JSLHR

Article

New Sentence Recognition Materials Developed Using a Basic Non-Native English Lexicon

Lauren Calandruccio[a] and Rajka Smiljanic[b]

Purpose: The objective of this project was to develop new sentence test materials drawing on a basic non-native English lexicon that could be used to test speech recognition for various listener populations. These materials have been designed to provide a test tool that is less linguistically biased, compared with materials that are currently available, for sentence recognition for non-native as well as native speakers of English.
Method: One hundred non-native speakers of English were interviewed on a range of 20 conversational topics. Over 26 hr of recorded non-native English speech were transcribed. These transcriptions were used to create a lexicon of over 4,000 unique words. The words from this lexicon were used to create the new materials based on a simple syntactic sentence structure frame.
Results: Twenty lists of 25 sentences were developed. Each sentence has 4 keywords, providing 100 keywords per list. Lists were equated for rate of occurrence of keywords in lexicon, high-frequency count (total number of affricates and fricatives), number of syllables, and distribution of syntactic structure. Listening-in-noise results for native-English-speaking, normal-hearing listeners indicated similar performance across lists.
Conclusion: The Basic English Lexicon materials provide a large set of sentences for native and non-native English speech-recognition testing.

Key Words: sentence testing, speech in noise, non-native English speaker

Over the past few decades, the study of second language (L2) speech perception has emerged as an expanding research area, the results of which have implications for a range of disciplines, including linguistics, speech science, audiology, and education. More recently, there has been an increased interest regarding the speech perception of listeners attending to speech in their non-native language in adverse listening conditions, such as in noise, which mimic real-life communication situations (e.g., Bradlow & Alexander, 2007; Cutler, Garcia Lecumberri, & Cooke, 2008; Pinet & Iverson, 2010; Shi, 2010; Smiljanic & Bradlow, 2011; van Wijngaarden, Steeneken, & Houtgast, 2002). This interest comes at a time when the U.S. population continues to become more diverse (U.S. Census Bureau, 2010). The increasingly diverse populace requires that speech-language pathologists and audiologists alike meet the demands of clients who have communication disorders and primarily communicate in English but do not speak English as their first language (Adler, 1990; Ballachanda, 2001a, 2001b). To properly assess communication disorders for this population as well as to eventually provide evidence-based practice for these clients, it is necessary to have a better understanding of speech perception of the non-impaired (i.e., those with normal audiometric thresholds, normal speech and/or language), non-native listener.

One facet that has hindered speech perception research with non-native speakers of English is the lack of test materials specifically designed to test the non-native English speaking population. The currently available test materials have typically been developed with other listener groups in mind (e.g., native listeners or children). They are often limited in the number of unique sentences or keywords that can be manipulated in experimental conditions. The existing test materials may thus be inappropriate for use with non-native adult listeners and can often lead to researchers developing their own project-specific materials. Creating novel stimuli is time consuming and may not allow for direct comparisons of the findings across various studies (see

[a]Queens College of the City University of New York, Flushing
[b]University of Texas at Austin
Correspondence to Lauren Calandruccio, who is now with the University of North Carolina Chapel Hill School of Medicine, Division of Speech and Hearing Sciences: lauren.calandruccio@med.unc.edu

Editor: Sid Bacon
Associate Editor: Eric Healy
Received September 16, 2011
Revision received January 27, 2012
Accepted February 21, 2012
DOI: 10.1044/1092-4388(2012/11-0260)

Bent and Bradlow, 2003, and Bradlow and Alexander, 2007, for discussions of such issues). With these issues in mind, the goal of the current project was to create a new, large set of test materials for sentence recognition that would be appropriate for both native- and non-native English speakers. These materials could then be used to investigate speech perception as well as to assess hearing abilities in the laboratory and clinics. Our long-term research goals, of which the new test materials are a crucial component, are to gain insight into the interaction of auditory and linguistic factors in shaping first and second language speech perception processes. As a part of this agenda, we have developed these sentences in the hope that they will aid in examining applied and theoretical issues regarding native and non-native speech processing.

This article describes the development of the new Basic English Lexicon (BEL) sentence recognition materials. Although these materials share some facets with existing sentence lists (e.g., Bamford-Kowal-Bench [BKB] and Speech Perception in Noise Test [SPIN]), they are distinguished by two important features. First, they are based on a basic English lexicon derived from naturalistic interactions with non-native speakers. We use "basic" here to indicate vocabulary used by non-native talkers in conversations discussing everyday topics. Second, they comprise a large set of original sentences (500 total with 2,000 keywords). These design features should allow for less linguistically biased speech recognition and hearing assessments of both native and non-native listener groups as well as for testing of multiple experimental conditions.

Speech recognition can be studied using many different components of speech, including phonemes, words, and sentences. Though research using each of these stimuli sheds light onto speech processing at distinct levels of linguistic structure, we opted to focus on the development of sentence materials since they reflect ecologically valid speech structures encountered in real-world listening environments (Villchur, 1982). A potential problem with using sentence materials for experimental and clinical testing is that once listeners hear the sentence, they become familiar with the test material. Presenting the listener with the same material again can bias recognition scores because the listener can rely partially on memory to complete the task (Yund & Woods, 2010). A limited number of sentences has proven to be challenging for researchers conducting experiments with multiple experimental conditions or with smaller populations that are difficult to recruit (e.g., hearing-impaired, cochlear-implant, and non-native English speaking listeners). An even more challenging problem when testing non-native speakers of English is that they may be unfamiliar with the vocabulary used within the sentences, and the sentence structure may be difficult for them to process. This may lead them to replace the target words with words that are more familiar to them or that make sense within the sentence even though it was not the signal they heard. The lack of familiarity and the inexperience with the various levels of L2 linguistic structure can thus further bias non-native listeners' recognition scores. It is therefore crucial to use sentence recognition materials that will aid researchers and clinicians in separating perception difficulties that are linguistically driven from those that are auditorily driven in non-native listener groups.

Many researchers have turned to the BKB (Bench, Kowal, & Bamford, 1979) sentence lists (or a revised version of these sentences) when testing non-native speakers of English (e.g., Crandell & Smaldino, 1996; Mayo, Florentine, & Buus, 1997; Nakamura & Gordon-Salant, 2011; Pinet & Iverson, 2010; Van Engen, 2010). The BKB sentence lists are composed of 21 lists of 16 sentences (1,050 keywords in total) and were originally developed using a lexicon from hearing-impaired British children to assess their auditory abilities. Bent and Bradlow (2003), for instance, used four lists from the revised BKB (Bamford & Wilson, 1979) lists to conduct English speech-perception testing for non-native speakers. There were 144 unique keywords across the four lists of sentences. After perception testing was completed, the non-native speakers of English performed a word familiarity test to determine how familiar the listeners were with the keywords. Listeners rated their familiarity on a 7-point scale ranging from 1 (*I don't know this word*) to 7 (*I know this word*). Additional words that previously have been shown to be less familiar with non-native speakers of English (Bradlow & Pisoni, 1999) were also included to allow the entire spectrum of the 7-point scale to be used. Results indicated that listeners ranked approximately 80% of the keywords as a 7 on the familiarity scale and very few keywords were rated below a 5 on the 7-point scale. This suggests that the listeners were familiar with and at least recognized the majority of the test words. These results highlight that, at a minimum, a portion of the BKB sentences are a good alternative when testing speech perception of non-natives (see appendix in Bent & Bradlow, 2003). However, the BKB sentences, which were aimed at testing children, include vocabulary and syntactic structures that are simpler than what non-native adults encounter in daily interactions.

Another set of test materials that are commonly used in speech perception research (e.g., Mayo et al., 1997; Tabri, Chacra, & Pring, 2011), SPIN sentences, were designed to separate differences in performance due to acoustic-phonetic versus "linguistic-situational" information. SPIN sentences are scored based on the listener's response to the perception of one noun that is in the final word position. In half of the sentences, the final

word is highly predictable based on the beginning of the sentence (e.g., "A bear has a thick coat of fur"), whereas the other half of the sentences contain a final word that is not predictable based on the beginning of the sentence (e.g., "They knew about the fur"). When SPIN sentences were created, the familiarity of the final words was controlled by using words from the Thorndike and Lorge lists (1944). These sentences were designed for use with a clinical population and initially included 250 low-predictability and 250 high-predictability sentences (a total of 250 keywords in which the same items were used for both sentence types). Later, Bilger, Nuetzel, Rabinowitz, and Rzeczkowski (1984) reported that only eight of the original 10 lists were equivalent in difficulty for a clinical population (a total of 200 keywords per sentence type), and these lists have become known as the Revised SPIN sentences. The Revised SPIN sentences have been used for non-native English sentence-recognition testing (Mayo et al., 1997), even though many of the final keywords may not be familiar to non-native speakers of English (e.g., "flock," "mast," "rim," "mist," "notch"). Though Mayo et al. (1997) reported that highly proficient bilinguals perform near native performance in quiet for the recognition of SPIN sentences, Bradlow and Alexander (2007) developed an alternative list of 120 sentences (60 high- and 60 low-predictability sentences) that used vocabulary (e.g., "paper," "bird," "trees," "mother") that would be more familiar to a broader range of proficiency levels for non-native speakers of English. To test the familiarity of the keywords for the high-predictability sentences, non-native English speakers were asked to fill in the blank for the final key word of the sentences they tested. Sentences that were most consistently completed were included in testing. Bradlow and Alexander developed these lists to investigate non-native listeners' ability to benefit from either semantic contextual cues or exaggerations of acoustic-phonetic cues. As already mentioned, an important caveat with using the SPIN materials (and the alternative lists provided by Bradlow and Alexander) is that the amount of materials available to test the subject is limited. This is challenging when there are multiple experimental conditions or a limited number of subjects available for testing across experiments. These projects also highlight the fact that in order to obtain meaningful data, researchers currently must resort to developing their own materials for specific experiments, which is both time consuming and costly.

Though the existing materials aided in providing us with important insights into sentence recognition for various populations, we undertook the task of developing new sentence materials that would more directly address the unique requirements posed by speakers of English as a second language. In doing this, we focused on the following features: first, designing a greater number of sentence materials to enable difficult-to-recruit listeners to complete tasks under multiple experimental conditions; second, developing sentence materials using the lexicon derived from the actual non-native speakers' spontaneous productions; third, creating sentences in which the majority of non-native speakers of English would be familiar with the vocabulary items; last, constructing sentences with a simple English structure to make the task less difficult for non-native listeners and older listeners. To this end, we collected a large non-native English lexicon from the naturalistic speech of 100 non-native English speakers. Using this lexicon, we created 500 sentences. This carefully designed set of 500 sentences aims to fill a gap in the resources available to researchers and clinicians for conducting studies and assessments with native and non-native speakers of English. In addition, high-quality recordings of the new materials are available. The development of the BEL sentence materials is described below.

Method

Basic Lexicon Development Procedures

The central consideration in developing the sentence materials was to use only lexical items with which a large number of non-native listeners, who may have limited knowledge of the English lexicon and syntax (among other L2 features), would be familiar. Native speakers, of course, would have familiarity with these more basic vocabulary items as well. For the purposes of ensuring the use of the lexical items that L2 speakers were likely to know in these materials, we adopted a novel approach in collecting the basic lexicon. Rather than using second language learning textbooks or assuming that certain items would be known to L2 learners, we conducted interviews with 100 non-native speakers of English. We focused on relatively proficient L2 learners with diverse language backgrounds who reside in the United States and are representative of the larger U.S. demographic. These materials are, therefore, most appropriate for a U.S. non-native English listener population. The 100 participants were part of the Queens College and the broader Queens County community. Queens is the easternmost of the five New York City boroughs. It has over 2 million residents, approximately half of whom were born outside of the United States. Approximately 138 languages are spoken in Queens (New York State Comptroller, 2000).

Conversations were elicited using 20 predetermined topics (see Table 1). On average, 40 participants discussed each topic (with a range of 29–53 participants per topic). Topics were designed to obtain a large lexicon while limiting the scope of the conversation. Also, many conversational topics were closely related to other topics. The specific number and types of topics we chose were

Table 1. The 20 topics used to elicit conversations with non-native speakers of English.

Topic	Number of participants discussing topic
Cooking	40
Travel	44
Children	39
Education	41
Consumption	41
Nationalism	40
Sports	35
Cost of living	41
Extended family	39
Holiday	40
Work	53
At the store	41
Medically related	29
Learning a language	39
Running errands	30
Dating	39
Leisure activities	38
Domestic pets	42
Music	41
Vacation	43

necessary to generate a large amount of vocabulary while at the same time trying to increase vocabulary repetition across participants. Research assistants, who were native speakers of American English, began the conversations by reading short, scripted introductions based on each topic. Several open-ended questions were used to prompt conversation. These questions were mainly used for those participants who were less talkative. For many participants, dialogue flowed naturally, and scripted questions were not necessary.[1] On average, each participant discussed eight conversational topics and spoke for approximately 2 min per topic. In total, 795 short conversations, totaling more than 26 hr of conversational speech, were digitally recorded using handheld SONY IC recorders with an attached lapel microphone. All procedures were approved by the Institutional Review Board at Queens College of the City University of New York. Participants were provided with a written statement regarding the research project. Written consent was not obtained as anonymity was maintained for all participants. Payment was not provided for this portion of the research project.

The next step in creating the lexicon involved orthographically transcribing the participants' speech during the interviews. Each conversation was transcribed by one transcriber and checked for reliability by a second transcriber. Words that could be identified clearly from their pronunciation and identified by the context in which they were used were included as lexical items. Even though some non-native productions differed from the native targets, we opted to include them because at this point we were concerned not with the accuracy of their pronunciation but rather with their active knowledge of these lexical items. Crucially, both transcribers had to agree that the words used revealed familiarity with the meaning and appropriate use in the sentence. From these data, a lexicon was created for each individual talker, including all the words used by that talker. Each lexical item was labeled for syllable count and part of speech. On average, each participant used 385 unique words per recording session. The first time a word was used, it was considered unique. All words (including every word spoken) used across all talkers were then combined into a master lexicon. In the master lexicon, each word was marked for the frequency with which it occurred (i.e., the number of times repeated) across all participants. The final lexicon included more than 200,000 spoken words, with 4,062 unique words (excluding function words) spoken with varying rates of occurrence. The words from the master lexicon were used as the vocabulary to develop the test sentences.

Background Information for Speakers Used to Create Lexicon

All participants completed a questionnaire modeled after the Northwestern University Subject Database questionnaire (Chan, 2012) developed by the Department of Linguistics at Northwestern University. Questions addressed the speakers' demographic information, language background, English experience, and so forth. Results from the questionnaire revealed that participants had a wide range of linguistic and educational backgrounds, English experience, and English proficiency levels. Of the 100 non-native English speakers, 55 participants were women, and 45 were men. Participants ranged between 18 and 73 years in age (M = 32 years, SD = 13 years). The participants represented 28 different nationalities and 16 native languages (see Tables 2 and 3, respectively). Sixty-nine of the participants spoke two languages (native language and English), 29 spoke three languages, and two spoke four languages. Not all 100 participants reported responses for their race and ethnicity; however, 70 of the participants reported

[1]As an example, the scripted introduction and the follow-up questions for the topic of consumption and consumerism are provided: "Many Americans love to shop. As a culture, we tend to buy too much and spend way too much money. We are also getting a bad reputation for wasting too much." Following this introduction, the below scripted questions might be used as prompts: (1) How important is shopping in [home country]? (2) Do people typically only buy what they need, or do they shop for fun? (3) Do you enjoy shopping? Why/why not? (4) After you've paid all your bills, what do you do with money you have left over? (5) Do you see a problem with people buying too much stuff? How does this affect the economy? The environment? (6) Do people in [home country] usually shop with credit cards or cash? Do you usually use a credit card or cash? Why?

Table 2. Number of participants from each country of origin for the 100 non-native speakers of English.

Country of origin	Number of participants from each country
Korea	23
China	18
Mexico	7
Peru	7
Uzbekistan	6
Colombia	4
Ecuador	4
Pakistan	4
Poland	3
Croatia	2
Germany	2
Italy	2
Hong Kong	2
Taiwan	2
Bangladesh	1
Bolivia	1
Chile	1
Costa Rica	1
Dominican Republic	1
El Salvador	1
Guatemala	1
Greece	1
Israel	1
Japan	1
Kazakhstan	1
Puerto Rico	1
Tajikistan	1
Thailand	1

Table 3. Native languages of the 100 non-native speakers of English who completed the conversations for lexicon development.

Native language	Number of participants speaking each native language
Spanish	29
Korean	23
Chinese	19
Russian	8
Fujianese	3
Polish	3
Urdu	3
Croatian	2
German	2
Italian	2
Arabic	1
Bengali	1
Greek	1
Hebrew	1
Japanese	1
Thai	1

being non-Hispanic or Latino, while 28 participants identified as Hispanic or Latino. Of those participants who reported their race, 50 identified as Asian, while 45 identified as White. The majority of our participants spoke Spanish, Mandarin, or Korean as their native language. It was important for us to have a very diverse group of cultures and native language backgrounds so that our sentence materials would not be biased toward one ethnic group or one native language background. The language background of our participants reflects demographic projections by the U.S. Census Bureau that between the years of 2000 and 2050, the Hispanic population in the United States will increase by 188%, while the Asian population will increase by 213%. Individual background information of the 100 participants is provided in Supplemental Material Table 1.

The average age at which participants began learning English was 12.8 years old (range = 8–39 years old). Participants had an average of 5.4 years of formal English language classes in their home country (range = 0–18 years). Forty-six participants had no formal English education; four began their English education in primary school, 15 in secondary school, 22 in college, and 13 in graduate school. Fifty-nine participants were full-time students (it should be noted that, of these, many were "nontraditional" undergraduates; Queens College has, for example, many students returning to school to change careers or who began college later in life), 16 identified as white-collar workers (including a microbiologist, a lawyer, an accountant, and an occupational therapist), 23 identified as blue-collar workers (including an electrician, a waiter, and a room attendant), and two were retired.

Sentence Development

A second major concern in developing these new materials was the need for the non-native listeners to be familiar with the syntactic structure in which the lexical items or keywords occurred. To ensure the relative simplicity of the syntactic structures and consistency across sentences, we used a predetermined syntactic frame to create the sentences. The basic syntactic frame that was used to develop the sentences consisted of the following obligatory and optional (in parentheses) word categories: (D), (A), N or Pro, (Adv), V, (Adv), (P), (D), (A), (N), where D = determiner, A = adjective, N = noun, Pro = pronoun, Adv = adverb, V = verb, and P = preposition. No complex syntactic frames with, for example, embedded or preposed dependent clauses, or with complementizer phrases, were included. In addition, only two members of the category could be coordinated using the form of *X and X,* such as Noun and Noun. Grammaticality, meaning, and the number of keywords (i.e., the words within the sentence that serve as the

targets for the sentence-recognition materials; these are shown in Table 4 and Supplemental Material Table 2 using capital letters) determined how many and which of the optional categories were included. There were 12 variants of the basic structure that were the most common and accounted for 70% of the BEL sentences. An example of the 12 variants is shown in Table 4. The remaining 20% of the sentences, though less common, were also similar types of variants of that original syntactic frame. Using the original syntactic frame provided us with sentences approximately equal in length (in terms of the number of total words) and complexity and an equal number of keywords. All keywords were derived from the master lexicon (described above). The complete list of the 500 sentences, each composed of five to seven words with four keywords, can be found in Supplemental Material Table 2.

Once 500 sentences were written, 16 native speakers of English (11 women, 5 men; mean age = 31 years old) were asked to read the sentences. They were asked to flag any sentences that were confusing, odd, did not make sense, or did not sound like a typical English sentence. Two lab members (including the first author) reviewed comments from the 15 native-English-speaking readers. Although all comments were taken into consideration, only those sentences that more than one reader critiqued were edited. An example of a native-reader critique was, "It sounds more natural to start this sentence with 'My Uncle' rather than 'The Uncle.'" A more substantive comment was made with respect to the following sentence: "The ARMY and GENERAL FOUGHT a BATTLE" (in which the keywords of the sentence are capitalized). Three separate readers commented that the combination of "the ARMY and GENERAL" sounded unnatural. This sentence was eventually edited to "The STRONG ARMY WON the BATTLE." More than one native-English-speaking reader critiqued 40 of the 500 sentences. All 40 sentences were edited or rewritten to address the readers' concerns.

Subsequently, a group of 15 non-native English speakers (10 men, 5 women; mean age = 45 years old) were asked to read the revised sentences. The average age at which these readers began learning English was 22 years old. The native languages of these readers included German, Igbo, Mandarin, Polish, and Spanish. These readers were also asked to flag any sentence that did not make sense to them or any sentence in which they did not understand the vocabulary used. All comments by these readers were taken into consideration. Words were replaced for any sentence that even one reader indicated he or she was not familiar with or if the reader indicated the sentence did not make sense. The majority of comments non-native readers made indicated difficulty with the whole meaning of the sentence and how it related to the real world, or at times the non-native readers pointed out incorrect use of grammar (which in actuality was correct).

One example of an original sentence with which several non-native readers had difficulty is "The FISH SWAM SLOWLY in the BOWL," which received comments such as "I don't understand the word 'bowl' used here" and "This sentence sounds strange." The sentence was edited to "The FISH SWAM SLOWLY in the LAKE." The common term "fishbowl" makes the use of "bowl" in

Table 4. Twelve syntactic structures accounting for 70% of the structures used in the 500 Basic English Lexicon sentences.

Syntactic structure	Example sentence using described structure	Number of times used
DANVDN	The DARK CLOUD COVERED the SKY.	47
DNVAN	The GOAT EATS DRY LEAVES.	47
DANVN	The GROCERY STORE SELLS FOOD.	38
DNVDAN	The PLAYER KICKED the SOCCER BALL.	36
DANVAandA	The PINK PIG is FAT and LAZY.	28
DANVPDN	The TIRED CHILD CRIED for her MOTHER.	28
DANVAdvA	A HAPPY MARRIAGE is VERY IMPORTANT.	25
DANVA	The BOILED FISH SMELLS BAD.	23
DNVPDAN	The ARTIST DREW on the YELLOW PAPER.	23
DNVDN	The EGGS NEED MORE SALT.	20
DNVPDN	The LADY WALKED DOWN the STREET.	20
DANVAdv	The ORANGE FIRE BURNED BRIGHTLY.	14

Note. The structure, an example sentence using the structure (with scoreable keywords shown in capital letters), and the number of times the structure was used are shown. Word types are as follows: D = determiner; A = adjective; N = noun; V = verb; P = preposition; Adv = adverb.

this context seem natural to a native speaker. However, if a non-native speaker is unfamiliar with the term "bowl," most likely understood as kitchenware, the use of this word in this context would seem very strange. Another example included the original sentence "The COOK MADE FLAT NOODLES." Three non-native readers stated the following comments: "I don't understand 'flat' in this sentence"; "'Cook' doesn't make sense in this sentence"; and "It should be 'cooker' instead of 'cook.'" The sentence was edited to "The CHEF MADE FRESH NOODLES." The use of "cook" as a noun is seemingly common to a native speaker of English but potentially confusing to a non-native as "cook" is often initially learned as a verb. In addition, although "flat" is a fairly easy adjective, its use with food may be a bit atypical. The substitution of "fresh" was chosen since this word is more commonly associated with food. A final example of a comment made by one reader that did not result in the sentence being edited was in response to the following sentence: "She WASHED and DRIED her CURLY HAIR." The non-native reader commented, "It should be 'hairs' instead of 'hair.'" No changes were made since this was an erroneous grammar edit, and the reader did not have difficulty understanding the meaning or vocabulary used within the original sentence. In total, 85 sentences were modified.

BEL Sentence List Development

The final 500 sentences were divided into 20 different test lists (each including 25 sentences and 100 keywords) based on vocabulary (including the rate of occurrence of each word in the master lexicon), syntactic structure (distributing different types of structures across lists), syllable counts, and high-frequency speech information (i.e., the distribution of fricative and affricate sounds to account for perception difficulties for listeners with high-frequency hearing loss). Function words were not included in the affricate and fricative count. Each list of 20 sentences contained 100 target words. Across the 2,000 keywords, 939 unique keywords were used. Ideally we desired that all 2,000 keywords be unique, that is, never repeated across any of the sentences. Having zero repeatability of keywords was a desired criterion because hearing a specific key word in one listening condition could prime a listener to correctly repeat that same word in another potentially more degraded listening condition. However, in reality with our constraints of using only vocabulary from our non-native elicited English lexicon, this criterion was, in fact, impossible. There were only a limited number of words we could use without repeating to generate a large number of sentences with four keywords that were also grammatically and semantically correct. This is a problem not specific to our sentences alone, however; it is a common problem both in BKB sentences (in which the developers of these sentences also had similar constraints that they were creating their stimuli based on naturally produced speech of hearing-impaired children) and the Harvard IEEE sentences (IEEE Subcommittee, 1969) (which did not necessarily have a lexicon restraint, but the developers were faced with trying to create 720 sentences with 3,600 total keywords). The amount that the keywords do repeat across our sentences is in alignment with these popular sentence lists in that 47% of our keywords are unique, whereas 47% and 52% of keywords are unique in the BKB and IEEE lists, respectively.

Recordings

Two native-English female talkers (26 and 27 years old) and one native-English male talker (age 34) each recorded the complete set of 500 sentences in a sound-treated room. Sentences were digitally recorded at a 44.1 kHz sampling rate with 16-bit resolution. A custom-designed software program created in MaxMSP (Cycling, 74' Version 5.0, 2008) using an Apple iMac computer connected to a MOTU Ultralite mK3 digital–analog convertor and a Shure SM81 cardioid condenser microphone with pop filter attached, placed 12 in. (parallel) from the talkers' lips, were used to record the sentences. The talkers were instructed to speak naturally. The text for each sentence was presented to the talker via a 27-in. Apple LED cinema display that could be visualized clearly through the double-paneled window of the sound-treated room. The talkers could also see a VU meter that indicated the target talking sound level. The 500 recordings were digitally edited using SoundStudio to remove silence at the beginning and at the end of each sentence. Some sentences were rerecorded because of speech disfluencies or extraneous noises in the recording. All files were then root-mean-square (RMS) equalized, using Praat (Boersma & Weenink, 2011), to the same pressure level of 0.1 Pa. The length (in seconds) of each wave file for all three talkers is reported in Supplemental Material Table 2. All sentence recognition testing reported in this article was conducted using the recordings obtained by the first female talker. No experimental testing has been conducted using the other two talkers at this point. All of the high-quality audio recordings are available for use by request to the first author.

Listening Test Procedure

To determine whether the 20 lists we created contained sentences and lexical items that were equally easy or difficult to recognize and that performance across lists would be correlated under identical listening conditions, we asked native speakers of English to listen to the sentences in noise. We decided to focus initially on

native speakers of English in order to obtain data about the relative difficulty across sentence lists on a more homogeneous participant pool (and therefore at this time, data for non-native speakers of English and hearing-impaired listeners will not be included). Seventeen normal-hearing native-English-speaking listeners (mean age = 23 years, *SD* = 4 years; 6 men, 11 women) participated. The Institutional Review Board at Queens College of the City University of New York approved all listening test procedures. Listeners were paid for their participation and provided written informed consent. Prior to participation, all listeners had an otoscopic evaluation to ensure clear ear canals. All listeners had normal hearing thresholds (<20 dB HL) bilaterally tested with standard clinical audiological procedures (American Speech-Language-Hearing Association, 2005) at octave frequencies between 250 and 8000 Hz on a two-channel Grason Stadler clinical audiometer and TDH headphones.

Listeners were seated in a comfortable chair in a double-walled, sound-treated room. Stimuli were passed to an MOTU Ultralite mK3 digital–analog convertor, passed through a HeadAmp 6 Pro headphone amplifier, and output to Etymotic ER1 insert earphones with disposable foam insert eartips (13 mm) attached. Each listener was presented with all 500 sentences, separated into 20 different lists of 25 sentences. The presentation order of the lists was randomly selected. The level of the target sentences was fixed at 65 dB SPL, measured using a GRAS Sound and Vibration IEC Ear Simulator coupled to a preamplifier. Sound pressure levels were based on the average RMS pressure of the sentence files. The sentences were presented in the presence of noise (16 bit, 22 kHz sampling rate) spectrally matched to the long-term average spectra (LTAS) of the 500 test sentences. The noise was generated using MATLAB by passing a Gaussian white noise through a Finite Impulse Response filter with a magnitude response equal to the LTAS of the 500 sentences spoken by the first female talker. The level of the competing noise was fixed at 70 dB SPL, providing a signal-to-noise ratio (SNR) of –5 dB. The target speech and noise masker stimuli were mixed in real time with custom software created using MaxMSP running on an Apple iMac computer. One target sentence was played on each trial, and a random portion of the 60-s noise masker was presented 1 s longer than the target sentence (500 ms prior to the beginning of the sentence and 500 ms at the end of the sentence).

Listeners were first presented with eight sentences (from the Revised BKB sentence materials) spoken by the same female talker used to record our target sentences. This allowed listeners a practice period listening in noise and familiarizing themselves with the talker's voice and the task before the experimental testing began. Listeners were presented the practice sentences at 5, 0, and –5 dB SNR. The number of sentences presented at each SNR varied depending on the participant's performance (i.e., if listeners were having difficulty with the task, they were allowed extra sentences at the easier SNR before moving to the more difficult SNR; otherwise, the first three sentences were presented at 5 dB SNR, the next three at 0 dB SNR, and the last four sentences at –5 dB SNR). Listeners came in on two separate occasions to complete all of the testing. All listeners completed their second test session within 7 days of the first test session, and on average the two test sessions were separated by 4 days. Subjects completed 10 lists per day.

Listeners were asked to repeat the sentences that they heard. Specifically, they were instructed to repeat any word that they heard, regardless of whether the sentence made sense or they had missed portions of the sentence. Listeners' responses were scored online by an examiner seated outside of the double-walled sound-treated room. Listeners' responses were also digitally recorded using a handheld Sony digital-voice recorder with an external lapel microphone. The recordings were scored for reliability purposes by a second examiner. Scores that were not in agreement between the two examiners were reassessed by a third examiner, and a score was agreed on.

Results

The specific question we addressed with this listening test was whether the word recognition success was similar across the 20 lists we created. In other words, we wanted to know whether the 20 lists were equally easy or difficult to recognize and that performance across lists was correlated under identical listening conditions.

As reported earlier, all online scoring was reliability checked, and agreement between the two testers occurred in 96% of the total trials. A third examiner was used to determine the final score in the event of scorer disagreements. Average data and standard deviations from the 17 listeners are presented in Table 5 along with the distribution of syllable count, high-frequency count, and rate of occurrence of keywords within the lexicon used to equally divide the sentences across lists. A repeated measures analysis of variance (ANOVA) was used to test differences in performance across lists. The ANOVA was significant at the $p < .05$ level, $F(19) = 11.15$, $p < .001$. To determine significant differences between the lists, we used post hoc pairwise comparisons based on estimated marginal means with a Bonferroni adjustment for multiple comparisons. These results indicated that 17 of the 20 lists were not significantly different from each other in difficulty with respect to mean performance for

Table 5. Distribution of syllable count, high-frequency count, rate of occurrence in lexicon of keywords, and performance scores (% correct) for normal-hearing, native-English-speaking listeners.

List number	Mean per list			Original list performance (n = 17)		Revised list performance (n = 20)	
	Syllable count	High frequency count (fricatives plus affricates)	Rate of occurrence of keywords	*M*	*SD*	*M*	*SD*
1	8.16	3.24	51.60	75.35	7.2	77.30	7.6
2	8.44	3.16	60.02	79.06	8.9	76.70	5.4
3	8.19	3.16	52.49	76.47	9.9	76.10	8.5
4	8.32	3.16	56.66	73.18	8.3	70.80	10.7
5	8.32	3.16	58.36	77.06	6.3	77.10	5.3
6	8.40	3.12	60.62	75.82	7.4	73.95	7.6
7	8.28	3.08	57.46	74.35	7.9	75.90	8.1
8	8.32	3.12	62.76	73.88	7.8	72.85	7.4
9	8.44[a]	3.04[a]	69.54[a]	63.35*	7.9	76.20	6.2
	8.56[b]	3.24[b]	75.80[b]				
10	8.40	3.20	59.66	77.24	9.3	75.50	8.4
11	8.36	3.08	65.42	78.59	7.8	78.60	6.2
12	8.32	3.08	64.46	80.29	7.4	77.45	6.8
13	8.44	3.04	60.66	74.94	8.7	72.75	6.1
14	8.44	3.20	59.86	73.88	8.9	70.50	9.8
15	8.40	3.20	60.94	71.65	8.6	71.35	8.2
16	8.36	3.12	54.56	79.53	8.7	77.80	5.9
17	8.28	3.20	65.56	78.53	7.0	75.40	8.8
18	8.48[a]	3.08[a]	53.78[a]	84.18*	7.5	73.20	8.5
	8.50[b]	3.04[b]	43.50[b]				
19	8.16	3.12	61.02	80.41	9.6		
20	8.36[a]	3.20[a]	62.76[a]	85.47*	8.3	77.00	5.6
	8.16[b]	3.04[b]	68.04[b]			72.90	8.2

*p < .0025.

[a]Data before revisions were made to these lists in order to equate list difficulty. [b]Data after revisions were made to these lists in order to equate list difficulty.

each list. Three of the lists were significantly different (List 9 yielded significantly lower scores compared with all lists except Lists 4 and 15, and Lists 18 and 20 yielded significantly higher scores compared with Lists 4, 8, 9, 14, and 15 and Lists 4, 7, 8, 9, 13, 14, and 15, respectively) than the other lists (p values ranging from < .001 to .044).

Redistribution of Sentence Lists 9, 18, and 20

We redistributed the sentences within Lists 9, 18, and 20 in an attempt to equate the difficulty of these three lists. Sentences were redistributed based on the performance of the 17 native-English-speaking listeners and the content of each sentence (syllable count, high-frequency count, syntactic structure, and rate of key word occurrence within the developed lexicon). That is, we used sentence performance to redistribute sentences that listeners performed consistently poorly on with sentences that many listeners consistently performed well on within Lists 9, 18, and 20, while maintaining similar distributions for syllable count, high-frequency count, syntactic structures, and rate of key word occurrence according to the overall design features. The distribution was based on modeling sentence recognition performance for each list, which resulted in mean performance scores based on the other 17 lists.

Perception Testing After Revision of Lists 9, 18, and 20

An additional 20 native-English speaking listeners (mean age = 23 years, *SD* = 5 years; 3 men, 17 women) with normal audiometric thresholds were tested on the recognition of all 500 sentences (20 lists). Average data and standard deviations from the 20 listeners are presented in the last two columns of Table 5. Table 6 illustrates mean and median performance, standard error, and the lower and upper bound of a 95% confidence interval for performance for each list.

Table 6. Mean and median performance of the final 20 lists.

List number	*M*	*Mdn*	*SE*	95% CI Lower bound	95% CI Upper bound
1	76.41	76.00	1.08	74.50	80.10
2	77.78	79.00	1.37	73.06	80.34
3	76.27	78.00	1.49	72.13	80.07
4	71.89	73.00	1.59	65.77	75.83
5	77.08	79.00	0.94	74.63	79.57
6	74.81	75.0	1.22	70.40	77.50
7	75.19	74.00	1.31	72.09	79.71
8	73.32	75.00	1.23	69.41	76.29
9	76.20	75.00	1.39	73.28	79.12
10	76.30	78.00	1.44	71.59	79.41
11	78.60	81.00	1.13	75.71	81.49
12	78.76	78.00	1.08	74.26	80.64
13	73.76	76.00	1.26	69.89	75.61
14	72.05	74.00	1.55	65.89	75.11
15	71.49	72.00	1.35	67.53	75.17
16	78.59	80.00	1.19	75.06	80.54
17	76.84	80.00	1.32	71.27	79.53
18	73.20	74.00	1.91	69.20	77.20
19	78.57	78.00	1.26	74.36	79.64
20	72.90	73.00	1.83	69.08	76.72

Note. No one list was significantly different from the rest in terms of mean performance. However, four pairs of lists were significantly different from each other. CI = confidence interval.

We conducted a repeated measures ANOVA to assess differences in mean performance across the 20 lists. The repeated measures ANOVA was significant at the $p < .05$ level, $F(19) = 4.409$, $p < .001$. To determine which list pairs had significantly different performance scores, we used post hoc pairwise comparisons based on estimated marginal means with a Bonferroni adjustment for multiple comparisons, which indicated that four list pairs were significantly different at the $\alpha = .05$ level. List Pairs 11 and 14 ($p < .001$), 11 and 15 ($p = .008$), 14 and 16 ($p = .009$), and 15 and 16 ($p = .010$) were significantly different from each other (significant differences for list pairs are shown in Table 7).

Table 7. List pairs that should not be used together.

	List 11	List 14	List 15	List 16
List 11		*	*	
List 14				*
List 15				*

Note. Asterisks denote two lists with significantly different mean performance scores.

To determine whether performance across lists was correlated under identical listening conditions, we calculated correlations between performance scores. Bivariate correlations for all 20 lists indicated that performance scores for the large majority of list pairs were significantly correlated. The performance scores of the two groups of listeners (those that listened to the original 20 lists and those that listened to the 20 lists with Lists 9, 18, and 20 revised) were used for these computations. Therefore, the sample size for these correlations was 37 for Lists 1–8, 10–17, and 19. A sample size of 20 was used for the correlations including Lists 9, 18, and 20. Pearson coefficients for all 210 list pairs are shown in Table 8.

General Discussion

The spontaneous speech of 100 non-native English speakers was used to create 500 new sentences for speech-recognition testing aimed at both native and non-native listeners. Though linguistic biases can never be removed from open-set speech recognition testing for non-native speakers of English, these materials were designed to provide a less linguistically confounding test tool. Twenty new test lists were created with 25 sentences each, providing 2,000 scoreable keywords. Recognition performance of native-English speaking, normal-hearing listeners at a fixed SNR of –5 dB indicated that overall the lists were equally difficult, and the majority of performance scores between lists were significantly correlated.

There has been tremendous progress understanding non-native speech perception in adverse listening conditions in recent years (e.g., Bradlow & Pisoni, 1999; Cutler et al., 2008); however, there are currently no standards for determining whether poor speech recognition scores obtained by non-native listeners on English-language speech recognition tests are due to their linguistic inexperience with the English language or various auditory impairments. It has been suggested by many (e.g., Gelfand, 2009; Nakamura & Gordon-Salant, 2011) that one way to deal with this issue is to test non-native speakers of English in their native language. To this end, new derivations of the Hearing in Noise Test were created in eight languages (Soli & Wong, 2008). This work is helping to fill a much-needed void in speech-in-noise assessments conducted in languages other than English. However, there are several disadvantages to this approach for testing in the United States. First, all clinics would need to be equipped with speech-recognition tests in the various languages. Second, test materials may not be available in a listener's native language. Third, the examiner would need to be fluent (or at least highly proficient) in the test language in order to score the listener's responses or at a minimum have access to a trained interpreter

Table 8. Correlations for performance between lists.

List	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	.538**	.458**	.500**	.541**	.473**	.538**	.507**	.759**	.561**	.408*	.259	.411*	.320	.546**	.359*	.460**	.433	.339*	.503*
2		—	.599**	.531**	.679**	.609**	.532**	.519**	.623**	.505**	.469**	.346*	.355*	.424**	.595**	.313	.589**	.463*	.486**	.321
3			—	.601**	.558**	.445**	.319	.513**	.441	.460**	.290	.406*	.311	.265	.432**	.300	.518**	.494*	.303	.602**
4				—	.666**	.645**	.662**	.731**	.499*	.455**	.530**	.489**	.508**	.577**	.668**	.425**	.646**	.298	.481**	.320
5					—	.706**	.657**	.645**	.554*	.592**	.529**	.414*	.458**	.514**	.640**	.386*	.509**	.324	.406*	.211
6						—	.634**	.587**	.634**	.474**	.500**	.349*	.206	.470**	.588**	.392*	.603**	.387	.419**	.331
7							—	.630**	.510*	.556**	.653**	.383*	.469**	.606**	.625**	.567**	.517**	.598**	.355*	.285
8								—	.563*	.613**	.485**	.486**	.558**	.549**	.584**	.434**	.505**	.220	.475**	.300
9									—	.743**	.534*	.364	.416	.477*	.635**	.395	.447*	.570**	.172	.432
10										—	.440**	.428**	.503**	.507**	.598**	.498**	.503**	.682**	.277	.555*
11											—	.442**	.692**	.744**	.869**	.805**	.601**	.507*	.510**	.704**
12												—	.586**	.598**	.475**	.421**	.577**	.476*	.451**	.500*
13													—	.665**	.697**	.631**	.521**	.434	.599**	.560*
14														—	.730**	.742**	.707**	.448*	.497**	.571**
15															—	.725**	.733**	.533*	.565**	.644**
16																—	.548**	.568**	.425**	.655**
17																	—	.507*	.634**	.610**
18																		—	.452*	.588**
19																			—	.559*
20																				—

Note. Pearson product coefficients are based on a sample of 37 normal-hearing listeners. Correlations including Lists 9, 18, and 20 were conducted with a sample size of 20. These three lists are highlighted in light gray.

*$p \leq .05$. **$p \leq .01$.

who could score the responses. Last, we argue that it is imperative to consider those listeners who communicate predominantly in English but are non-native speakers of the language (e.g., someone born and reared outside of the United States but who now works, shops, dines, and so forth in an English-speaking community). For these listeners, their communication disorder is most likely being experienced predominantly in English because they are using English, not their native language, throughout the course of their day. Therefore, evaluating speech recognition in English, rather than their native language, may be functionally more relevant.

As a first step in assessing list equivalency for the new recognition materials, we conducted perception testing on native-English speakers with normal hearing. We opted to focus on native speakers of English in order to gain an insight into the relative difficulty across sentence lists on a more homogeneous subject pool. The use of normal-hearing, monolingual English speakers allowed us to probe whether there were any material- and/or signal-related variations within the sentences or recordings that could cause significant differences in performance across lists.

Correlations between performance on one list compared with performance on a different list showed that out of 210 list pairs, 87% of the correlations were significant at an alpha of .05 or greater (see Table 8), indicating equal perceptual ease or difficulty across the majority of lists. The sample size used for correlations for Lists 9, 18, and 20 was smaller ($n = 20$) than the sample size used for the remaining lists ($n = 37$). When these lists were not included in the analyses, significant correlations between list pairs increased to 92%. These data can be used when deciding on lists for research designs that maximize the equivalency across lists.

It is important to note that the original SPIN sentences, though equivalently difficult for normal-hearing listeners, were not found to be equivalent in difficulty for hearing-impaired listeners (Bilger et al., 1984). This resulted in a reduced number of revised materials that provided equal difficulty across lists for those listeners with hearing loss. However, we opted not to test non-native English listeners or hearing-impaired listeners in this initial phase. Both groups are heterogeneous by nature, and for a proper analysis with these types of listeners, a large sample size is needed that is beyond the scope of this article. The BEL sentences remain to be tested on a clinical population and a large non-native English speaking population. Until these data are collected, we cannot confirm that all 20 lists are equally difficult for these listener groups. However, throughout the development of these sentences, the demands of hearing-impaired and older listeners were taken into account. We evenly distributed the number of high-frequency phonemes (fricatives plus affricates) across lists to account for perception difficulties for those listeners with high-frequency hearing loss (see Hedrick, 1997; Zeng & Turner, 1990). We also kept the length of the sentences within seven words to help make these lists a potential test tool for older listeners or other populations with cognitive or memory limitations that may lead to greater difficulty recalling longer strings of words (see Wingfield, Poon, Lombardi, & Lowe, 1985). It is our hope that these considerations will allow us to use the BEL sentences for perceptual testing with diverse listener groups. However, it may be that after more data are collected, a revised version of the BEL sentences will be needed for use with specific clinical populations.

A possible shortcoming of the BEL sentences is that all non-native speakers used to create the lexicon were residents of Queens, New York. Therefore, their vocabulary could be specific to American English speakers and non-natives living in urban areas. However, during the creation of the sentences, we tried to avoid words that participants used that were specific to the New York City metropolitan area, such as "subway" or "transit." We also intentionally asked our talkers to elaborate on experiences within their home country in the hope that this would broaden the range of vocabulary used. We believe that the resulting lexicon from which we derived keywords for BEL sentences reflect more general knowledge and experiences that would be common to many U.S. residents.

Planned future work in our laboratories includes the collection of a large set of normative data from 100 non-native speakers of English for all 20 lists. The findings from this project extension will be able to directly address some of the issues raised above. As expected, our pilot data show a large variability between participants, as a result of differences in linguistic history and English experience. It is important to note, though, that a similar pattern of mean performance between lists was observed for non-native listeners compared with the native results reported above.

Finally, it is important to keep in mind that the perception data reported were collected using only one talker (SR). No experimental testing has been completed with the other two talkers, although recordings are available for all three talkers. The data reported above are specific to the acoustics of talker SR's recordings, and it cannot be assumed that the reported results will be replicated with the other talkers. Out of our three talkers, SR had the slowest rate of speech (see Supplemental Material Table 2), which may aid in non-native speech perception. Furthermore, all perception testing was conducted at one level of distortion (–5 dB SNR). It is possible that at other SNRs or in other types of background noise (e.g., informational maskers), list equivalency may

vary from that reported above. Healy and Montgomery (2007) suggested that sentences that are more or less difficult with respect to recognition in noise tend to retain their relative level of intelligibility even when overall performance level changes significantly and for different levels of listener ability. They caution, however, that sentences may not retain their relative level of intelligibility when sentence intelligibility is compared across different types of signal distortion. It is therefore crucial to make these sentences available to be tested in other experimental paradigms such that we can gain more insight into the various talker-, listener-, and signal-related factors that affect word recognition for these materials.

Conclusions

The goal of this project was to develop new sentence materials that could be used to test speech recognition for various listener populations. In our approach, we were guided by two major considerations: one, to use lexical items and syntactic structures appropriate for use with non-native listeners and, two, to generate a large set of materials that could be used in multiple experimental conditions and with difficult-to-recruit listener populations. We were able to address these considerations by deriving a large basic lexicon from ecologically valid communicative situations (i.e., spontaneous dialogues) with non-native speakers and embedding these lexical items in simple syntactic frames. The resulting set of materials consists of 20 lists of 25 sentences for a total of 2,000 target keywords. The development of these test materials was a major methodological goal of the present study. An important finding reported in this article is that native-English listeners found these lists to be equivalent in difficulty. Although these results provide important knowledge about the test materials in this stage of the project development, we are hopeful that similar results will be obtained for non-native and other clinical populations when tested using the BEL sentences.

Acknowledgments

This work was supported in part by the Grant Program for Projects on Multicultural Affairs from the American Speech-Language-Hearing Association and the Undergraduate Research/Mentor Experience at Queens College. This project could not have been completed without the tremendous commitment from the undergraduate and graduate students in the Speech and Auditory Research Lab at Queens College and the UTsoundLab at the University of Texas at Austin. The authors are especially thankful to Stacey Rimikis for her incredible dedication to this project. Portions of this project were presented at the 2010 American Speech-Language-Hearing Association convention in Philadelphia, PA; the 160th meeting of the Acoustical Society of America in Cancun, Mexico, in 2010; and the 2011 New York State Speech-Language-Hearing Association convention in Saratoga Springs, NY. A preliminary report of this project was published by Lauren Calandruccio in *The ASHA Leader* (October 12, 2010, edition).

References

Adler, S. (1990). Multicultural clients: Implications for the SLP. *Language, Speech, and Hearing Services in Schools, 21,* 135–139.

American Speech-Language-Hearing Association. (2005). *Guidelines for manual pure-tone threshold audiometry.* Available from www.asha.org/policy/.

Ballachanda, B. B. (2001a, April). Audiological assessment of "We, the foreign born." *Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse Populations, 7*(1), 11–13.

Ballachanda, B. B. (2001b). Meeting the needs of multicultural clients. *Advance for Audiologists, 3*(5), 50–53.

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. In J. Bench & J. Bamford (Eds.), *Speech-hearing tests and the spoken language of hearing-impaired children* (pp. 148–187). London, England: Academic Press.

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.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114,* 1600–1611.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27,* 32–48.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.2.39) [Computer program]. Available from www.praat.org/.

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121,* 2339–2349.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America, 106*(4, Pt. 1), 2074–2085.

Chan, C. L. (2012). NU-subdb: Northwestern University Subject Database [Web application]. Chicago, IL: Department of Linguistics, Northwestern University. Retrieved from https://babel.ling.northwestern.edu/nusubdb2.

Crandell, C. C., & Smaldino, J. J. (1996). Speech perception in noise by children for whom English is a second language. *American Journal of Audiology, 5,* 47–51.

Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America, 124,* 1264–1268.

Gelfand, S. A. (2009). *Essentials of audiology.* New York, NY: Thieme.

Healy, E. W., & Montgomery, A. A. (2007). The consistency of sentence intelligibility across three types of signal

distortion. *Journal of Speech Language, and Hearing Research, 50,* 270–282.

Hedrick, M. (1997). Effect of acoustic cues on labeling fricatives and affricates. *Journal of Speech, Language, and Hearing Research, 40,* 925–938.

IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics*, *17,* 227–246.

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.

Nakamura, K., & Gordon-Salant, S. (2011). Speech perception in quiet and noise using the Hearing in Noise Test and the Japanese Hearing in Noise Test by Japanese listeners. *Ear and Hearing, 32,* 121–131.

New York State Comptroller. (2000). *Queens: An economic review.* Available from www.osc.state.ny.us/osdc/rpt1100/rpt1100.htm.

Pinet, M., & Iverson, P. (2010). Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience. *The Journal of the Acoustical Society of America, 128,* 1357–1365.

Shi, L. F. (2010). Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *Journal of Speech, Language, and Hearing Research, 53,* 821–835.

Smiljanic, R., & Bradlow, A. (2011). Bidirectional clear speech perception benefit for native and high proficiency non-native 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 the Hearing in Noise Test. *International Journal of Audiology, 47,* 356–361.

Tabri, D., Chacra, K. M., & Pring, T. (2011). Speech perception in noise by monolingual, bilingual and trilingual listeners. *International Journal of Language & Communication Disorders, 46,* 411–422.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.

U.S. Census Bureau. (2010). *Census 2000 redistricting data (Public Law 94-171) summary file, Tables PL1 and PL2; and 2010 census redistricting data (Public Law 94-171) summary file, Tables P1 and P2. Statistical abstract of the United States, Section 1*. Available from http://2010.census.gov/2010census/data/.

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication, 52,* 943–953.

van Wijngaarden, S. J., Steeneken, H. J., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America, 111,* 1906–1916.

Villchur, E. (1982). The evaluation of amplitude-compression processing for hearing aids. In G. A. Studebaker & F. H. Bess (Eds.), *The Vanderbilt hearing-aid report* (pp. 141–143). Upper Darby, PA: Monographs in Contemporary Audiology.

Wingfield, A., Poon, L. W., Lombardi, L., & Lowe, D. (1985). Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *Journal of Gerontology, 40,* 579–585.

Yund, E. W., & Woods, D. L. (2010). Content and procedural learning in repeated sentence tests of speech perception. *Ear and Hearing, 31,* 769–778.

Zeng, F. G., & Turner, C. W. (1990). Recognition of voiceless fricatives by normal and hearing-impaired subjects. *Journal of Speech and Hearing Research, 33,* 440–449.