has a high level of English proficiency and has been speaking the language for a significant amount of time (Bergman, 1980; Mayo, Florentine, & Buus, 1997). Due to relatively poorer performance, speech audiometry becomes problematic in the assessment of nonnative speakers of English. The main difficulty arises from an inability to distinguish what role auditory impairment as opposed to linguistic inexperience plays in poor performance. Speech-recognition testing is a critical part of the audiological testing battery that provides valuable information about the nature of a person's hearing loss and how it affects individual

<sup>a</sup>Queens College of the City University of New York, Flushing

<sup>b</sup>University of Texas, Austin

Correspondence to: Lauren Calandruccio:

Lauren\_Calandruccio@med.unc.edu

Editor: Sid Bacon

Associate Editor: Eric Healy

Received June 7, 2012

Accepted September 28, 2012

DOI: 10.1044/1092-4388(2012)12-0178)

communicative functioning. Unfortunately, many audiologists omit speech-recognition testing when working with subjects from linguistically diverse backgrounds (ASHA, 2004). If speech-recognition testing is to be included in the test battery for nonnative speakers of English, clinicians have two options: (a) test the listeners using recognition materials in the listener's native language or (b) use English recognition materials.

Recently, the Hearing in Noise Test (HINT; Nilsson, Soli, & Sullivan, 1994), which was originally designed for clinical use with native speakers of English, has been modified for use with several different languages, such as Cantonese, Mandarin, and Japanese (see Soli & Wong, 2008, for a description). Although these new materials provide clinicians with an alternative to English-recognition testing, several problems persist. Language-specific materials for many languages have yet to be created, and the language-specific materials that have been created may not be readily available in all testing centers. Furthermore, in order to accurately score these materials, the tester must be functionally proficient in the given language or have access to an interpreter who understands proper scoring for speech-recognition testing. Because the majority of audiologists in the United States provide only monolingual services in the clinic (ASHA, 2000), most clinicians report using interpreters when serving linguistically diverse populations (ASHA, 2004). This can be problematic if the interpreters are not trained to administer speech-recognition testing and rather serve only as direct translators, adding an element of unreliability to results. Finally, many people who immigrate to the United States use English as their primary language of communication even though it is not their native language, and they are likely to experience their communication disorder in English. Therefore, testing in English may provide a greater level of ecological validity than native-language testing for nonnative individuals who communicate on a day-to-day basis in English.

Testing sentence recognition in English for nonnative speakers, however, gives rise to a separate set of challenges. Foremost, available recognition materials may be linguistically biased for nonnative speakers of English due to unfamiliar vocabulary and complex grammatical structures (see Bradlow & Alexander, 2007). In addition, similar to listeners with hearing loss, nonnative speakers of English are a very heterogeneous group. No two experiences learning English are exactly the same, and the inherent differences in linguistic and demographic experience that make up an individual's background can create large variations in performance. A large body of second-language-learning literature has identified a host of factors that may impact adult learners' ability to attain nativelike levels in second-language (L2) speech production and perception (e.g., Bialystok & Hakuta, 1999; Flege, 1995; Flege et al., 2006; Flege, Yeni-Komshian, & Liu, 1999; MacKay, Flege, & Imai, 2006; Piske, MacKay, & Flege, 2001; Smiljanic, 2012). These studies underscore the importance of better understanding the role of linguistic variables, such as the age of acquisition of L2, length of residence in L2-speaking country, relative

amount of native language (L1) and L2 use, quantity and quality of input from native L2-speakers, motivation, social stigma associated with speaking with an accent, and so forth. These environment- and learner-related parameters contribute to the difficulty in interpreting the results of English speech-recognition tests of nonnative English speakers.

In order to move toward inclusion of speech-recognition testing as a standard part of the audiological test battery for the nonnative English speaker, there is a strong need to better understand English speech-in-noise recognition for the nonnative speaker of English. A crucial step toward this goal is the use of English speech-recognition materials that are linguistically appropriate for nonnative speakers of English. To that end, Calandruccio and Smiljanic (2012) published a new corpus of sentence-recognition materials designed specifically for use with nonnative speakers of English. A distinguishing factor in creating the BEL sentences in comparison to the other existing word-recognition materials was that the BEL sentences were developed based on conversational samples, that is, the actual lexical items used by 100 nonnative speakers of English. Language samples were collected using a set of predetermined, everyday conversational topics, such as travel or cooking. These language samples were analyzed in order to compile a lexicon of all unique words most commonly used across speakers. The lexicon was used in conjunction with simple syntactic frames to create 500 sentences, each containing four keywords (2,000 keywords total). These sentences were divided into 20 lists of 25 sentences each, with lists being relatively equal on the following counts: number of syllables, number of high-frequency phonemes (fricatives and affricates), types of syntactic structures, and median frequency of keyword occurrence in the collected lexicon.

In an initial probe to determine equal list difficulty of the BEL sentence lists, Calandruccio and Smiljanic (2012) reported data for monolingual native speakers of English with normal hearing. Listeners were presented all 20 lists in the presence of spectrally matched noise at a fixed signal-to-noise ratio (SNR) of  $-5$  dB. Results demonstrated similar speech-in-noise performance for all but four sentence list pairs (see Calandruccio & Smiljanic, 2012, Table VII). That is, the majority of BEL lists showed equivalence of difficulty, suggesting that BEL test materials can be used to reliably test speech recognition in native English speakers. It has been reported that difficulty among sentence list materials may vary depending upon the listening population (e.g., sentence-recognition materials that were equally difficult to one another for listeners with normal hearing were not equally difficult for listeners with hearing loss; see Bilger, Nuetzel, Rabinowitz, & Rzezakowski, 1984). Therefore, the next step involves determining whether BEL sentence materials show equivalent difficulty across lists for nonnative listeners of English and would, therefore, be appropriate for use with this diverse population. The purpose of the present study was to test the difficulty of all 20 BEL sentence lists for a large cohort of nonnative English speakers with diverse linguistic and cultural backgrounds. Furthermore, we examined in detail how individual listeners' linguistic experience and

proficiency correlate with speech-in-noise performance on BEL sentences. This allowed us to assess whether BEL sentences would be most appropriate for use with nonnative listeners who have a certain linguistic profile.

## Method

### Participants

The BEL sentences were presented to 102 nonnative speakers of English (72 women, 30 men). Subjects ranged in age from 18 to 50 years with a mean age of 25 years. The mean age at immigration to the United States was 16.94 years ( $SD = 6.84$  years). A wide range of geographic and linguistic backgrounds was represented, with subjects emigrating from 33 nations and representing 24 different native languages (see Tables 1 and 2 for the national and linguistic distributions of the 102 subjects). All subjects were at a minimum bilingual (speaking their native language and English); however, more than half of the subjects identified some level of proficiency in three languages ( $n = 42$ ) or four or more languages ( $n = 16$ ). The mean age at which subjects first began to study English was 11.42 years ( $SD = 5.37$  years), and the mean number of years of English experience was 13.95 ( $SD = 7.29$  years). Subjects were recruited on the campuses of the Queens College of the City University of New York and the University of Texas—Austin using flyers soliciting nonnative speakers of English. Although the majority of participants were recruited on college campuses, a large portion of the Queens College student population comprises nontraditional students (e.g., part-time students, those returning to school after time in the workforce, etc.), allowing for a more diverse subject pool. No subjects

**Table 1.** Native languages of the 102 nonnative English speakers.

Language	<i>n</i>
Spanish	30
Mandarin	15
Cantonese	8
Korean	8
Russian	6
Bengali	4
French	4
Farsi	3
Hebrew	3
Portuguese	3
German	2
Greek	2
Polish	2
Urdu	2
Albanian	1
Arabic	1
Armenian	1
Bulgarian	1
Japanese	1
Romanian	1
Tagalog	1
Thai	1
Ukrainian	1
Vietnamese	1

**Table 2.** Countries of origin of the nonnative English-speaking listeners who reported their birth country.

Country	<i>n</i>
Armenia	1
China	14
Colombia	10
South Korea	8
Ecuador	5
Taiwan	5
Hong Kong	4
Mexico	4
Bangladesh	3
Brazil	3
Dominican Republic	3
Iran	3
Israel	3
Uzbekistan	3
Venezuela	3
Argentina	2
Germany	2
Greece	2
Haiti	2
Peru	2
Poland	2
Ukraine	2
Bulgaria	1
France	1
India	1
Pakistan	1
Philippines	1
Romania	1
Saudi Arabia	1
Spain	1
Thailand	1
Vietnam	1
Yugoslavia	1

*Note.* Of the 102 nonnative English-speaking participants, 97 reported their birth country.

reported history of hearing, language, or communication disorders. Subjects were paid for their participation.

### Procedure

Prior to testing, each subject provided written, informed consent. Subjects also completed a computer-based questionnaire developed by the Linguistics Department at Northwestern University (Chan, 2012) that provided a variety of demographic and linguistic background information about each subject. An otoscopic evaluation and hearing screening was performed on each subject using clinical guidelines for hearing screening procedures (ASHA, 1997). Hearing was screened bilaterally at 20 dB HL at 1000, 2000, and 4000 Hz using TDH headphones and a Grason Stadler audiometer. Subjects were seated in a double-walled, sound-attenuated room for the hearing screening, as well as for the subsequent BEL sentence-in-noise listening test.

### Stimuli: BEL Sentences

The BEL sentences were recorded by a monolingual female speaker of General American English (age 28). An

example of the BEL sentences (with keywords in capital letters) is “The EGGS NEED MORE SALT.” The recordings were made in a double-walled, sound-attenuated booth at a 44.1-kHz sampling rate with 16-bit resolution. For further description of the development of these sentences and how they were recorded, see Calandruccio and Smiljanic (2012).

All sentence files were root-mean-square (RMS) equalized to the same pressure level using Praat (Boersma and Weenink, 2012). For sentence-recognition testing, all stimuli were controlled using Cycling 74’s MaxMSP computer software and were presented to the listeners via Etymotic ER1 insert earphones. Sentences were presented binaurally in the presence of noise spectrally matched to the long-term average spectra (LTAS) of the 500 target sentences. The noise was generated in MATLAB by passing a Gaussian white noise through a finite impulse response (FIR) filter with a magnitude response equal to the LTAS of 500 BEL sentences spoken by the target talker. The long-term average level of the sentences was fixed at 65 dB SPL, whereas the long-term average level of the noise was fixed at 68 dB SPL, providing an SNR of –3 dB. The target sentence and the noise masker were mixed in real time via the MaxMSP software interface. On each trial, one of the target sentences was played with a random portion of the 60-s noise masker. The noise masker began 500 ms prior to the beginning of the sentence and ended 500 ms after the end of the sentence. Subjects’ responses were recorded using a Shure SM81 cardioid-condenser microphone for offline scoring.

Subjects were told that they would hear a series of sentences presented with noise and were instructed to try and ignore the noise and repeat exactly what they heard following each sentence. In order to familiarize subjects with the task as well as the voice of the talker, pretest sentences were presented at gradually decreasing SNRs (three to five sentences presented at each SNR of 3 dB, 0 dB, and –3 dB). Sentences used for familiarization were taken from the Bamford-Kowal-Bench (BKB) sentences (Bench, Kowal, & Bamford, 1979) and were spoken by the same female talker. Following the pretest, subjects were presented with the BEL sentence materials. Sentence-in-noise testing was conducted in two sessions. In each session, listeners heard 10 lists of the BEL sentences (25 sentences/list) presented in a randomly generated order, with sentences blocked by list. Subjects’ responses were scored based on their response to the four keywords in each sentence (100 keywords/list). Criteria for a correct response involved a complete, accurate repetition of a keyword, including morphological affixes. Twenty-five percent of subjects completed the two sessions on the same day with a break between sessions. For the remaining subjects, an average of 12 days passed between the completion of Sessions 1 and 2. If sessions were completed on separate days, an otoscopic evaluation and warm-up were completed prior to the presentation of the second half of the BEL sentence materials.

Following the completion of the second testing session, subjects completed the Versant English Test (Pearson), an automated assessment of spoken English proficiency

completed over the telephone. This assessment provided scores (on a scale from 20 to 80 points) on several facets of spoken English proficiency, including sentence mastery, fluency, vocabulary, and pronunciation, as well as a total proficiency score. Average, minimal, and maximal scores for individual participants can be found in Table 3.

## Scoring

Due to the varying level of accents among subjects, online scoring was not used for computing the final word-recognition scores. Each subject’s performance was instead scored offline using the recorded oral responses made during testing. Two experimenters scored each subject’s responses. For any keywords on which the two scorers differed, a third scorer reviewed the audio recording and made a final judgment.

## Results

### List Equivalency

The primary goal of the perception testing was to determine whether the 20 lists of the BEL sentences were equated in difficulty with respect to speech-in-noise recognition for nonnative speakers of English. Furthermore, we wanted to know whether performance scores between lists were correlated under equivalent listening conditions, as was demonstrated for native speakers of English (Calandruccio & Smiljanic, 2012). The results showed that the mean sentence-recognition accuracy for nonnative listeners ranged between 63 (List 18) and 68 (List 5) percentage points, with an average performance of 65 percentage points across lists. Means, medians, standard errors, and confidence intervals for the 20 lists are shown in Table 4.

A repeated-measures analysis of variance (ANOVA) for percentage correct score as a dependent variable was conducted. List number was the within-subjects factor with 20 levels. The analysis indicated a significant main effect for list number,  $F(1, 19) = 9.43$ ,  $p < .001$ . Post hoc pairwise comparisons based on estimated marginal means using a Bonferroni correction to account for the multiple comparisons were used to determine significant differences in performance among lists. The post hoc testing revealed that although there were significant differences in difficulty among the 20 lists, the lists could be divided into three groups of statistically equal performance. Word-recognition performance among the lists within each of the three groups

**Table 3.** Means (and SDs) and minimum and maximum Versant scores for the 102 nonnative English speakers.

Versant score	<i>M (SD)</i>	Minimum	Maximum
Overall	64.77 (11.14)	41	80
Sentence Mastery	65.51 (12.01)	39	80
Vocabulary	64.86 (10.36)	41	80
Fluency	62.85 (13.20)	36	80
Pronunciation	64.53 (11.47)	36	80



**Table 4.** Mean and median performance on all 20 Basic English Lexicon (BEL) sentence lists for 102 nonnative English speakers with normal hearing.

List	M	Median	SE	95% CI	
				LL	UL
1	66.91	67.50	1.36	64.25	69.58
2	67.02	67.00	1.37	64.33	69.71
3	65.21	68.00	1.54	62.18	68.23
4	62.87	62.50	1.39	60.15	65.60
5	67.92	68.00	1.34	65.30	70.55
6	64.31	63.00	1.56	61.25	67.38
7	66.92	68.50	1.51	63.97	69.88
8	65.99	65.50	1.40	63.25	68.73
9	66.91	69.00	1.44	64.09	69.74
10	67.41	69.00	1.64	64.20	70.62
11	66.84	66.00	1.39	64.13	69.56
12	64.36	65.00	1.49	61.45	67.28
13	62.62	63.50	1.43	59.81	65.43
14	64.94	65.00	1.46	62.07	67.81
15	62.70	62.50	1.32	60.11	65.28
16	64.21	66.00	1.66	60.96	67.45
17	67.74	67.00	1.28	65.23	70.24
18	62.51	62.00	1.64	59.30	65.72
19	65.10	66.50	1.43	62.29	67.90
20	63.75	64.50	1.44	60.92	66.57

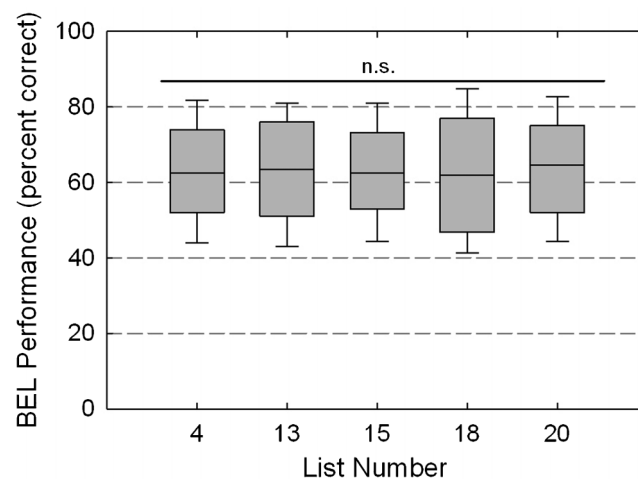
Note. CI = confidence interval; LL = lower limit; UL = upper limit.

was statistically equivalent. Group 1 included five lists: Lists 4, 13, 15, 18, and 20. These lists were the most difficult of the 20 lists, with mean performance ranging from 63 (List 18) to 64 (List 20) percentage points. A second post hoc repeated-measures ANOVA indicated no significant differences in performance among these five lists,  $F(1, 4) = 0.617, p = .650$ . Group 2 included 6 lists: Lists 3, 6, 12, 14, 16, and 19, with performance ranging between 64 (List 16) and 65 (List 3) percentage points. A post hoc repeated-measures ANOVA confirmed no significant differences in performance among these lists,  $F(1, 5) = 0.628, p = .678$ . Last, Group 3 included 9 lists with best overall performance: Lists 1, 2, 5, 7, 8, 9, 10, 11, and 17; mean performance ranged between 65.9 (List 8) and 67.9 (List 5) percentage points. A post hoc repeated-measures ANOVA confirmed no significant differences in performance scores,  $F(1, 8) = 1.125, p = .343$ . Figures 1–3 illustrate performance ranges for the 102 nonnative speakers of English across the 20 lists for the three list groups that allowed for equivalent performance.

### Linguistic Predictors of Performance

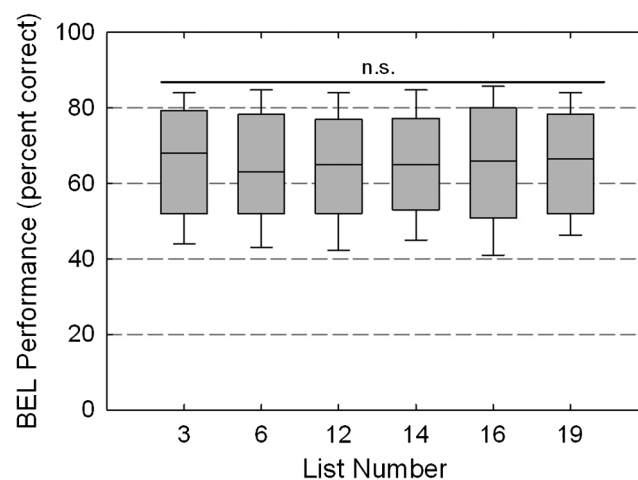
In the following analyses, we attempted to get a better sense of a nonnative demographic for which the BEL sentences would be most appropriate. That is, we examined in detail which environment- and language background-related variables would serve as good predictors for BEL sentence performance. We collected a detailed linguistic history of all participants. Information was obtained regarding the age the participant immigrated to the United States, the percentage of time daily spent speaking English, the number of years lived in the United States, the language

**Figure 1.** Box plots for the 102 nonnative English-speaking, normal-hearing listeners for five BEL sentence lists: Lists 4, 13, 15, 18, and 20 (Group 1). n.s. = nonsignificant statistical groupings.

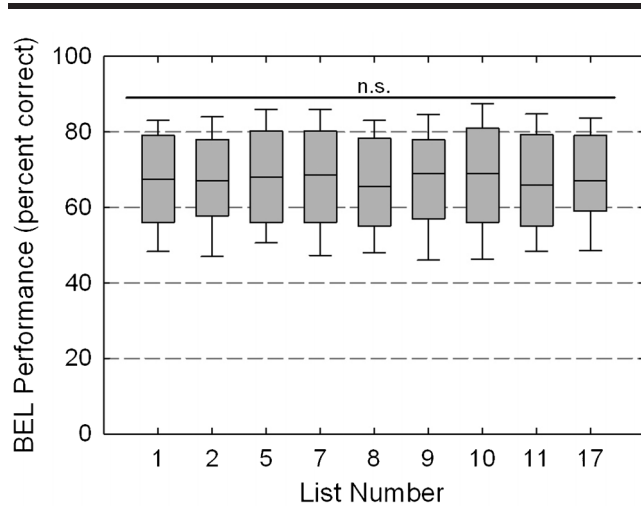


spoken in primary and secondary school (dichotomous reporting, English or other language), the language spoken at home (dichotomous reporting, English or other language), the number of years of English speaking experience, and the age of English acquisition. All of these variables were significantly correlated with BEL performance ( $r$  range =  $-.248$ – $.754, p < .001$ ), except for years of English speaking experience ( $r = .092, p = .193$ ). The seven background variables that were significantly correlated with BEL performance as well as the participant's overall performance score on the Versant test of spoken language proficiency were used as predictor variables in a multiple regression analysis to predict the criterion variable of overall BEL

**Figure 2.** Box plots for the 102 nonnative English-speaking, normal-hearing listeners for six BEL sentence lists: Lists 3, 6, 12, 14, 16, and 19 (Group 2).



**Figure 3.** Box plots for the 102 nonnative English-speaking, normal-hearing listeners for nine BEL sentence lists: Lists 1, 2, 5, 7, 8, 9, 10, 11, and 17 (Group 3).



performance. The regression model indicated (using the stepwise method) an adjusted  $R^2 = .625$ ;  $F(2, 98) = 74.301$ ,  $p < .001$ . Significant predictor variables for the regression model were the overall Versant score and age of immigration (Table 5). The percentage of time daily spent speaking English, the number of years lived in the United States, the language spoken in primary and secondary school, the language spoken at home, the number of years of English speaking experience, and the age of English acquisition were all excluded from the model because they did not significantly add to the strength of the model.

The results thus showed that spoken language proficiency, as indicated by the Versant score, was a highly significant predictor of the sentence-recognition performance. However, as the Versant test is costly and time consuming, it may not be a feasible addition for many clinicians and researchers. Therefore, an additional multiple

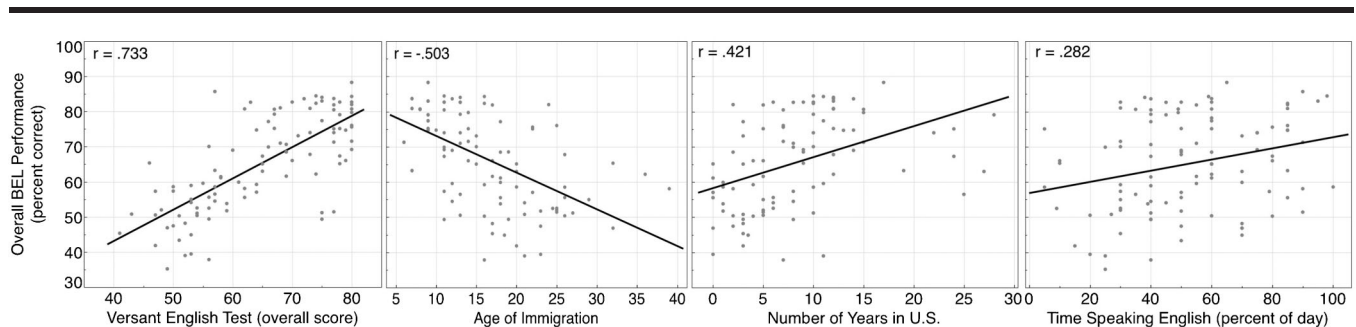
regression analysis with the same predictor variables as stated above minus the overall Versant score was conducted. After excluding the Versant score, the regression model indicated (using the stepwise method) an adjusted  $R^2 = .330$ ;  $F(3, 98) = 9.903$ ,  $p < .001$ . Age at the time of immigration, number of years spent in the United States, and percent of daily English use were significant predictors of the overall sentence-recognition performance (Table 6). The language spoken in primary and secondary school, the language spoken at home, the number of years of English speaking experience, and the age of English acquisition were all excluded from the model. See Figure 4 for correlation plots of the four factors that were included in one of the two stepwise regression models.

The model that included the Versant score accounted for 63% of the variance in the criterion variable. This was greatly reduced by using only self-reported linguistic information as predictor variables (accounting for 33% of the variance in BEL performance).

In addition, the effect of specific L1 on BEL performance was examined. That is, we wanted to know if BEL materials are biased toward a particular language background. Out of the 102 nonnative speaking subjects, the two largest native-language groups that were represented were Spanish speakers ( $n = 30$ ) and speakers of one of the Chinese dialects (e.g., Mandarin, Cantonese;  $n = 23$ ). Additional analyses were conducted to ensure that the three groups of lists were equivalently difficult for the Spanish and Chinese speakers when their performance scores were not grouped with the large subject pool. The results showed that there were no significant differences within list groups for either native language background. Repeated-measures ANOVAs indicated equivalent performance for sentences within Group 1, 2, and 3 each for the Spanish speakers ( $p$  range = .194–.779) and for the Chinese speakers ( $p$  range = .116–.945).

Just over half of the nonnative subjects ( $n = 58$ ) were multilingual (i.e., spoke more than two languages). In order to recruit a diverse pool of subjects, we did not control for linguistic background. It was, therefore, of interest to examine whether the participants who were multilingual

**Figure 4.** Correlation plots for four factors that were included in one of two stepwise regression models. Overall BEL performance (across all 20 lists) at an SNR of  $-3$  dB is shown correlated with overall Versant score, age of immigration, the number of years in the United States, and the percentage of time (daily) speaking English. Pearson correlation coefficients for each factor and BEL performance are shown on each individual plot. All correlations reached significance at an alpha of .01.



**Table 5.** Significant predictor variables for the multiple regression analysis to predict the criterion variable of BEL performance (overall Versant score included in model).

Predictor variable	$\beta$	$p$
Overall Versant score	.661	< .001
Age at immigration	-.271	< .001

performed differently compared to our bilingual listeners. A repeated-measures ANOVA found no significant effect of group: bilingual vs. multilingual,  $F(1, 19) = 1.021$ ,  $p = .315$ . Even though no significant difference was found between groups, it is possible that differences in word-recognition-in-noise scores could arise between these two groups when factors such as English experience or English proficiency (as represented by Versant scores) are taken into account.

Finally, to determine if performance scores among lists for all 102 listeners within the three aforementioned list groups were correlated under identical listening conditions (i.e., in steady-state, spectrally matched noise at  $-3$  dB SNR), correlations between performance scores were calculated. Significant bivariate correlations were observed for all list pairs within each of the three list groups at a significance level of .01 (see Tables 7–9). In addition, an analysis looking at the top 20% of our subjects ( $n = 20$ ) with respect to their performance across all 20 BEL lists was conducted, in which we aimed to examine whether these more proficient nonnative listeners would perform equally on a larger set of BEL sentence lists compared to less proficient listeners. The overall repeated-measures ANOVA was significant,  $F(1, 19) = 3.141$ ,  $p < .001$ . Post hoc pairwise comparisons utilizing a Bonferroni correction indicated no significant differences across the majority of list pairs with the exception of Lists 5 and 4, Lists 5 and 13, and Lists 5 and 15 ( $p$  range = .010–.047). This analysis showed that the more proficient nonnative listeners performed similarly to the native English listeners in our previous study. That is, unlike lower-proficiency nonnative listeners, whose speech-in-noise performance indicated equivalent difficulty for three distinct sets of BEL sentence lists, they found most BEL sentence lists to be of equivalent difficulty.

## Discussion

### List Equivalency

Calandruccio and Smiljanic (2012) showed that performance scores across all 20 lists of the BEL sentences

**Table 6.** Significant predictor variables for the multiple regression analysis to predict the criterion variable of BEL performance (overall Versant score excluded from model).

Predictor variable	$\beta$	$p$
Age at immigration	-.354	< .001
Number of years in United States	.285	.004
Percent daily time speaking English	.220	.015

**Table 7.** Pearson product coefficients for correlations between list pairs for Group 1.

List	4	13	15	18	20
4	—	.877	.751	.856	.827
13		—	.807	.824	.832
15			—	.804	.787
18				—	.811
20					—

Note. All correlations were significant at an alpha level of .01 for the total sample of 102 nonnative English-speaking subjects.

were equivalent in terms of difficulty for native speakers of English with normal hearing, with the exception of four list pairs. The present results provide performance scores for the same materials from 102 normal-hearing listeners with diverse linguistic backgrounds and a range of English proficiencies. Nonnative listeners' word-recognition performance indicated three groups of lists within which performance was equivalent, the largest of which included nine lists (225 sentences, 900 keywords). The current results thus demonstrate that the BEL sentence materials can be used successfully to examine speech recognition in nonnative English-speaking listeners. Unlike native-English-speaking listeners, experimental designs should be limited to nine conditions, as nine was the largest number of lists observed to have equal performance scores. The results also demonstrated that more proficient nonnative listeners performed equally on a larger number of BEL sentence lists, more closely resembling the native English listener performance. It is important to note that similar performance patterns were found for speakers with different L1s, suggesting that BEL materials are not biased toward a particular L1 but are appropriate for use with a variety of language backgrounds and proficiency levels.

One difference between the data for native English speakers and the data reported in this article is the overall accuracy levels of the two groups. The native English listener group reported in Calandruccio and Smiljanic (2012) had a mean performance range between 71 and 79 percentage points from the most difficult to the easiest BEL list. The nonnative English listener performance had a mean range of 63–68 percentage points from the most difficult to the easiest BEL list, respectively. Although nonnative listeners performed more poorly on average, some individual nonnative listeners were within the performance range of native

**Table 8.** Pearson product coefficients for correlations between list pairs for Group 2.

List	3	6	12	14	16	19
3	—	.860	.829	.866	.847	.877
6		—	.839	.862	.847	.875
12			—	.864	.877	.822
14				—	.902	.904
16					—	.893
19						—

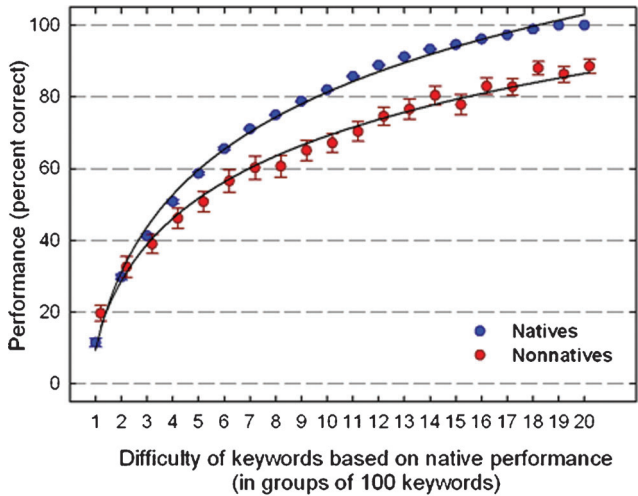
**Table 9.** Pearson product coefficients for correlations between list pairs for Group 3.

List	1	2	5	7	8	9	10	11	17
1	—	.868	.840	.852	.890	.844	.871	.853	.842
2		—	.869	.842	.859	.868	.901	.830	.859
5			—	.867	.852	.869	.868	.822	.871
7				—	.858	.868	.881	.832	.813
8					—	.889	.889	.844	.852
9						—	.869	.830	.830
10							—	.871	.831
11								—	.869
17									—

listeners. The SNR used in this experiment had an added benefit of 2 dB as compared with that used in our previous 2012 study (a -3 dB SNR condition for nonnative listeners vs. a -5 dB SNR condition used for native listeners). The overall accuracy results show that more proficient nonnative listeners can be brought within the performance range of native listeners with the added benefit of a better (in this case, 2 dB) SNR, but for other listeners, a greater improvement in SNR may be needed. These data are in line with previous findings indicating that even for highly proficient nonnative listeners, the added difficulty in processing a second language must be offset by an improved SNR (the amount of which may vary for different materials and listening conditions; Rogers, Lister, Febo, Besing, & Abrams, 2006; Smiljanic & Bradlow, 2011).

In order to further examine similarities and differences in native (from Calandruccio & Smiljanic, 2012) and nonnative listener performance on BEL sentence lists, we compared mean performance for each of the 2,000 keywords. Keywords were ranked based on native-listener performance and then broken into 20 consecutive groups of 100 keywords each. Nonnative performance was then examined following

**Figure 5.** Keyword performance for native and nonnative subjects. All 2,000 keywords were arranged in order of native-listener performance and divided into 20 groups of 100 keywords.



the same native-based keyword ranking (Figure 5). A significant correlation was found between individual keyword performance for native and nonnative listeners ( $r = .819, p < .001$ ). That is, the groups of keywords that were the most difficult for native listeners were the most difficult for nonnative listeners as well. This indicates that the distribution of keyword difficulty is similar between native and nonnative listeners.

In addition to the keyword difficulty among the two groups of listeners, we examined the sentence difficulty for low and high nonnative performers. Work by Healy and Montgomery (2006) showed that a spoken sentence maintains its intelligibility compared to other sentences in a group when manipulated under similar acoustic conditions and/or for various listeners. They further argued that such a sentence effect could still be realized even when large differences in performance (more than 30 percentage points) were observed. Recall that the heterogeneity of the 102 nonnative subjects was large, allowing for a wide range in performance scores at a fixed SNR. The performance scores of the top 10% of performers ( $n = 11$ ) and the bottom 10% of performers ( $n = 10$ ) were analyzed. The top 10% of performers had a mean performance score of 84%, whereas the bottom 10% of performers had a mean performance score of 42%, resulting in a difference in mean performance of 42 percentage points between these two groups. The mean performance of the top 10% and bottom 10% of performers was then analyzed for each of the 500 BEL sentences in order to determine if the same sentences were found to be easy or difficult to process by our nonnative listeners regardless of overall performance score. A strong positive correlation ( $r = .593, p < .001$ ) was found between the intelligibility of the top 10% and bottom 10% of performers for each of the 500 BEL sentences. This correlation indicates that the sentence effect is unaffected by listener overall performance in word recognition, further demonstrating that the BEL sentences would be appropriate for use with nonnative listeners of various proficiency levels provided the listening difficulty (e.g., SNR level) was adapted to their individual needs.

### ***BEL Performance Correlation With Linguistic and Demographic Background***

Our data set included a large number of nonnative speakers with varied linguistic experiences and proficiency levels. The most speech-in-noise performance variance



observed across the nonnative subjects on the BEL sentences was accounted for by performance on the automated Versant test. These data indicate a strong linear relationship between spoken language proficiency, as measured by Versant scores, and BEL performance in steady-state noise. The multiple linear regression model that included the Versant scores was strongly predictive of BEL performance in noise. Nevertheless, the cost and time associated with administering the Versant test may be impractical for use outside of the research laboratory. For research purposes, however, conducting a Versant test on a large subject pool could help recruit subjects for smaller experiments that will have similar speech-in-noise performance levels. This could help alleviate some of the natural variance that tends to occur across subjects when conducting nonnative speech perception. In addition, the Versant test requires only a single administration; subjects do not need to retake the assessment prior to each speech-perception testing session. Therefore, the benefit of using the test may outweigh its cost if the same subjects are used in multiple experimental conditions.

Age of immigration to the United States was another strong predictor of BEL sentence-in-noise performance. Earlier age of immigration correlates strongly with performance on speech-in-noise tasks and English proficiency (Hakuta, Bialystok, & Wiley, 2003). Smaller correlation was found between BEL performance and age of English acquisition. This difference may be due to the fact that the age one immigrates to the United States can be quantified precisely. However, the age of English acquisition may be more subjective in terms of reporting a number for this marker. That is, some people may include the years they spent learning basic vocabulary words whereas others may only count the years in which they felt proficient enough to converse in English. One possible way to strengthen these correlations may be for researchers to generate very specific questions when collecting linguistic histories of nonnative English-speaking subjects (e.g., “How old were you when you were able to comfortably hold a conversation in English?” as opposed to “How old were you when you learned English?”). Similarly, the relationship between BEL performance and age of immigration was also stronger than the relationship between BEL performance and the reported number of years spent in the United States. One possible reason for the stronger relationship may be the earlier age of English immersion and its effect on proficiency. Past research has shown a correlation in age of English immersion and speech-perception ability (Flege, 1995). Many people may immigrate to the United States in adulthood and continue to live here for the remainder of their lives, equaling a high number of years spent in the United States. These subjects may have a lower level of proficiency in comparison to a younger person who immigrated in childhood but has lived fewer years in the United States. These data highlight the importance of better understanding the influence of specific demographic and linguistic experience factors on speech processing for such a heterogeneous listener population. They further demonstrate the importance of taking a detailed linguistic history

when conducting clinical work with a nonnative speaking population to be able to better understand test results (see also von Hapsburg & Peña, 2002).

## Scoring

Due to the varying levels of accents among subjects, we did not use online scoring to compute the final word recognition levels. Each subject's performance was instead scored offline using the recorded oral responses made during testing. Two experimenters scored each subject's responses. For any keywords on which the two scorers differed, a third scorer reviewed the audio recording and made a final judgment.

An analysis investigating the difference in scores between the two offline testers indicated a mean difference of 3.74%. This percent difference between scorers is higher than what is typically observed in similar analyses on data from native-English-speaking listeners (e.g., see Calandruccio, Dhar, & Bradlow, 2010; Calandruccio, Van Engen, Dhar, & Bradlow, 2010, for reported discrepancies between online and offline scoring of 1.4% and 1%, respectively). This is expected given that scoring for nonnative speakers of English is inherently more challenging than scoring for native speakers of English, due to varied English production. For our data set, we found a significant bivariate correlation between Versant pronunciation scores and the percentage differences between the two offline testers,  $r(101) = -.439$ ,  $p < .001$ . In other words, the tester variability was seen to increase with higher levels of accent and overall lower spoken language proficiency.

Two separate post hoc analyses were conducted to determine the most effective way to score these data when variable levels of accent and performance were present. The first analysis compared the original online scoring with the final data. This was conducted by comparing the online scores for each individual keyword for each subject with the final, offline scores and calculating the total number of differences found. A mean difference of 4.66% was found between online and final, offline scores. The second analysis compared the scores of the two offline testers to one another by comparing scores for each keyword and calculating the number of words on which the two testers disagreed. A mean difference of 3.74% was found between the two testers.

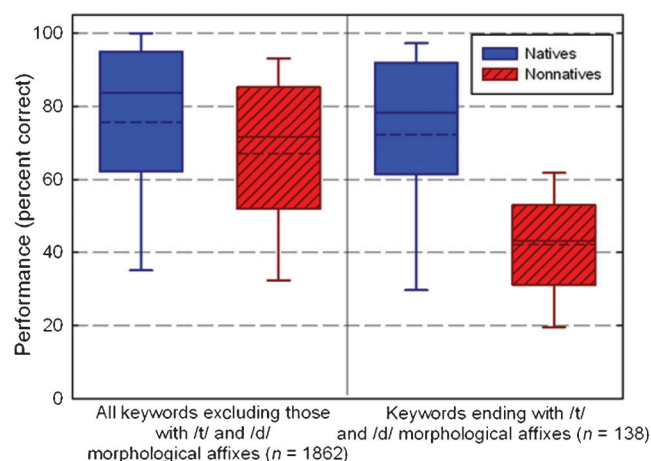
Although variability was indicated between online and offline scoring, the difference between the two offline testers suggests that variability in scoring remains a factor whether scoring is performed online or offline. Given the considerable time and effort needed to conduct full offline scoring, online scoring may be more cost effective considering the small gain in accuracy. However, as with any open-set speech-recognition testing with nonnative speakers of English, a wider margin of error should be expected than would occur when testing native English speakers. The level of reliability can be affected by a range of factors, including speaker accentedness level, tester familiarity with speaker's type of accent (Bradlow & Bent, 2008), and spoken English proficiency.

Another issue regarding scoring arose from the strict scoring criterion we used. A target word was scored as

correct only if all morphological endings were present (e.g., *jump* or *jumps* were not accepted if the target word contained the past tense marker *jumped*). To gain a better insight into the effect of this strict scoring criterion (which would affect nonnative listeners more detrimentally compared to native listeners), keywords from the BEL sentences were categorized based on the presence of certain morphological affixes (see Table 10), and performance was analyzed based on these categories. Of the affixes analyzed, nonnative English speaking subjects performed significantly more poorly on keywords with nonsyllabic past-tense morphological affixes (i.e., *-ed* endings pronounced as /t/ or /d/, such as *kicked* or *seemed*, as opposed to *-ed* endings pronounced as /Id/, such as *painted*) in comparison to performance on the lexicon as a whole. When we analyzed performance for keywords that did not contain nonsyllabic past-tense morphological affixes (a total of 1,862 keywords), nonnative English-speaking subjects showed a mean performance score of 67%. However, when we analyzed performance only on those keywords that contained nonsyllabic, past-tense morphological affixes (a total of 138 keywords), nonnative English-speaking subjects showed a mean performance score of 42%. In contrast, native English-speaking subjects reported in Calandruccio and Smiljanic (2012) showed a much smaller affix-based discrepancy in performance on BEL keywords. These native English subjects showed a mean performance score of 76% on keywords that did not contain nonsyllabic past-tense morphological affixes and a mean performance score of 72% on keywords that did contain nonsyllabic past-tense morphological affixes (see Figure 6).

The morphological affix scoring discrepancy was also analyzed for the two largest linguistic groups of nonnative English speakers (subjects who identified their native language as Spanish and subjects who identified their native language as a Chinese dialect). Native Spanish speakers showed a mean performance score of 71% on keywords that did not contain nonsyllabic past-tense morphological affixes in comparison to 38% on words that contained nonsyllabic, past-tense morphological affixes. Native speakers of a Chinese dialect showed a mean performance score of 59% on keywords that did not contain nonsyllabic past-tense morphological affixes in comparison to a mean performance score of 29% on words that contained nonsyllabic, past-tense morphological affixes. It is important to note that the difference in performance between native Spanish and

**Figure 6.** Mean performance for all keywords, excluding those with /t/ and /d/ morphological affixes ( $n = 1,862$ ). The two box plots on the left indicate mean performance for all keywords excluding those with /t/ and /d/ morphological affixes. The two box plots on the right indicate performance for all keywords ending with /t/ and /d/ morphological affixes ( $n = 138$ ).



Chinese speakers is due to the overall proficiency levels between the two groups (with Spanish speakers overall being more proficient) rather than the increased sentence difficulty for a particular L1. Thus, our findings support previous research on nonnative English speaker production of nonsyllabic past-tense morphological affixes in that the omission of these affixes has been found to be a common phenomenon in nonnative English speech from a variety of linguistic backgrounds and proficiency levels (Klein et al., 2004; Lardiere, 2003; Solt et al., 2004).

In light of the difficulty of producing nonsyllabic past-tense morphological affixes, nonnative speakers could be penalized in scoring in open-set speech-recognition testing for what is actually an accent-based production error incorrectly interpreted as an auditory misperception. In order to reduce this effect, the possibility of using root-word-only scoring for nonnative speakers of English may be considered for future speech-recognition testing. However, one consideration with this approach is the high incidence of high-frequency phonemes in morphological endings (e.g., the /s/ in plural nouns and present-tense verbs). As high-frequency

**Table 10.** Morphological affixes analyzed in BEL keyword performance.

Phonology	Part of speech	Examples from BEL sentences
/t/	verb (past tense)	<i>looked, walked, practiced</i>
/d/	verb (past tense)	<i>seemed, played, learned</i>
/Id/	verb (past tense)	<i>sounded, ended, tasted</i>
/s/	verb (present tense, third person singular)	<i>sleeps, drinks, asks</i>
/z/	verb (present tense, third person singular)	<i>warms, answers, lays</i>
/Iz/	verb (present tense, third person singular)	<i>brushes, chases, watches</i>
/s/	noun (plural, regular)	<i>jokes, nights, books</i>
/z/	noun (plural, regular)	<i>buttons, songs, trees</i>
/Iz/	noun (plural, regular)	<i>nurses, colleges, beaches</i>

phonemes can be valuable in assessing certain auditory impairments such as presbycusis, utilizing the root-word-only scoring model only for words with nonsyllabic past-tense /t/ and /d/ endings may be an option. A pressing goal of future research remains to determine the potential interaction between linguistic factors (i.e., nonnative language speech processing) and hearing loss for L2 listeners.

Another option in testing is closed-set scoring. With a decrease in spoken language fluency, foreign-accented speech is less reliably interpreted and understood by others. Recall the correlation reported above between Versant pronunciation score and offline tester agreement. For those speakers with heavily accented speech, closed-set testing may serve to alleviate tester difficulty in scoring. However, closed-set testing may present other drawbacks, namely that the task is not equal to open-set testing with respect to information processing demands and level of chance performance (Luce & Pisoni, 1998). Therefore, this type of testing most likely should be considered only for groups for which it may present a marked benefit in scoring (i.e., those who speak English on a daily basis, but have heavily accented speech).

It may be argued that in testing the speech recognition of nonnative speakers of English, written scoring may alleviate the difficulty of tester interpretation of subject accent. However, in an informal analysis in our lab in which both written and verbal responses were collected during speech-recognition testing, it was found that spelling errors mimicked verbal-accent errors, in that many subjects orthographically transcribed their phonetic pronunciation. For example, the keywords *dull* and *urge* were written as “dool” and “earch,” respectively. Furthermore, English spelling is highly irregular and may pose additional problems for nonnative listeners, especially those who may not have had formal schooling. Thus written responses presented similar difficulties to verbal scoring, in that there was tester disagreement concerning at which point such errors would be interpreted as incorrect responses. In addition, written scoring is significantly less time-effective than verbal scoring, in that gathering written responses increases the time required for testing at least twofold.

Finally, we considered word frequency patterns for BEL sentence lists. Previous research established that word frequency (how often a word occurs in a lexicon) is an important variable in speech processing. High-frequency words are perceived and produced more quickly and more efficiently and are easier to recall than low-frequency words (e.g., Balota & Chumbley, 1984; Glanzer & Bowles, 1976; Jescheniak & Levelt, 1994; Rayner & Duffy, 1986). Even though second-language learners may not share all lexical knowledge with monolingual speakers, it is likely that some frequency effects play a role in speech processing for nonnative speakers (especially those with higher proficiency). In order to assess whether word-recognition results on the BEL sentences bear any relationship to lexical frequency, we analyzed the BEL sentence keywords according to frequency of occurrence in the English language using a measure of frequency per million. For frequency counts, we used the SUBTLEXus which is based on the subtitles from American

films and television series (for a total of 51 million words; Brysbaert & New, 2009). A value was assigned to each individual keyword, and medians (based on 100 keywords) were calculated for each of the 20 lists of the BEL sentences. When we compared these medians for frequency per million to performance on the BEL sentence lists, we found no significant relationship between group difficulty and keyword frequency ( $r = .402, p = .079$ ). The three groups of BEL sentence lists with equivalent difficulty all had similar word frequency distribution, and the different word-recognition performance of our nonnative listeners between the three list groups cannot be accounted for by the different frequencies of the keywords across lists.

Keywords were also analyzed according to frequency of occurrence within the original lexicon used to develop the BEL sentences. Recall that the original lexicon was derived from the conversational speech samples of 100 nonnative speakers of English, and, as such, it reflects familiarity with the target words of a large number of nonnative speakers. Again, a value was assigned to each individual keyword and medians were calculated for each of the 20 lists of the BEL sentences. For the list medians, a relationship was found between list difficulty and frequency of keyword occurrence within the BEL nonnative lexicon, such that the better word-recognition score (for each list) was significantly correlated with high word-frequency occurrence in the lexicon ( $r = .475, p = .034$ ). This significant correlation between recognition performance and BEL lexicon word frequency may indicate that the original lexicon used to create the BEL sentences is a more accurate reflection of the nonnative population's familiarity with some of the target keywords.

In analyzing the current data, we opted to use statistical equivalence to determine differences in performance accuracy across lists. This analysis is aimed at establishing the appropriateness of the BEL sentences for use in research where this type of analysis is most typically used. Establishing critical difference values for the BEL sentence lists will be an important next step in examining the clinical use of these materials (e.g., Cox, Alexander, & Gilmore, 1987; Thornton & Raffin, 1978).

## Conclusion

In this study, our goal was to extend the findings of Calandruccio and Smiljanic (2012), who found that the newly developed BEL sentence materials presented equivalent list difficulty for normal-hearing, native English listeners. Here we showed that these materials are suitable for testing the speech-recognition abilities of nonnative English speakers as well. It is important to note that the results showed that these materials are appropriate for use with a wide variety of nonnative speakers, including different L1s and various levels of proficiency. The results from a large number (102) of nonnative listeners provided insights into the factors that must be taken into consideration when testing this target population. A detailed linguistic and demographic history should be considered when interpreting nonnative listeners' speech-recognition results, including



information regarding the listener's age of immigration to the United States and percent of time daily speaking English. For research purposes, the results revealed that a quick, automated spoken language proficiency test (Versant) can aid in selecting a desired and less variable nonnative demographic. Differences in list equivalencies in nonnative English speaker performance on BEL materials in relation to native English speakers were also found. The list groupings indicated by the current data should be taken into consideration when testing nonnative English speakers on multiple BEL lists. Nevertheless, the various analyses on the characteristics of the BEL sentence materials, which were created using lexical items and syntactic structures that would reflect the actual English experience of nonnative speakers, suggest that these materials may be used to test speech recognition for various listener populations. Ultimately, more research of nonnative speech-in-noise recognition on these and other test materials will allow us to move closer to an end goal of including speech recognition in the standard battery of audiological assessment for the nonnative English-speaking population.

## Acknowledgments

This study was funded, in part, by The Capita Foundation. We are indebted to the undergraduate and graduate research assistants in the Speech and Auditory Research Laboratory at Queens College and the UTSoundLab at the University of Texas—Austin. Portions of these data were presented at the 2011 American Speech-Language-Hearing Association convention in San Diego, California, the 2011 meeting of the Acoustical Society of America in San Diego, California, and the 2012 New York State Speech-Language-Hearing Association convention in Saratoga Springs, New York.

## References

- American Speech-Language-Hearing Association.** (1997). *Guidelines for audiologic screening*. Available from [www.asha.org/policy](http://www.asha.org/policy).
- American Speech-Language-Hearing Association.** (2000). *2000 Omnibus survey: Practice trends in audiology*. Rockville, MD: Author.
- American Speech-Language-Hearing Association.** (2004). *Audiology survey: Practice trends in audiology*. Rockville, MD: Author.
- American Speech-Language-Hearing Association.** (2011). *Cultural competence in professional service delivery* [Professional Issues Statement]. Available from [www.asha.org/policy](http://www.asha.org/policy).
- Balota, D. A., & Chumbley, J. I.** (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 340–357.
- Bench, J., Kowal, A., & Bamford, J.** (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially hearing children. *British Journal of Audiology*, 13, 108–112.
- Bergman, M. (Ed.).** (1980). *Aging and the perception of speech*. Baltimore, MD: University Park Press.
- Bialystok, E., & Hakuta, K.** (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis*. Mahwah, NJ: Erlbaum.
- Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C.** (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 27, 32–48.
- Boersma, P., & Weenink, D.** (2012). Praat: Doing phonetics by computer (Version 5.3.15) [Computer program]. Retrieved from <http://www.praat.org>.
- Bradlow, A. R., & Alexander, J. A.** (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and nonnative listeners. *The Journal of the Acoustical Society of America*, 121, 2339–2349.
- Bradlow, A. R., & Bent, T.** (2008). Perceptual adaptation to nonnative speech. *Cognition*, 106, 707–729.
- Bradlow, A. R., & Pisoni, D. B.** (1999). Recognition of spoken words by native and nonnative listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, 106, 2074–2085.
- Brysbaert, M., & New, B.** (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977–990.
- Calandruccio, L., Dhar, S., & Bradlow, A. R.** (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*, 128, 860–869.
- Calandruccio, L., & Smiljanic, R.** (2012). New sentence-recognition materials developed using a basic nonnative English lexicon. *Journal of Speech, Language, and Hearing Research*, 55, 1–34.
- Calandruccio, L., Van Engen, K., Dhar, S., & Bradlow, A. R.** (2010). The effectiveness of clear speech as a masker. *Journal of Speech, Language, and Hearing Research*, 5, 1458–1471.
- Chan, C. L.** (2012). NU-subdb: Northwestern University Subject Database [Web application]. Department of Linguistics, Northwestern University. <https://babel.ling.northwestern.edu/nusubdb2/>.
- Cox, R. M., Alexander, G. C., & Gilmore, C.** (1987). Development of the Connected Speech Test (CST). *Ear and Hearing*, 8, 119S–126S.
- Flege, J. E.** (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–272). Timonium, MD: York.
- Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K.** (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 33, 153–175.
- Flege, J. E., Yeni-Komshian, G. H., & Liu, S.** (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, 41, 78–104.
- Glanzer, M., & Bowles, N.** (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 21–31.
- Hakuta, K., Bialystok, E., & Wiley, E.** (2003). Critical evidence: Test of the critical period hypothesis for 2nd language acquisition. *Psychological Science*, 14, 31–38.
- Healy, E. W., & Montgomery, A. A.** (2006). Consistency of sentence intelligibility across difficult listening situations. *Journal of Speech, Language, and Hearing Research*, 49, 823–834.
- Jescheniak, J. D., & Levelt, W. J. M.** (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 824–843.
- Klein, E., Stoyneshka, I., Adams, K., Rose, T., Pugash, Y., & Solt, S.** (2004). Past tense affixation in L2 English: The effects of lexical



- aspect and perceptual salience. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *BUCLD 28: Proceedings of the 28th Annual Boston University Conference on Language Development: Online supplement*. Available from [www.bu.edu/buclid/proceedings/supplement/](http://www.bu.edu/buclid/proceedings/supplement/).
- Lardiere, D.** (2003). Second language knowledge of [+/-Past] and [+/-Finite]. In J. M. Liceras et al. (Eds.), *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference (GASLA 2002)*; pp. 176–189). Somerville, MA: Cascadilla Proceedings Project.
- Luce, P. A., & Pisoni, D. B.** (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36.
- MacKay, I. R. A., Flege, J. E., & Imai, S.** (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics*, 27, 157–183.
- Mayo, L. H., Florentine, M., & Buus, S.** (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, 40, 686–693.
- Nilsson, M., Soli, S. D., & Sullivan, J. A.** (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95, 1085–1099.
- Pearson Education.** (2010). *Versant English Test description and validation manual*. Palo Alto, CA: Author. Retrieved from [www.versanttest.com/products/english.jsp](http://www.versanttest.com/products/english.jsp)
- Piske, T., MacKay, I.R.A., & Flege, J. E.** (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, 29, 191–215.
- Rayner, K., & Duffy, S. A.** (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14, 191–201.
- Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B.** (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27, 465–485.
- Smiljanic, R.** (2012). Second language learning. In A. C. Cohn, C. Fougerson, & M. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (pp. 417–426). Oxford, United Kingdom: Oxford University Press.
- Smiljanic, R., & Bradlow, A.** (2011). Bidirectional clear speech perception benefit for native and high proficiency nonnative talker-listener pairs: Intelligibility and accentedness. *The Journal of the Acoustical Society of America*, 130, 4020–4031.
- Soli, S. D., & Wong, L. L. N.** (2008). Assessment of speech intelligibility in noise with hearing in noise test. *International Journal of Audiology*, 47, 356–361.
- Solt, S., Pugach, Y., Klein, E. C., Adams, K., Stoyneshka, I., & Rose, T.** (2004). L2 perception and production of the English regular past: Evidence of phonological effects. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *BUCLD 28: Proceedings of the 28th annual Boston University Conference on Language Development* (pp. 553–564). Somerville, MA: Cascadilla Press.
- Takata, Y., & Nabelek, A. K.** (1990). English consonant recognition in noise and in reverberation by Japanese and American listeners. *The Journal of the Acoustical Society of America*, 88, 663–666.
- Thornton, A. R., & Raffin, M. I.** (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research*, 21, 507–518.
- U. S. Census Bureau.** (2007). *Language use in the United States: 2007* [American Community Survey Reports]. Retrieved from <http://www.census.gov/prod/2010pubs/acs-12.pdf>.
- von Hapsburg, D., & Peña, E. D.** (2002). Understanding bilingualism and its impact on speech audiometry. *Journal of Speech, Language, and Hearing Research*, 45, 202–213.