Published in final edited form as:

J Am Acad Audiol. 2015 June ; 26(6): 582–594. doi:10.3766/jaaa.14082.

List Equivalency of PRESTO for the Evaluation of Speech Recognition

Kathleen F. Faulkner[*,†], **Terrin N. Tamati**[*], **Jaimie L. Gilbert**[‡], and **David B. Pisoni**[*,†]

[*]Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University, Bloomington, 1101 East Tenth Street, Bloomington, IN 47405

[†] DeVault Otologic Research Laboratory, Department of Otolaryngology, Head and Neck Surgery, Indiana University School of Medicine, 699 Riley Research Drive, RR044, Indianapolis, IN 46202

[‡] Department of Communicative Disorders, University of Wisconsin, Stevens Point, 1901 Fourth Avenue, Stevens Point, WI 54481

Abstract

Background: There is a pressing clinical need for the development of ecologically valid and robust assessment measures of speech recognition. Perceptually Robust English Sentence Test Open-set (PRESTO) is a new high-variability sentence recognition test that is sensitive to individual differences and was designed for use with several different clinical populations. PRESTO differs from other sentence recognition tests because the target sentences differ in talker, gender, and regional dialect. Increasing interest in using PRESTO as a clinical test of spoken word recognition dictates the need to establish equivalence across test lists.

Purpose: The purpose of this study was to establish list equivalency of PRESTO for clinical use.

ResearchDesign: PRESTO sentence lists were presented to three groups of normal-hearing listeners in noise (multitalker babble [MTB] at 0 dB signal-to-noise ratio) or under eight-channel cochlear implant simulation (CI-Sim).

Study Sample: Ninety-one young native speakers of English who were undergraduate students from the Indiana University community participated in this study.

Data Collection and Analysis: Participants completed a sentence recognition task using different PRESTO sentence lists. They listened to sentences presented over headphones and typed in the words they heard on a computer. Keyword scoring was completed offline. Equivalency for sentence lists was determined based on the list intelligibility (mean keyword accuracy for each list compared with all other lists) and listener consistency (the relation between mean keyword accuracy on each list for each listener).

Results: Based on measures of list equivalency and listener consistency, ten PRESTO lists were found to be equivalent in the MTB condition, nine lists were equivalent in the CI-Sim condition, and six PRESTO lists were equivalent in both conditions.

Corresponding author: Kathleen F. Faulkner, Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405; katieff@indiana.edu.

Conclusions: PRESTO is a valuable addition to the clinical toolbox for assessing sentence recognition across different populations. Because the test condition influenced the overall intelligibility of lists, researchers and clinicians should take the presentation conditions into consideration when selecting the best PRESTO lists for their research or clinical protocols.

Keywords

speech perception; speech recognition; speaker variation; individual differences; TIMIT; cochlear implants

INTRODUCTION

spoken word recognition is one of the most robust and highly adaptive information processing skills that humans have developed over the course of evolution (Moore, 2007a, 2007b). Human listeners adapt and compensate rapidly and effortlessly to variations within and across different speakers, including changes in the speaker's gender, age, regional dialect, speaking rate, and speaking style (e.g., Nygaard et al, 1994; Sommers et al, 1994; Nusbaum and Magnuson, 1997). Human listeners are also able to adjust to multiple sources of acoustic degradation in the speech signal such as noise, filtering, and reverberation in their immediate listening environment (Remez et al, 1981; Shannon et al, 1995; Warren, et al, 1995; Stickney and Assmann, 2001; Mattys et al, 2012). The robust ability to rapidly adapt and compensate in adverse listening conditions allows human listeners to successfully recognize and understand speech produced by novel unfamiliar talkers from different geographic regions of origin and talkers with different native languages under an enormously wide range of adverse and challenging conditions (Bradlow and Pisoni, 1999; Clarke and Garrett, 2004; Floccia et al, 2006; Tamati et al, 2013). How listeners accomplish this task with relative ease so quickly and efficiently with few errors or loss of the talker's intended message is a major focus of this research (e.g., Magnuson and Nusbaum, 2007; Ro¨nnberg et al, 2010; Zekveld et al, 2014; Johnsrude et al, 2013). Furthermore, this question is of great clinical importance to individuals with hearing loss who use hearing aids and/or cochlear implants (CIs), clinical populations that are particularly vulnerable to the presence of noise and variability in the vocal sound source (e.g., Fu and Nogaki, 2005; George et al, 2010).

Conventional speech recognition tests were originally designed to assess the limitations of communication equipment used in early telephone systems (Hudgins et al, 1947; Egan, 1948). As pointed out more than 60 yr ago by Licklider and Miller (1951), none of these speech recognition tests were actually designed to assess the robust highly adaptive information processing skills of human listeners or to study the enormous individual differences in human listeners in a wide range of demanding speech communication tasks. Historically, tests of spoken word recognition and sentence understanding were developed with materials that eliminated as many sources of variability as possible, by using a singletalker with a standard unmarked regional dialect and simple materials that were carefully spoken at a slow articulation rate.

One example of a conventional speech recognition test is the Hearing in Noise Test (HINT) developed by Nilsson et al (1994). The HINT consists of 25 lists of short, uniform, high-context meaningful sentences originally designed for use with children (Bench et al, 1979). All of the sentences were spoken by a single male speaker. The HINT was originally designed as an adaptive speech-in-noise procedure yielding the signal-to-noise ratio (SNR) required for a predetermined proportion correct. However, clinicians often administer the HINT by list in quiet or under fixed noise conditions, reporting percent words correct in each condition. Spoken word recognition and sentence perception are routinely assessed in adult users of CIs with the HINT to document outcome and benefit. The HINT was also originally included as part of the Minimum Speech Test Battery (MSTB) for adult CI users (Nilsson et al, 1996; Luxford et al, 2001).

Performance with CIs has improved over time due to advancements in device technology, signal-processing strategies, and the expansion of implant candidacy criteria. Furthermore, many patients are now being considered for implantation with greater residual hearing (Gifford et al, 2010). As a result, many CI candidates and patients easily reach ceiling levels of performance on conventional speech recognition tests like the HINT, therefore requiring more sensitive assessment tools to establish baseline performance (Gifford et al, 2008). The HINT generally overestimates sentence recognition performance because listeners adapt rapidly and benefit from hearing the same talker presented repeatedly in a test list (e.g., Gifford et al, 2008). And, the use of easy, high-context sentence materials provide an artificial boost in sentence recognition performance beyond perceptual processing of the elementary acoustic cues encoded in the speech signal. For example, Gifford et al (2008) compared performance on several conventional speech recognition tests in adult users of CIs, and found that all measures of speech recognition were highly correlated. However, HINT in quiet was a poor predictor of performance on the other tests. Specifically, individual CI users could achieve a score of 100% correct on a list of HINT sentences while their scores on the other tests encompassed the entire distribution of scores (e.g., from 20% to 94% on Consonant-Nucleus-Consonant word lists and 3.75–18.5 dB SNR on the Bamford-Kowal-Bench Speech-in-Noise test). Similar findings have been reported in NH and hearing-impaired listeners with the HINT (Wilson et al, 2007). Because of ceiling effects with HINT sentences in quiet, more robust and challenging test materials are needed for evaluating candidacy and tracking outcome and benefit after implantation.

To address this need, several new sentence recognition tests have been developed that are more difficult for CI users (Spahr and Dorman, 2004; Spahr et al, 2012; King et al, 2012; Boyle et al, 2013). These assessment instruments incorporate a greater range of stimulus variability to eliminate ceiling effects thereby allowing for the measurement of changes in performance over time, especially in higher performing patients. The MSTB for adult CI users has been updated recently (MSTB, 2011) to reflect changes in the performance of current CI users and includes the AzBio test as the standard sentence recognition test presented in quiet and in multitalker babble (MTB) noise. The goal of the revised MSTB was to eliminate ceiling effects and to set standards for best practices and drive uniformity in assessment allowing for more accurate across-clinic comparisons and longitudinal tracking of outcomes.

Everyday speech recognition under real-world conditions requires listeners to rapidly adapt to changes in the talker and background listening conditions. Further, it is well known that individual talkers are not equally intelligible, and that speech recognition varies as a function of the speaker's gender, native accent, regional dialect, rate of speech, and intensity level. One of the reasons the AzBio test is more difficult compared with the HINT is the use of multiple talkers. The AzBio test has four talkers: two male and two female. However, the same four talkers are used repeatedly throughout the test and listeners may learn the dynamics of the talker's vocal tract transfer function and vocal source features after only a few exposures to the same talker (e.g., Nusbaum and Morin, 1992; Nygaard and Pisoni, 1998; Bradlow and Pisoni, 1999; Rosenblum et al, 2007; Levi et al, 2011).

The Perceptually Robust English Sentence Test OpenSet (PRESTO) is a new high-variability sentence recognition test that maximizes talker variability (Gilbert et al, 2013; Tamati et al, 2013). PRESTO uses sentences that were read aloud by different talkers. PRESTO was originally designed to assess the capacity to rapidly learn and adapt to high-variability speech. The PRESTO lists contain multiple sources of variability, including talker, gender, and regional dialect, while controlling for several lexical characteristics, such as frequency and familiarity that have been found to affect spoken word recognition. PRESTO is more challenging than conventional sentence recognition tests because the multiple talker variability requires additional controlled attentional processing from the listener (Mullennix et al, 1989; Nusbaum and Morin, 1992; Wong et al, 2004). PRESTO is an example of a family of theoretically motivated perceptually robust tests of speech perception and spoken word recognition (Kirk et al, 1995; 1997; Pisoni, 1998; Holt et al, 2011). A key feature of perceptually robust sentence recognition tests is the inclusion of stimulus variability, for example: single versus multiple talkers, variation in speaking rate, lexical competition, or sentence predictability (Kirk et al, 1995; 1997; Sommers 1996; Krull et al, 2010; Tamati et al, 2013; Tamati and Pisoni, 2014). The multitalker variability in PRESTO was included to create a more challenging test of sentence recognition that may better reflect a patient's everyday experience. Further, manipulating the difficulty of the test through the use of high-variability test sentences eliminates ceiling effects under quiet testing conditions for clinical populations.

The original PRESTO sentence test was developed in the Speech Research Laboratory at Indiana University (Gilbert et al, 2013; Tamati et al, 2013). The PRESTO test consists of 19 lists created from sentences selected from the Texas Instruments/Massachusetts Institute of Technology (TIMIT) Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al, 1993). The TIMIT corpus was developed to create a database of materials suitable for acoustic-phonetic research and for the development and testing of automatic speech recognition algorithms (see Klatt, 1979). This database includes recordings of 630 speakers representing eight major dialects of American English, each reading ten phonetically rich sentences. The sentences chosen for the PRESTO lists followed specific guidelines: No speaker was repeated within any test list and no sentences were repeated within or across lists. The gender of the speaker was balanced; half the speakers in each list were male and half female. The talkers' regional dialects were selected from eight different North American geographic regions (New England, Northern, North Midland, South Midland, Southern, New York City, Western, Army Brat). The difficulty of the sentences varied within lists, but not between lists

(measured by word frequency, word familiarity, number of words in the sentences, and number of content words). Each list was composed of 18 sentences with a total of 76 keywords, with an average keyword familiarity of 6.9 (of 7) and log keyword frequency of 2.5 (Nusbaum et al, 1984). Within a list, each sentence was composed of 5–10 total words and 3–6 content words. The keywords in each sentence had a minimum average word familiarity rating of 6.5 (of 7) and a minimum average log frequency score of 1.0 (Nusbaum et al, 1984). Although sentences differed in syntactic structure, all sentences contained one verb phrase.

PRESTO has been shown to have high test–retest reliability (Gilbert et al, 2013) and it has been used recently to explore high-variability sentence recognition abilities in several different populations (Faulkner, Kidd, et al, 2014; Faulkner, Twiggs, et al, 2013; 2014; Tamati et al, 2013; Tamati and Pisoni, 2014). Even within NH listeners, PRESTO has been useful in assessing individual and group differences in high-variability speech recognition. Gilbert et al (2013) demonstrated that PRESTO was sensitive to individual differences in young, NH native adult speakers of English, with overall PRESTO accuracy scores ranging from 40% to 76%. Furthermore, these individual differences were found to be related to severalunderlying,coreneurocognitiveand perceptualabilities (Tamati et al, 2013). PRESTO was also used in a study designed to explore how language background and developmental history affect recognition of high-variability speech (Tamati and Pisoni, 2014). Tamati and Pisoni reported that nonnative speakers of English performed much more poorly on both PRESTO and HINT compared with native speakers of English, but were more adversely affected by the highvariability nature of PRESTO sentences.

While PRESTO is currently in use in research and clinical protocols and has shown high test–retest reliability, list equivalency has not yet been established. Given its current design, the recommended use of PRESTO is by test list. As such, it is now important to establish list equivalency so that performance can be compared both between and within listener groups and to ensure that differences in scores across conditions or test sessions do not simply reflect differences in test lists. Furthermore, because PRESTO was constructed to be useful in evaluating the candidacy of borderline CI candidates as well as measuring outcome and benefit with a CI over time, establishing list equivalency under CI simulation conditions was also an additional important objective. The purpose of this study was to evaluate the PRESTO sentence lists for equivalency with young adult NH listeners to establish list recommendations for clinical and research use.

METHODS

Participants

This study consisted of three phases (specific phase details provided in the following section) with 91 total participants recruited from the Indiana University undergraduate student community: Phase I-A: N 5 21 (mean age: 19.4, range 18–26 yr old, 18 female), Phase I-B: N 5 18 (mean age: 19.6, range 18–22 yr old, 13 female), Phase II: N 5 26 (mean age: 21.9, range 18– 36 yr old, 13 female), and Phase III: N 5 26 (mean age: 22.4, range 20–30 yr old, 21 female). All participants were native speakers of American English, and reported no history of hearing or speech disorders at the time of testing. All participants also

passed a pure-tone hearingscreening test at 25 dB HL from 250–8000 Hz for both ears. Participants in Phases I and II were given course credit for their participation; participants in Phase III received $15 for 1.5 h of participation. All participants were informed about the test protocol and procedures, and written consent was obtained prior to participation. The institutional review board approved the informed consent and procedures employed in this study and all testing took place in the Speech Research Laboratory at Indiana University in Bloomington, Indiana.

Materials and Procedures

Participants were tested in small groups of four or fewer, seated in a quiet room in an enclosed testing carrel. All test stimuli were presented through PowerMacG4 computers running MacOS 9.2, using experimental programs controlled by PsyScript 5.1d3 scripts. Immediately after the hearing screening, participants completed the sentence recognition task. Participants were presented with sentence materials binaurally through Beyer Dynamic DT-100 circum-aural headphones. Output levels of the target sentences were calibrated to be approximately 64 dB sound pressure level. All sentences used in Phases I and II were mixed with MTB and presented at 0 dB SNR. Sentences used in Phase III were processed through an eight-channel noise-band vocoder using previously established methods (Shannon et al, 1995) using Tiger CIS (Tigerspeech Technology, Qian-Jie Fu, House Ear Institute). To familiarize participants in Phase III with CI simulated speech (CI-Sim), one practice list of eight-channel vocoded HINT sentences preceded testing with the PRESTO test lists.

Sentences were presented individually, followed by a pop-up dialog box where the participant was required to type in what they heard. Partial answers and guessing were encouraged. The experiment was self-paced, without time limits, and breaks were taken between lists. Testing took approximately 1–1.25 h. All listeners heard each test sentence only once. Scoring was completed offline for keywords correct. Keywords were scored according to the instructions in the PRESTO manual. Exact word order was not required, but plural or possessive morphological markers were required to match the keyword. Deconstructed contractions (e.g., “there is” for “there’s”) and homophones (e.g., “bear” for “bare”) were acceptable responses. Minor spelling errors were also acceptable as long as the error did not result in an entirely different word. Hyphenated keywords were considered to be one keyword, and to be acceptable the response was required to contain all component words. Additionally, partial embedded keywords were incorrect (e.g., “economic” for “socioeconomic”).

The sentences used in Phase I-A were PRESTO Lists: 2, 7, 10, 12, 13, 14, 15, 17, 18, and 22. The sentences used in Phase I-B included the remaining PRESTO Lists: 3, 4, 5, 8, 11, 17, 19, 20, 21, and 23. List 17 was selected as the tenth list for Phase I-B because it was correlated with and was not significantly different from three other lists in Phase I-A. The sentence lists used in Phase II and III were the test lists from Phase I-A and I-B that were determined to be perceptually equivalent and consistent across listeners (criteria are described below in the data analysis section): PRESTO Lists 2, 3, 4 7, 8, 10, 11, 13, 14, 15, 17, 21, and 23. List presentation order was randomized for each participant, but all sentences

in a list were presented in the same order. Figure 1 summarizes the overall study design and shows which lists were assigned to each phase.

Data Analysis

Equivalency for sentence lists was evaluated in two ways: list intelligibility and listener consistency. List intelligibility was determined by comparing the mean keyword accuracy and variance for each list with all other lists—lists that had mean accuracy scores and variances that were similar were deemed to be equivalent. To determine the most equivalent lists within each phase of testing, one-way repeated measures analysis of variances (ANOVAs) and post hoc tests with Bonferroni corrections for multiple comparisons were carried out on the mean scores for all possible list pairings. Listener consistency was assessed by examining the correlations between mean keyword accuracy on each list for individual listeners—lists that were correlated across listeners were deemed to have high listener consistency. To determine the consistency of individual listener performance across lists, individual Pearson correlations were computed.

RESULTS

Phase I

Phase I-A consisted of 10 PRESTO lists selected from Gilbert et al (2013) to minimize repeated talkers. Phase I-B included the remaining nine PRESTO lists and List 17. Within each phase, each list was evaluated for overall list intelligibility and listener consistency compared with the other nine lists. Phase I-A mean PRESTO accuracy was 70.3% (standard deviation [SD] = 4.28). Overall accuracy ranged from 55.5% to 95.0% across all listeners. Listeners were the least accurate on List 5 with mean accuracy at 65.6% (SD = 6.7%), and most accurate on List 21 with mean accuracy at 84.1% (SD = 6.1%). For Phase I-A, a one-way repeated measures ANOVA on keyword accuracy with list as the factor revealed a significant main effect of list [$\boldsymbol{F}(9,180) = 16.0$, $\boldsymbol{p} < 0.001$]. Bonferroni corrected post hoc tests and correlational analyses were performed on all possible list pairings in order to determine the most similar lists. Lists 12, 18, and 22 were removed after this process because they were significantly different from and were not correlated with several other Phase I-A lists. These three lists were the only Phase I-A lists that were not both equally intelligible and consistent with another list in the set.

Phase I-B mean PRESTO accuracy was 74.9% (SD = 6.1%). Overall accuracy ranged from 43.4% to 89.5% across all listeners. Listeners were the least accurate on List 18 with mean accuracy at 62.5% (SD = 8.8%), and most accurate on List 13 with mean accuracy at 76.4% (SD = 7.2%). For Phase I-B, a one-way repeated measures ANOVA on keyword accuracy with list as the factor revealed a significant main effect of list [$\boldsymbol{F}(9,153) = 31.2$, $\boldsymbol{p}<0.001$]. Bonferroni corrected post hoc tests and correlational analyses were also carried out on all possible list pairings in order to determine the most similar lists. Phase I-B lists were more similar to each other than Phase I-A lists. Lists 5, 19, and 20 were removed because they were significantly different from and were not correlated with several other Phase I-B lists. Although List 21 was more intelligible than other lists, it was preferred to List 20 because it was more consistent. Finally, List 17 was not excluded because it met the inclusion criteria

based on the Part I-A analyses. In total, 13 lists were selected for further testing in Phase II and III; six lists were selected from Phase I-A (Lists 2, 7, 10, 13, 14, 15), six lists were selected from Phase I-B (3, 4, 8, 11, 21, 23), and List 17, which was included in both Phase I-A and I-B. For the interested reader who may wish to obtain additional information about the correlations and differences among all the lists used in Phase I, please contact the authors.

Phases II and III

Phase II (MTB) mean PRESTO accuracy was 76.0% (SD = 6.5%). Overall performance ranged from 59.3% to 85.0% across all listeners, with highly consistent mean accuracy across all 13 lists. Listeners were the least accurate on List 3 with mean accuracy at 69.5% (SD = 7.9%), and most accurate on List 21 with mean accuracy at 84.6% (SD = 6.1%). Phase III (CI-Sim) mean PRESTO accuracy was 76.9.0% (SD = 5.6%). Overall accuracy ranged from 66.7% to 86.1% across all listeners, with highly consistent mean accuracy across all 13 lists. Listeners were least accurate on List 10 with mean accuracy at 70.7% (SD = 7.8%), and most accurate on List 4 with mean accuracy at 82.4% (SD = 6.5%). Figure 2 displays mean keyword accuracy and individual scores on all thirteen PRESTO lists tested in MTB (Figure 2A) and CI-Sim (Figure 2B) conditions.

All test lists that were within 65% of the mean keyword accuracy were considered to be equally intelligible. For MTB, List 3 was .5% below the mean keyword accuracy, and List 21 was .5% above mean keyword accuracy. For CI-Sim, Lists 4 and 15 were .5% below the mean keyword accuracy, and List 10 was .5% above mean keyword accuracy. Additional statistical analyses were carried out to quantify list intelligibility and listener consistency.

List intelligibility was assessed by comparing mean keyword accuracy across lists for both MTB and CISim. A one-way repeated measures ANOVA on keyword a SpeechLang Hear ccuracy on lists from MTB revealed a significant main effect of list [$\boldsymbol{F}(12,325)$ $\boldsymbol{p}$ <.29, p , 0.001]. Bonferroni corrected post hoc tests on all possible MTB list pairings revealed that Lists 3 and 21 were significantly different than several of the other MTB lists (see details included in Appendix A). A one-way repeated measures ANOVA on keyword accuracy of lists from CI-Sim also revealed a significant main effect of list [$\boldsymbol{F}(12,325) = 5.55$, $\boldsymbol{p} <$ 0.001]. Bonferroni corrected post hoc tests on all possible CI-Sim list pairings revealed that Lists 4, 7, 10, and 14 were also significantly different from several of the other CI-Sim lists (see details included in Appendix A).

Listener consistency was assessed using correlational analyses on mean PRESTO accuracy on each list. Listener performance was highly consistent across all lists in both conditions (see details included in Appendix B). For MTB, all possible list pairs were highly correlated ($\boldsymbol{r}$= 0.41 to 0.79, all $\boldsymbol{p}\leq$ 0.05), except for List 11, which showed the least consistency compared with the other MTB lists ($\boldsymbol{r}$ = 0.38–0.68, all $\boldsymbol{p} \leq$ 0.055). For CI-Sim, all possible list pairs were also highly correlated ($\boldsymbol{r}$ = 0.39–0.77, all $\boldsymbol{p}\leq$ 0.05), except for List 2, which showed the least consistency compared with the other CI-Sim lists ($\boldsymbol{r}$ = 0.35–0.639, all $\boldsymbol{p} \leq$ 0.079).

Combining the results obtained for list intelligibility and listener consistency, in MTB, the following lists were deemed to be equivalent: Lists 2, 4, 7, 8, 10, 13, 14, 15, 17, and 23. In CI-Sim, the following lists were also deemed to be equivalent: Lists 3, 7, 8, 11, 13, 15, 17, 21, and 23. Finally, six PRESTO lists were found to be consistent across both MTB and CI-Sim: Lists 7, 8, 13, 15, 17, and 23.

DISCUSSION

The most common methodology used for creating new sentence recognition tests is to first measure the intelligibility for a large number of sentences and then compile equally intelligible sentences into individual lists (e.g., Bilger et al, 1984). However, because PRESTO lists were constructed with very specific design characteristics in mind, including high levels of sentence-to-sentence indexical (e.g., talkers, regional dialect) and linguistic (e.g., syntactic structure, lexical characteristics) variability, individual test sentences were not equally intelligible. Lists were balanced for keyword frequency and familiarity because the lexical characteristics of words in sentences have been shown to affect speech intelligibility (e.g., Bell and Wilson, 2001; Dirks et al, 2001). However, not all sources of variability were accounted for in the construction of the lists. Given these design considerations, it was therefore important to establish equivalency empirically for the PRESTO lists.

List equivalency was established using measures of both list intelligibility and listener consistency. While list equivalency is typically determined by measures of the average intelligibility of each test list, based upon group mean proportion correct, we also wanted to include a measure of listener consistency by examining performance of individual listeners to ensure that each listener performed consistently across lists. This approach provides confidence that differences in scores from one condition or test session to the next cannot be attributed to differences in test lists. Although listener consistency is not explicitly discussed in the literature on speech discrimination testing, new tests of sentence recognition account for differences across listeners by modeling their sentence recognition scores using a binomial distribution (e.g., Spahr et al, 2012; Spahr et al, 2014). The number of test items influences the confidence intervals for determining critical differences (Thornton and Raffin, 1978; Raffin and Thornton, 1980). PRESTO includes 76 keywords per list, resulting in greater confidence that any differences observed in scores are not merely reflecting differences between lists. A binomial distribution table (see Appendix C) is based on 76 items that can be used as a guideline for determining significant differences across lists; however, it should be interpreted with caution because performance has not yet been validated with clinical populations. It is also possible to improve the test–retest reliability by presenting more than one list per condition or session.

PRESTO was originally designed to be a more challenging sentence recognition test to assess benefit and track performance over time for adult CI users that have scores near ceiling on conventional sentence tests, such as the HINT or even AzBio. Therefore, it was important to include a condition that approximated CI sound processing. Other speech recognition tests designed for use with CI users also used CI simulation conditions for establishing equivalency, including the adult and pediatric versions of the AzBio (Spahr et

al, 2012; Spahr et al, 2014) and the lists of TIMIT sentences constructed by Dorman et al (2005).

Another example of this approach is the Quick-SIN test (Killion et al, 2004), which was designed for use with hearing-impaired patients to identify the listeners who have difficulty understanding speech in the presence of background noise (SNR loss). Equivalency of the Quick-SIN lists was established under both noise and low-pass filtered conditions to approximate highfrequency hearing loss. Further, specific recommendations were made for administering the Quick-SIN test based on a reassessment of list equivalency with hearingimpaired listeners (Bentler, 2000).

Mixing sentences with MTB or spectrally degrading sentences with an eight-channel noise-band vocoder (CI-Sim) are two experimental manipulations that reduce the intelligibility of speech and may have differential effects on speech perception (e.g., Healy and Montgomery, 2007; Mattys et al, 2012). In this study, these two methods were employed to reduce ceiling effects in NH listeners. It was important to ensure that there were no differences across lists under these two manipulations. While the mean recognition across the MTB and CI-Sim conditions was similar (MTB: 76% and CI-Sim: 72%), the variability across and within a list was found to be consistently smaller under CI-Sim and the rank ordering of the intelligibility of the lists was different. For example, while List 3 was the easiest and List 21 the most difficult list in the MTB condition, scores on these two lists fell in the middle of the range in the CI-Sim condition. The differences observed in variability and rank ordering of lists were likely a result of the type of signal degradation used and the listeners' prior experience with both types of degraded speech signals. All of the participants have had real-world experience listening to speech in background noise, and it is possible that they developed different listening strategies based on their own unique developmental histories. In contrast, the participants in the CI-Sim condition were perceptually nïve to these conditions because they had no prior experience listening to CI simulated speech and it is unlikely they developed any consistent perceptual strategies for adapting to this type of signal degradation. An additional reason for the differences in variability may be that while all of the participants were drawn from the same undergraduate student population, participants in Phases I and II (MTB) were given course credit, while those in Phase III (CI-Sim) were paid $10 per hour for their time. Therefore, list equivalency is not invariant and may depend on the specific testing conditions, listeners, and tasks carried out (Jenkins, 1979; Roediger, 2008). Further, the rank ordering of the lists were found to be different in Phase II (noise) and Phase III (CI-Sim). Had all 19 lists been originally evaluated under CI simulation, it is possible that an entirely different group of lists may have emerged. Therefore, while we are confident that this study provides clinicians and researchers with equivalent lists based on MTB at 0 dB SNR, this may not be the only attempt at establishing equivalency with these materials in other conditions.

For clinical purposes, PRESTO provides an additional measure of sentence recognition under more challenging conditions, which assesses real-world, high-variability listening strategies. In addition, because of the design characteristics of the test, PRESTO may also be used to assess speech recognition abilities in quiet conditions in clinical populations. While testing in noise may also reduce ceiling effects, some clinical populations may be

differentially affected by the presence of background masking noise and/or competition for limited processing resources leading to results that may be difficult to interpret. Because PRESTO can be used without producing any ceiling effects in both quiet and noise conditions, this perceptually robust sentence recognition test provides the clinician with the ability to identify patients who may struggle more with speech in noise relative to their performance with speech in quiet (Neff and Dethlefs, 1995; Richards and Zeng, 2001; Wightman et al, 2010).

It is very likely that no single sentence recognition test can be universally applied to all populations and all testing conditions. Many factors affect an individual listener's ability to recognize and comprehend speech. Speech recognition performance in any condition depends on the background environment (e.g., type of noise and/or background competition, reverberation), the content of the target signal (e.g., talker(s) and talker attributes, linguistic material), and, task goals (e.g., keyword recognition, isolated word recognition, true/ false judgments) (Gilbert et al, 2013). Furthermore, in assessing variability in listener performance, numerous factors affect the performance of the individual listener, including his or her native language, linguistic background, developmental history, and cognitive skills (e.g., Tamati et al, 2013; Tamati and Pisoni, 2014). These performance factors, among others, and the interactions among these factors may contribute to the difficulty of a particular speech recognition test, reflecting an individual listener's speech recognition capabilities. Thus, to better understand and assess a patient's speech recognition abilities, it is important to consider these factors, and encourage clinicians to choose testing materials appropriate for each specific patient and/or issue from a well-stocked toolbox of tests.

Although the CI simulation was used as an approximation of listening with a CI, this condition was not intended to replace testing with CI patients. Further, the design of this study allowed us to evaluate differences in performance as a result of two experimental manipulations, such as masking noise and signal degradation, which were expected to have differential effects on perception (Mattys et al, 2012). As discussed in King et al (2012), a study evaluating the equivalency of the lists of TIMIT sentences constructed by Dorman et al (2005) in adult CI users, predicting CI patient performance based on NH young adults listening under CI simulation may have limited clinical utility and PRESTO has not yet been validated with patients with CIs. However, establishing list equivalency with CI listeners may be problematic because of the large degree of variability among patients and the many individual patient or device factors that may influence performance for any given listener. If clinics or research groups assess sentence recognition in CI patients with PRESTO lists, validating test lists may be helpful to provide initial expectations for the range of overall performance within the clinical population of interest. Further, for tests like PRESTO to be considered more representative of real-world speech understanding because of the increased variability in stimulus materials, an important next step will be to demonstrate relations between performance on PRESTO and self-assessed listening difficulties with measures of self-report, such as the widely used Speech, Spatial, and Qualities of Hearing Questionnaire (SSQ), developed by Gatehouse and Noble (2004).

SUMMARY

In this article, we report the results of an initial study designed to establish list equivalency for PRESTO test lists in two test conditions in order to provide the best recommendations for clinical and research use. Based on list intelligibility and listener consistency criteria in MTB, 10 PRESTO lists were determined to be equivalent: Lists 2, 4, 7, 8, 10, 13, 14, 15, 17, and 23. Under CI simulation, nine PRESTO lists were determined to be equivalent: Lists 3, 7, 8, 11, 13, 15, 17, 21, and 23. Combining the results across both test conditions, six PRESTO lists were determined to be equivalent: Lists 7, 8, 13, 15, 17, and 23. Although these six lists were determined to be equivalent overall, the testing conditions clearly influenced the intelligibility and consistency of the lists. Because of these effects, researchers and clinicians should take the specific presentation conditions into consideration when selecting the best PRESTO lists for use in their particular research or clinical protocols.

Acknowledgments

We thank Lauren Gowdy, Lindsay Stone, and Taylor Twiggs for their assistance on this project.

This study was supported in part by NIH NIDCD Training Grant T32DC00012 and NIH NIDCD Research Grant R01-DC00111 to Indiana University.

Appendix A

List intelligibility for Phases II and III. Results from Bonferroni corrected post hoc comparisons on all possible Phases II and III PRESTO list pairs. Resulting ***p*** values are reported. Empty cells indicate that the comparison was not significant at the 0.05 level. Phase II: MTB (upper right, no fill), Phase III: CI-Sim (bottom left, gray fill).

		Phase II (MTB) Lists												
Phase III	*p*	2	3	4	7	8	10	11	13	14	15	17	21	23
(CI-Sim)	2		<0.001							0.005			0.031	
Lists	3					<0.001	0.019	<0.001	<0.001		<0.001		<0.001	<0.001
	4	0.001	<0.0011										<0.001	
	7			0.013				0.008	0.011				<0.001	
	8			0.011									<0.001	
	10			0.010	<0.001	<0.001		0.009					<0.001	
	11	0.028		0.013			<0.001			0.001		0.002		
	13			0.012			<0.001			0.033			0.012	
	14			0.013	<0.001	<0.001		0.008	0.003				<0.001	
	15			0.016	0.003	0.031							<0.001	
	17			0.011			0.005			0.032				
	21			0.012			<0.001			<0.001				0.011
	23			0.010			<0.001							

Appendix B

Listener consistency for Phases II and III. **Results of correlational analyses between all possible Phases II and III PRESTO list pairs. Resulting r and *p* values are reported. Values from comparisons that did not reach significance at the 0.05 level are in bold. Phase II: MTB (upper right, no fill), Phase III: CI-Sim (bottom left, gray fill)

		Phase II (MTB) Lists												
Phase III	*r*, *p*	2	3	4	7	8	10	11	13	14	15	17	21	23
(CI-Sim)	2		0.555	0.668	0.646	0.676	0.733	0.553	0.735	0.676	0.683	0.549	0.553	0.539
Lists			0.003	<0.001	<0.001	<0.001	<0.001	0.003	<0.001	<0.001	<0.001	0.004	0.003	0.005
	3	0.514		0.665	0.623	0.697	0.647	0.480	0.707	0.608	0.707	0.630	0.602	0.672
		0.007		<0.001	0.001	<0.001	<0.001	0.013	<0.001	0.001	<0.001	0.001	0.001	<0.001
	4	0.534	0.543		0.713	0.753	0.678	0.469	0.789	0.767	0.699	0.794	0.612	0.723
		0.005	0.004		<0.001	<0.001	<0.001	0.016	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
	7	0.561	0.453	0.620		0.702	0.687	0.503	0.788	0.687	0.660	0.641	0.772	0.572
		0.003	0.020	0.001		<0.001	<0.001	0.009	<0.001	<0.001	<0.001	<0.001	<0.001	0.002
	8	0.429	0.524	0.581	0.681		0.569	0.603	0.772	0.655	0.742	0.549	0.648	0.709
		0.029	0.006	0.002	<0.001		0.002	0.001	<0.001	<0.001	<0.001	0.004	<0.001	<0.001
	10	0.639	0.487	0.741	0.713	0.767		0.410	0.653	0.761	0.770	0.713	0.547	0.582
		<0.001	0.012	<0.001	<0.001	<0.001		0.037	<0.001	<0.001	<0.001	<0.001	0.004	0.002
	11	0.629	0.392	0.581	0.573	0.712	0.749		0.676	**0.381**	0.576	0.533	0.496	0.541
		0.001	0.047	0.002	0.002	<0.001	<0.001		<0.001	**0.055**	0.002	0.005	0.010	0.004
	13	0.561	0.403	0.585	0.686	0.592	0.734	0.574		0.627	0.825	0.719	0.708	0.710
		0.003	0.041	0.002	<0.001	0.001	<0.001	0.002		0.001	<0.001	<0.001	<0.001	<0.001
	14	**0.360**	0.575	0.533	0.560	0.474	0.585	0.434	0.451		0.700	0.696	0.489	0.490
		0.071	0.002	0.005	0.003	0.014	0.002	0.027	0.021		<0.001	<0.001	0.011	0.011
	15	**0.369**	**0.265**	0.582	0.701	0.725	0.689	0.719	0.627	0.410		0.573	0.554	0.605
		0.064	**0.191**	0.002	<0.001	<0.001	<0.001	<0.001	0.001	0.038		0.002	0.003	0.001
	17	0.594	0.408	0.699	0.567	**0.287**	0.580	0.495	0.583	0.466	0.490		0.556	0.678
		0.001	0.038	<0.001	0.003	**0.155**	0.002	0.010	0.002	0.016	0.011		0.003	<0.001
	21	0.497	0.514	0.573	0.639	0.612	0.646	0.488	0.684	0.635	0.544	0.425		0.497
		0.010	0.007	0.002	<0.001	0.001	<0.001	0.011	<0.001	<0.001	0.004	0.030		0.010
	23	**0.351**	0.458	0.707	0.502	0.604	0.717	0.550	0.571	0.425	0.518	0.550	0.589	
		0.079	0.019	<0.001	0.009	0.001	<0.001	0.004	0.002	0.031	0.007	0.004	0.002	

Appendix C

Binomial Distribution

Binomial distribution table. Upper and lower 95% confidence intervals for the PRESTO sentence lists computed using the binomial distribution of 76 items per list (Thornton and Raffin, 1978)

	Limits	
% Score	Lower	Upper
0	0	0
5	1	13
10	3	21
15	7	28
20	9	34
25	13	39
30	17	45
35	21	50
40	26	55
45	30	61
50	36	64
55	39	70
60	43	74
65	50	79
70	55	83
75	61	87
80	66	91
85	72	93
90	79	97
95	86	99
100	100	100

Abbreviations:

ANOVA	analysis of variance
CI	cochlear implant
CI-Sim	cochlear implant simulation
HINT	Hearing in Noise Test
MSTB	minimum speech test battery
MTB	multitalker babble
NH	normal-hearing
PRESTO	Perceptually Robust English Sentence Test Open-Set

SD	standard deviation
SNR	signal-to-noise ratio
TIMIT	Texas Instruments/Massachusetts Institute of Technology

REFERENCES

Bell TS, Wilson RH. (2001) Sentence recognition materials based on frequency of word use and lexical confusability. J Am Acad Audiol 12(10):514–522. [PubMed: 11791938]

Bench J, Kowal A, Bamford J. (1979) The BKB (Bamford-KowalBench) sentence lists for partially-hearing children. Br J Audiol 13(3):108–112. [PubMed: 486816]

Bentler RA. (2000) List equivalency and test-retest reliability of the Speech in Noise test. Am J Audiol 9(2):84–100. [PubMed: 11200196]

Bilger RC, Nuetzel JM, Rabinowitz WM, Rzeczkowski C. (1984) Standardization of a test of speech perception in noise. J Speech Hear Res 27(1):32–48. [PubMed: 6717005]

Boyle PJ, Nunn TB, O'Connor AF, Moore BCJ. (2013) STARR: a speech test for evaluation of the effectiveness of auditory prostheses under realistic conditions. Ear Hear 34(2): 203–212. [PubMed: 23135616]

Bradlow AR, Pisoni DB. (1999) Recognition of spoken words by native and non-native listeners: talker-, listener-, and itemrelated factors. J Acoust Soc Am 106(4):2074–2085. [PubMed: 10530030]

Clarke CM, Garrett MF. (2004) Rapid adaptation to foreignaccented English. J Acoust Soc Am 116(6): 3647–3658. [PubMed: 15658715]

Dirks DD, Takayana S, Moshfegh A. (2001) Effects of lexical factors on word recognition among normal-hearing and hearingimpaired listeners. J Am Acad Audiol 12(5):233–244. [PubMed: 11392435]

Dorman MF, Spahr AJ, Loizou PC, Dana CJ, Schmidt JS. (2005) Acoustic simulations of combined electric and acoustic hearing (EAS). Ear Hear 26(4):371–380. [PubMed: 16079632]

Egan JP. (1948) Articulation testing methods. Laryngoscope 58(9): 955–991. [PubMed: 18887435]

Faulkner KF, Kidd GR, Humes LE, Pisoni DB. (2014) Sentence recognition in older adults. J Acoust Soc Am 136(4):2314.

Faulkner KF, Twiggs T, Pisoni DB. (2013) PRESTO: Preliminary findings with a new high-variability sentence recognition test in patients with cochlear implants. Poster session presented at the 16th Biennial Conference on Implantable Auditory Prostheses; 7 14–19; Lake Tahoe, CA.

Faulkner KF, Twiggs T, Tamati T, Pisoni DB. (2014) Speech recognition performance with PRESTO in four clinical populations. Poster session presented at the 41st Annual Meeting of the American Auditory Society; Mar 6–8; Scottsdale, AZ.

Floccia C, Goslin J, Girard F, Konopczynski G. (2006) Does a regional accent perturb speech processing? J Exp Psychol Hum Percept Perform 32(5):1276–1293. [PubMed: 17002537]

Fu Q-J, Nogaki G. (2005) Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. J Assoc Res Otolaryngol 6(1):19–27. [PubMed: 15735937]

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. (1993) The DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium.

Gatehouse S, Noble W. (2004) The speech, spatial and qualities of hearing scale (SSQ). Int J Audiol 43(2):85–99. [PubMed: 15035561]

George ELJ, Goverts ST, Festen JM, Houtgast T. (2010) Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. J SpeechLang Hear Res 53(6): 1429–1439.

Gifford RH, Shallop JK, Peterson AM. (2008) Speech recognition materials and ceiling effects: considerations for cochlear implant programs. Audiol Neurootol 13(3):193–205. [PubMed: 18212519]

Gifford RH, Dorman MF, Shallop JK, Sydlowski SA. (2010) Evidence for the expansion of adult cochlear implant candidacy. Ear Hear 31(2):186–194. [PubMed: 20071994]

Gilbert JL, Tamati TN, Pisoni DB. (2013) Development, reliability, and validity of PRESTO: a new high-variability sentence recognition test. J Am Acad Audiol 24(1):26–36. [PubMed: 23231814]

Healy EW, Montgomery AA. (2007) The consistency of sentence intelligibility across three types of signal distortion. J SpeechLang Hear Res 50(2):270–282.

Holt RF, Kirk KI, Hay-McCutcheon M. (2011) Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants. J Speech Lang Hear Res 54(2):632–657. [PubMed: 20689028]

Hudgins CV, Hawkins JE, Karlin JE, Stevens SS. (1947) The development of recorded auditory tests for measuring hearing loss for speech. Laryngoscope 57(1):57–89. [PubMed: 20287775]

Jenkins JJ. (1979) Four points to remember: a tetrahedral model of memory experiments In: Cermak LS, Craik FIM, eds. Levels of Processing in Human Memory, 429–446. Hillsdale, NJ: Erlbaum Associates.

Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP. (2013) Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci 24(10):1995–2004. [PubMed: 23985575]

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. (2004) Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 116(4):2395–2405. [PubMed: 15532670]

King SE, Firszt JB, Reeder RM, Holden LK, Strube M. (2012) Evaluation of TIMIT sentence list equivalency with adult cochlear implant recipients. J Am Acad Audiol 23(5):313–331. [PubMed: 22533975]

Kirk KI, Pisoni DB, Osberger MJ. (1995) Lexical effects on spoken word recognition by pediatric cochlear implant users. Ear Hear 16(5):470–481. [PubMed: 8654902]

Kirk KI, Pisoni DB, Miyamoto RC. (1997) Effects of stimulus variability on speech perception in listeners with hearing impairment. J Speech Lang Hear Res 40(6):1395–1405. [PubMed: 9430759]

Klatt DH. (1979) Speech perception: a model of acoustic-phonetic analysis and lexical access. J Phonetics 7:279–312.

Krull V, Choi S, Kirk KI, Prusick L, French B. (2010) Lexical effects on spoken-word recognition in children with normal hearing. Ear Hear 31(1):102–114. [PubMed: 19701087]

Levi SV, Winters SJ, Pisoni DB. (2011) Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? J Acoust Soc Am 130(6):4053–4062. [PubMed: 22225059]

Licklider JCR, Miller GA. (1951) The perception of speech In: Stevens SS, ed. Handbook of experimental psychology, pp. 1040–1074. Oxford, England: Wiley.

Luxford WM, Ad Hoc Subcommittee of the Committee on Hearing and Equilibrium of the American Academy of OtolaryngologyHead and Neck Surgery. (2001) Minimum speech test battery for postlingually deafened adult cochlear implant patients. Otolaryngol Head Neck Surg 124(2):125–126. [PubMed: 11226944]

Magnuson JS, Nusbaum HC. (2007) Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. J Exp Psychol Hum Percept Perform 33 (2):391–409. [PubMed: 17469975]

Mattys SL, Davis MH, Bradlow AR, Scott SK. (2012) Speech recognition in adverse conditions: A review. Lang Cogn Process 27(7–8):953–978.

Minimum Speech Test Battery (MSTB). Auditory Potential, LLC. Retrieved July 23, 2014, from http://www.auditorypotential.com/MSTB_Nav.html.

Moore RK. (2007a) Spoken language processing: Piecing together the puzzle. Speech Commun 49:418–435.

Moore RK. (2007b) PRESENCE: A human-inspired architecture for speech-based human-machine interaction. IEEE Trans Comput 56(9):1176–1188.

Mullennix JW, Pisoni DB, Martin CS. (1989) Some effects of talker variability on spoken word recognition. J Acoust Soc Am 85(1): 365–378. [PubMed: 2921419]

Neff DL, Dethlefs TM. (1995) Individual differences in simultaneous masking with random-frequency, multicomponent maskers. J Acoust Soc Am 98(1):125–134. [PubMed: 7608391]

Nilsson JM, McCaw VM, Soli SD. (1996) Speech test battery for adult cochlear implant users. Los Angeles, CA: House Ear Institute.

Nilsson M, Soli SD, Sullivan JA. (1994) Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am 95(2):1085–1099. [PubMed: 8132902]

Nusbaum HC, Pisoni DB, Davis CK. (1984) Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. Research on Speech Perception Progress Report No. 10 pp. 357–376. Indiana University, Bloomington: Speech Research Laboratory.

Nusbaum HC, Morin TM. (1992) Paying attention to differences among talkers In: Tohkura Y, Sagisaka Y, Vatikiotis-Bateson E, eds. Speech Perception, Speech Production, and Linguistic Structure. pp. 113–134. Tokyo: OHM.

Nusbaum HC, Magnuson JS. (1997) Talker normalization: Phonetic constancy as a cognitive process In: Johnson K, Mullennix JW, eds. Talker variability in speech processing. San Diego, CA: Academic Press, 109–132.

Nygaard LC, Sommers MS, Pisoni DB. (1994) Speech perception as a talker-contingent process. Psychol Sci 5(1):42–46. [PubMed: 21526138]

Nygaard LC, Pisoni DB. (1998) Talker-specific learning in speech perception. Percept Psychophys 60(3):355–376. [PubMed: 9599989]

Pisoni DB. (1998) Development of new perceptually robust tests (PRTs) of speech discrimination. Paper presented at Catch the Rising Star. 10th Annual Convention and Exposition of the American Academy of Audiology; Apr 2–5; Los Angeles, CA.

Raffin MJ, Thornton AR. (1980) Confidence levels for differences between speech-discrimination scores. A research note. J Speech Hear Res 23(1):5–18. [PubMed: 7442184]

Remez RE, Rubin PE, Pisoni DB, Carrell TD. (1981) Speech perception without traditional speech cues. Science 212(4497):947–949. [PubMed: 7233191]

Richards VM, Zeng T. (2001) Informational masking in profile analysis: comparing ideal and human observers. J Assoc Res Otolaryngol 2(3):189–198. [PubMed: 11669393]

Roediger HL, 3rd. (2008) Relativity of remembering: why the laws of memory vanished. Annu Rev Psychol 59:225–254. [PubMed: 18154501]

Rönnberg J, Rudner M, Lunner T, Zekveld AA. (2010) When cognition kicks in: working memory and speech understanding in noise. Noise Health 12(49):263–269. [PubMed: 20871181]

Rosenblum LD, Miller RM, Sanchez K. (2007) Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. Psychol Sci 18(5):392–396. [PubMed: 17576277]

ShannonRV ZengFG, Kamath V, Wygonski J, Ekelid M (1995)Speech recognition with primarily temporal cues. Science 270(5234):303–304. [PubMed: 7569981]

Sommers MS, Nygaard LC, Pisoni DB. (1994) Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. J Acoust Soc Am 96(3):1314–1324. [PubMed: 7962998]

Sommers MS. (1996) The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. Psychol Aging 11(2):333–341. [PubMed: 8795062]

Spahr AJ, Dorman MF. (2004) Performance of subjects fit with the advanced bionics CII and nucleus 3G cochlear implant devices. Arch Otolaryngol Head Neck Surg 130(5):624–628. [PubMed: 15148187]

Spahr AJ, Dorman MF, Litvak LM, et al. (2012) Development and validation of the AzBio sentence lists. Ear Hear 33(1):112–117. [PubMed: 21829134]

Spahr AJ, Dorman MF, Litvak LM, et al. (2014) Development and validation of the pediatric AzBio sentence lists. Ear Hear 35(4):418–422. [PubMed: 24658601]

Stickney GS, Assmann PF. (2001) Acoustic and linguistic factors in the perception of bandpass-filtered speech. J Acoust Soc Am 109(3):1157–1165. [PubMed: 11303929]

Tamati TN, Gilbert JL, Pisoni DB. (2013) Some factors underlying individual differences in speech recognition on PRESTO: a first report. J Am Acad Audiol 24(7):616–634. [PubMed: 24047949]

Tamati TN, Pisoni DB. (2014) Non-native speech recognition in adverse listening conditions. J Am Acad Audiol 25(9):869–892. [PubMed: 25405842]

Thornton AR, Raffin MJ. (1978) Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res 21(3): 507–518. [PubMed: 713519]

Warren RM, Riener KR, Bashford JA, Brubaker BS, Jr (1995) Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Percept Psychophys 57(2):175–182. [PubMed: 7885815]

Wightman FL, Kistler DJ, O'Bryan A. (2010) Individual differences and age effects in a dichotic informational masking paradigm. J Acoust Soc Am 128(1):270–279. [PubMed: 20649222]

Wilson RH, McArdle RA, Smith SL. (2007) An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. J Speech Lang Hear Res 50(4):844–856. [PubMed: 17675590]

Wong PCM, Nusbaum HC, Small SL. (2004) Neural bases of talker normalization. J Cogn Neurosci 16(7):1173–1184. [PubMed: 15453972]

Zekveld AA, Rudner M, Kramer SE, Lyzenga J, Ro¨nnberg J. (2014) Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. Front Neurosci 8:88. [PubMed: 24808818]

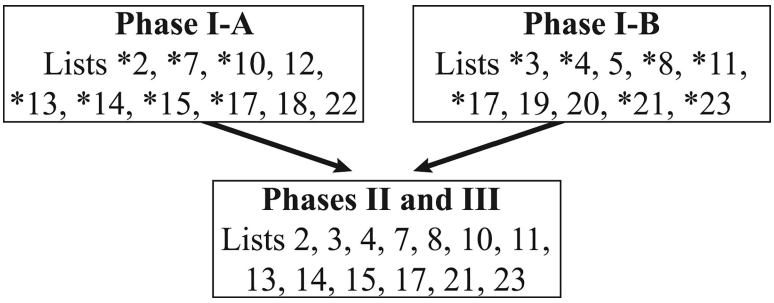

Figure 1.

Schematic of the study design. Each box contains the lists tested in each of the three phases (Phases I–III). Starred lists in Phases I-A and I-B were chosen for further evaluation in Phases II and III.

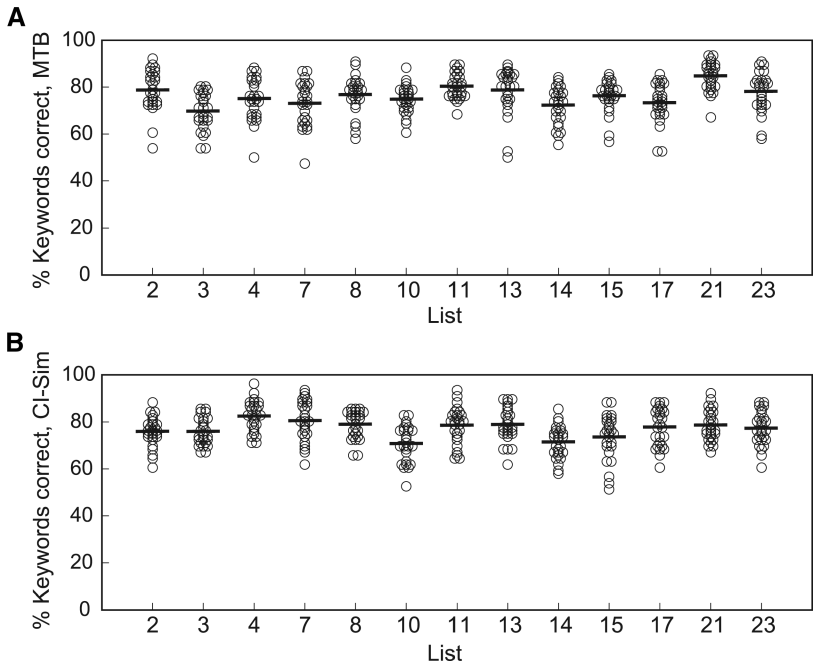

Figure 2.

Mean and individual percent keywords correct on 13 PRESTO lists in Phase II: MTB (A) and Phase III: CI-Sim (B). The mean percent correct for each list is indicated by horizontal bar. Unfilled circles represent individual performance.