Published in final edited form as:
Ear Hear. 2014 ; 35(4): 418–422. doi:10.1097/AUD.0000000000000031.

Development and validation of the Pediatric AzBio sentence lists

Anthony J. Spahr, Ph.D.,
Advanced Bionics Corporation; Arizona State University; Vanderbilt University

Michael F. Dorman, Ph.D.,
Arizona State University

Leonid M. Litvak, Ph.D.,
Advanced Bionics Corporation

Sarah Cook, B.S.,
Arizona State University

Louise M. Loiselle,
Arizona State University

Melissa D. DeJong, Au.D.,
Mayo Clinic

Andrea Hedley-Williams, Au.D.,
Vanderbilt University

Linsey S. Sunderhaus, Au.D.,
Vanderbilt University

Catherine A. Hayes, Au.D., and
Vanderbilt University

René H. Gifford, Ph.D.
Vanderbilt University

Abstract

Objectives—The goal of this study was to create and validate a new set of sentence lists that could be used to evaluate the speech perception abilities of listeners with hearing loss in cases where adult materials are inappropriate due to difficulty level or content. Our intention was to generate a large number of sentence lists with an equivalent level of difficulty for the evaluation of performance over time and across conditions.

Design—The original Pediatric-AzBio sentence corpus included 450 sentences recorded from one female talker. All sentences included in the corpus were successfully repeated by kindergarten and first grade students with normal hearing. The mean intelligibility of each sentence was estimated by processing each sentence through a cochlear implant simulation and calculating the mean percent correct score achieved by 15 normal-hearing listeners. After sorting sentences by

Correspondence Anthony J. Spahr, Advanced Bionics Corporation, 28515 Westinghouse Place, Valencia, CA 91355, Phone: (661)362-4583, Fax: (661)362-7709, tspahr@asu.edu.

mean percent correct scores, 320 sentences were assigned to 16 lists of equivalent difficulty. List equivalency was then validated by presenting all sentence lists, in a novel random order, to adults and children with hearing loss. A final-validation stage examined single-list comparisons from adult and pediatric listeners tested in research or clinical settings.

Results—The results of the simulation study allowed for the creation of 16 lists of 20 sentences. The average intelligibility of each list ranged from 78.4% to 78.7%. List equivalency was then validated, when the results of 16 adult cochlear implant users and 9 pediatric hearing aid and cochlear implant users revealed no significant differences across lists. The binomial distribution model was used to account for the inherent variability observed in the lists. This model was also used to generate 95% confidence intervals for one and two list comparisons. A retrospective analysis of 361 instances from 78 adult cochlear implant users and 48 instances from 36 pediatric cochlear implant users revealed that the 95% confidence intervals derived from the model captured 94% of all responses (385/409).

Conclusions—The cochlear implant simulation was shown to be an effective method for estimating the intelligibility of individual sentences for use in the evaluation of cochlear implant users. Further the method used for constructing equivalent sentence lists and estimating the inherent variability of the materials has also been validated. Thus, the AzBio Pediatric Sentence Lists are equivalent and appropriate for the assessment of speech understanding abilities of children with hearing loss as well as adults for whom performance on AzBio sentences is near the floor.

Introduction

Though there are many methods to evaluate the success of any rehabilitative therapy for hearing loss, speech understanding is, perhaps, the most common and best understood evaluation option available to clinicians and researchers. This evaluation process often requires a comparison of outcomes across conditions or over time. Thus, it is imperative that speech understanding materials contain multiple versions (i.e. lists) of materials that are of equal difficulty and appropriate for the population tested. With respect to sentence tests, compiling lists of equal difficulty is facilitated with an estimate of the intelligibility level of individual sentences within each list. Spahr and Dorman (2004) reported a novel application of a cochlear implant simulation to obtain an estimate of sentence intelligibility from normal-hearing listeners. This estimate of sentence intelligibility was then used to group sentences into lists of equivalent difficulty (Spahr and Dorman, 2004; Spahr et al., 2012). The resulting AzBio Sentence Test has since been adopted by many clinicians and implant programs for clinical evaluation of adult cochlear implant candidates—as outlined in the minimum speech test battery (MSTB, 2011; Fabry et al., 2009).

The AzBio materials were created to provide a realistic estimate of patient performance in real-world listening environments. To this end, each list contains samples of four different talkers producing relatively complex sentences using a conversational speaking style and rate.Gifford et al. (2008) demonstrated that the AzBio sentence lists were not as sensitive to ceiling effects as other speech materials and were highly correlated with monosyllabic word scores.

After releasing these materials for evaluation of adult listeners, several clinics opted to include them in the evaluation of younger listeners. This transition proved to be challenging for a number of reasons, not the least of which was the appropriateness of content for children (e.g. “I’ll have a cranberry and vodka, please”). Thus, several pediatric audiologists requested a set of materials that would be appropriate for use with younger listeners. A number of audiologists also registered concerns that the difficulty level of the AzBio sentence test was not appropriate for poorer-performing adult patients. To address these concerns, the process of creating a new set of materials appropriate for children and poorer-performing adults was begun.

Methods

Sentence Construction

A total of 450 sentences were created for this project. Sentences were constructed to have a length between 3 and 12 words (mean = 7.0, s.d. 1.5) and to be without proper nouns. The sentences were documented *as spoken by children* working with pediatric audiologists and by the audiologists’ own children. The children’s utterances were written down in a logbook that was kept on hand for approximately one month during which the sentences were accumulated for analysis. Given that the sentences were originally generated by children ranging in age from 5 to 12 years (both with and without hearing loss), the sentences are believed to be appropriate for gauging speech recognition abilities of children as well as being age appropriate with respect to vocabulary, grammar, and communicative abilities. It was necessary, however, to make minor corrections to the sentence utterances allowing for grammatical correctness where it may have been lacking in the original sample. Since the sentences were constructed on the basis of spoken utterances of native English speaking children, it should be noted that the corpus is heavily influenced by Western culture—specifically by those residing in North America.

Sentence Recordings

The 450 sentences were recorded by a single, female talker[1] (age = 24 years). The talker was seated in a sound booth and spoke into an AKG C2000B condenser microphone. Signals from the microphone were captured using Cool Edit 2000 software (sample frequency = 22050 Hz, resolution = 16 bit), a laptop computer, and an M-Audio Audiophile USB soundcard. Given that methods of sentence recording were identical to that completed for the AzBio Sentence Test, please refer toSpahr et al. (2012) for greater detail.

Sentence Repeatability Screening

The 450 sentence files were screened for inclusion in a pediatric sentence test by presenting them, unprocessed, to children with normal hearing (n=30). Subjects were recruited from the kindergarten and first grade classes of Phoenix Christian Elementary School. All listeners were screened at 20 dB HL at test frequencies of 0.5, 1, 2, and 4 kHz under headphones in a

[1]Though sentences were recorded with both female and male talkers, during pilot testing, it was found that the participants were more responsive to sentences spoken by the female talker. It is planned, however, to develop additional multi-talker lists such as those offered by other pediatric speech recognition metrics [e.g., multi-syllabic lexical neighborhood test (MLNT) and lexical neighborhood test (LNT), Kirk et al., 1995].

quiet room. For speech testing, sentences were randomly grouped into 9 blocks of 50 sentences. Sentences were presented at a comfortable level determined by the individual listener (60 dB SPL with minor volume adjustment, if requested) using Sennheiser HD 20 Linear II[2] headphones.

Each listener completed a practice session consisting of 20 sentences from an untested block before being tested on 1–2 blocks of 50 sentences. Each sentence was scored as the number of words repeated correctly. This process was repeated with 30 listeners in order to obtain 5 unique responses on each sentence. The total number of word errors was calculated for each sentence. Sentences with total word errors greater than 1 were deemed inappropriate for inclusion and discarded from the corpus. The remaining corpus included 388 unique sentences.

Sentence Intelligibility Estimation

The 388 repeatable sentence files and 2 non-repeatable sentence files were processed through a 15-channel noiseband vocoder and randomly assigned to 39 blocks of 10 sentences. Output noise bands had logarithmically spaced center frequencies and symmetrical roll-off of 10 dB per octave (see Litvak et al. 2007 for review). The degree of roll-off, which affects intelligibility, was selected to avoid ceiling effects when testing normal-hearing adults. A total of 15 listeners with normal hearing completed the study. Sentences were presented at a comfortable level (60 dB SPL), over headphones (Sennheiser HD 20 Linear II) to the listener, who was comfortably situated in a sound booth. Instructions were given to repeat all words that were understood and, in unsure cases, to make their best guess. Listeners were trained on the cochlear implant simulation by repeating 50 AzBio sentences (Spahr et al., 2012) processed through progressively more difficult simulations (i.e. roll-off of 40, 20, and 10 dB/octave) prior to testing.

The presentation order of sentence blocks was randomized for each listener and all sentences were scored as the number of words correctly repeated by the listener. For each sentence, a mean percent correct score was calculated (words correct / words presented). This mean percent correct score was then used as the estimate of intelligibility. Averaged across all listeners, mean sentence intelligibility ranged from 13% – 100% (mean = 72.5%, s.d = 18.7).

Sentence Selection and List Formation

It was decided that, like the AzBio Sentence Test, the pediatric test would include 20 sentences per list. As with the adult materials, all sentences were rank ordered by the mean intelligibility score. An analysis of the results revealed 320 sentences with intelligibility scores between 52% and 100% (mean = 78.6%, s.d. = 12.8). These 320 sentences, still rank ordered by mean percent correct scores, were then sequentially assigned to 16 lists. The first 16 sentences were assigned, in order, to lists 1–16 and the next 16 sentences were assigned, in reverse order, to each list (e.g. 1, 2, 3, …3, 2, 1). This sentence-to-list assignment produced 16 lists of 20 sentences with a mean intelligibility score of 78.6 percent correct (s.d. = 0.07, range 78.4% to 78.7%). The number of words per list ranged from 127 to 150

[2]Though these headphones have been discontinued, the frequency response was 20 to 20,000 Hz.

(mean = 138, s.d. = 6.4). The range of sentence intelligibility scores for the processed stimuli within lists averaged 52.9% (s.d. 1.2) to 99.2% (s.d. 0.7). Individual sentence scores and mean list scores are shown in Figure 1.

List Equivalency Validation

To evaluate equivalency of lists and characterize the inherent variability of the sentence materials, validation studies were conducted first with adult listeners and then with pediatric listeners. Given the complexity of pediatric testing, adult testing was used to initially validate equivalency and provide an opportunity to discard outlier lists prior to pediatric evaluation. A total of 16 adult, experienced cochlear implant users, were tested at Arizona State University, and 9 pediatric listeners (4 hearing aid users and 5 cochlear implant users) were tested at Desert Voices Elementary School. Pediatric listeners ranged in age from 3 to 7 years of age (mean = 4 years, 8 months) and were deemed to be appropriate for sentence testing by classroom teachers or the school speech language pathologist. All testing was completed with approval of the Institutional Review Board.

For both studies, the objective was to obtain scores from each listener on all 16 sentence lists. Adult listeners were tested in quiet or in noise (20-talker babble), based on their Consonant Nucleus Consonant (CNC, Peterson and Lehiste, 1962) monosyllabic word score. The signal-to-noise ratio used for sentence testing was determined by the individual's CNC word score. Sentences were presented in quiet, +10 dB SNR, and +5 dB SNR for individual's with word scores at or below 65% (n=13), between 66% and 85% (n=3), and between 86% and 100% (n=0), respectively. All pediatric listeners were tested in quiet. Adult testing was conducted in a double-walled sound booth. Pediatric testing was conducted in a quiet room[3]. In both cases, sentences were presented through a loudspeaker at a calibrated presentation level of 60 dB SPL at the position of the listener's head. Listeners completed a practice session (2 AzBio lists and 5 sentences from the unused pediatric corpus) prior to testing. Test lists were presented in a novel random order for each listener. To avoid fatigue, listeners were encouraged to take short breaks between lists and required to take a break after every 4th test list. All adult listeners completed testing within a single session. Pediatric listeners were evaluated over multiple sessions and not all listeners were able to complete all lists. Testing the pediatric listeners across sessions can add more variability despite randomization of list order; however, given that pediatric implant recipients will also be assessed across clinical testing sessions, it could be considered to more closely reflect real-world design.

Results and Discussion

Adult Validation Study

The mean level of performance achieved by individual adult CI listeners ranged from 37 to 90 percent correct (mean = 74%, s.d. = 14). The distribution of list scores for all 16 CI listeners is shown in Figure 2. Averaged scores for the 16 lists ranged from 69 to 78 percent

[3]It should be recognized that the use of a quiet room, instead of a sound treated booth, could potentially introduce another source of variation in performance across sentence lists. These listeners, however, all had normal hearing and were listening to materials presented at suprathreshold levels. All experimental listeners were tested in a sound treated booth.

correct (mean = 74%, s.d. = 2.4). A repeated-measures ANOVA revealed no significant main effect of list number ($F_{(15,15)} = 1.55$, $p > 0.05$).

List scores obtained from the adult CI listeners were used to identify potential outliers. To correct for individual differences in performance across all listeners, scores were normalized prior to analysis by setting all single list scores relative to the individual's mean score from all lists (list score – mean score). Normalized scores for all 16 lists are shown in Figure 3.

Pediatric Validation Study

As no significant differences were observed for the 16 sentence lists, all lists were included in the pediatric validation study. Of the 9 pediatric listeners tested, 7 completed at least 15 of 16 lists, one completed 12 lists, and one completed only 7 lists. Individual performance, averaged across all completed lists, ranged from 44% to 87% (mean = 71%, s.d. = 15). Individual scores from pediatric listeners are shown in Figure 2. The mean scores for the pediatric listeners fell within the range of scores observed for the adult listeners (37% to 90%). The standard deviation of list scores for individual listeners, a simple estimate of variability, ranged from 3.7 to 10.4 for pediatric listeners. This was below the maximum variability observed for adult CI listeners (max = 15.1).

Because not all subjects completed all lists, the data set was reduced for analysis. The two subjects completing the fewest lists were removed and two lists (2 & 12) with missing values from some subjects were also removed. For the limited data set, the group mean score for all lists was 77%, with a range of 71% to 83%. Data from the 7 pediatric listeners was then added to the existing adult data set. The 7 pediatric listeners consisted of 2 HA users (P-02 & P-09) and 5 CI users. The age of these listeners ranged from 3–7 years (mean = 5.1, s.d. = 1.7). A repeated-measures ANOVA on the complete supplemented data set revealed no significant list differences ($F_{(22,13)} = 1.749$, $p > 0.05$). Normalized list scores for all pediatric listeners are displayed in Figure 3.

Variability of Materials

Though no statistical differences were found among the 16 sentence lists on average, individual listeners all demonstrate a range of performance across lists. As with other speech materials, this variability is expected and can be modeled. The same method to characterize variability in the adult AzBio materials has been applied to the pediatric materials. That method, originally described by Thorton and Raffin (1978), was used to demonstrate how variability of performance on a monosyllabic word tests of different length could be predicted. This binomial distribution model states that variability is influenced by two factors: (i) number of independent test items, and (ii) the starting level of performance on the task. They demonstrated that variability is negatively correlated with the number of independent test items (i.e. increasing the number of test items will reduce test-retest variability) and that the greatest variability should be expected for mid-range scores (i.e. 50%). In the current study, background noise was used, when necessary, to keep performance in this mid-range. So, the model would predict that a relatively high level of variability in list scores should be observed.

Similar to the adult materials, each list contained 20 unique sentences, with an average of 138 words per list (range 127 to 150). Variability across list scores is affected by the number of independent items within each list. Calculating performance as words correct, not sentences correct, means that each scored item (word) is not likely to be completely independent due to contextual or other cues within the sentence. To determine how to best estimate the variability in list scores, mean and standard deviations were calculated from list scores of all listeners and all lists. These results were compared to the variability predicted by the binomial distribution model for lists with different numbers of independent items. This comparison revealed that the same 40-item model used to model variability in adult materials was the best fit for the pediatric materials. This result would suggest that variability in performance across these lists of sentence materials should be just slightly higher than that observed on a monosyllabic word test with 50 independent items.

List Variability

In order to reduce variability, researchers and clinicians will commonly increase the number of independent test items by administering multiple test lists within a condition. The change in expected variability when testing one or two lists per condition is shown in Table 1. For these sentence lists, the maximum variability, observed for scores of 50% correct, is reduced from 15 percentage points to 11 percentage points by adding a second test list. Though this table can be used to estimate the significance of changes across conditions or over time, caution should be exercised when scores fall too near the ceiling (> 85%) or the floor (< 15%), as quite small changes in performance will appear significant.

Clinical data collected at Arizona State University, Vanderbilt University Medical Center, and Mayo Clinic, Rochester were plotted against these confidence intervals in an effort to validate the model. These centers provided 361 instances from 78 adult cochlear implant users and 48 instances from 36 pediatric cochlear implant users where the listener had been tested on two lists in the same condition. The average age of the pediatric listeners was 10.3 years (s.d. = 3.5, range 5 to 19). The scores from the first (A) and second (B) list tested within each condition are shown in Figure 4. Given that these scores were collected under the same listening conditions, the model would predict that that the data points would fall within the confidence intervals for a single list comparison. The results revealed that 94% of scores (385/409) fell within the 95% confidence intervals. Thus, the model appears to adequately predict the variability of the sentence lists. Further, the model appears to account for both the pediatric and adult listeners with similar accuracy.

Conclusion

The goal of this project was to create a set of sentence materials appropriate for use with children and poorer-performing adults. Our approach was a slight modification to that used in creation of the AzBio Sentence Test for adult listeners. The modifications included generating materials specific to pediatric listeners, using a single, consistent voice throughout the entire list, and screening all sentences through young children with normal hearing to verify that they were repeatable. The intelligibility of the remaining sentences and

list construction was then very similar to the process used in the creation of the adult materials.

At the conclusion of this process, 16 equivalent lists of 20 sentences remained. The pediatric lists have a word count that is similar to the adult materials and the inherent variability is accurately fit by the same 40-item model. The number of lists should allow clinicians and researchers the opportunity to assess changes in performance across conditions or over time. The variability model should provide a reasonable guide for assessing the significance of performance differences within or across conditions.

Though the assessment of performance variability was conducted with both adult and pediatric cochlear implant users, the applicability of these stimuli will most generally be in the pediatric implant population. There is no reason to believe, however, that variability would differ across the adult and pediatric populations for measures of speech recognition. In fact,McCreery et al. (2010) showed no differences in the range of variability between adults and children of various ages on measures of speech recognition in noise.

Though the indication from our initial pediatric study is that the variability of pediatric CI users is very similar to that of pediatric hearing aid users, additional work should be conducted to verify that the list equivalency and the variability model hold across populations. Indeed, we expect that these materials will undergo additional validation including not only the applicability for hearing aid wearers, but also included a broader age range for the pediatric listeners.

Acknowledgments

The authors would like to thank Phoenix Christian Elementary School (Goodyear Campus) and Desert Voices Pre-School & Early Elementary Program for their assistance in screening and evaluating the sentence materials. The copyright for the sentence materials is held by the Arizona Board of Regents. The licensing rights to the materials are held by Auditory Potential, LLC, which is owned by co-authors Anthony J. Spahr and Michael F. Dorman.

References

Fabry D, Firszt JB, Gifford RH, Holden LK, Koch D. Evaluating speech perception benefit in adult cochlear implant recipients. Audiology Today. 2009; 21:37–42.

Gifford RH, Shallop JK, Peterson AM. Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. Audiology & Neuro-Otology. 2008; 13(3):193–205. [PubMed: 18212519]

Litvak LM, Spahr AJ, Saoji AA, Fridman, Fridman GY. Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. Journal of the Acoustical Society of America. 2007; 122(2):982–991. [PubMed: 17672646]

McCreery R, Ito R, Spratford M, Lewis D, Hoover B, Stelmachowicz PG. Performance-intensityfunctionsfornormal-hearingadultsandchildrenusingcomputer-aidedspeechperceptionassessment. Ear and Hearing. 2010; 31:95–101. [PubMed: 19773658]

MSTB. [Accessed February 22, 2013] The new minimum speech test battery. 2011. http://auditorypotential.com/MSTB.html.

Nilsson M, Soli S, Sullivan J. Development of the Hearing in Noise test for the measurement of speech reception thresholds in quiet and in noise. Journal of the Acoustical Society of America. 1994; 95:1085–1099. [PubMed: 8132902]

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

Spahr A, Dorman M, Litvak L, Van Wie S, Gifford R, Loiselle L, Oakes T, Cook S. Development and validation of the AzBio sentence lists. Ear and Hearing. 2012; 33(1):112–117. [PubMed: 21829134]

Thornton AR, Raffin MJ. Speech-discrimination scores modeled as a binomial variable. Journal of Speech and Hearing Research. 1978; 21(3):507–518. [PubMed: 713519]

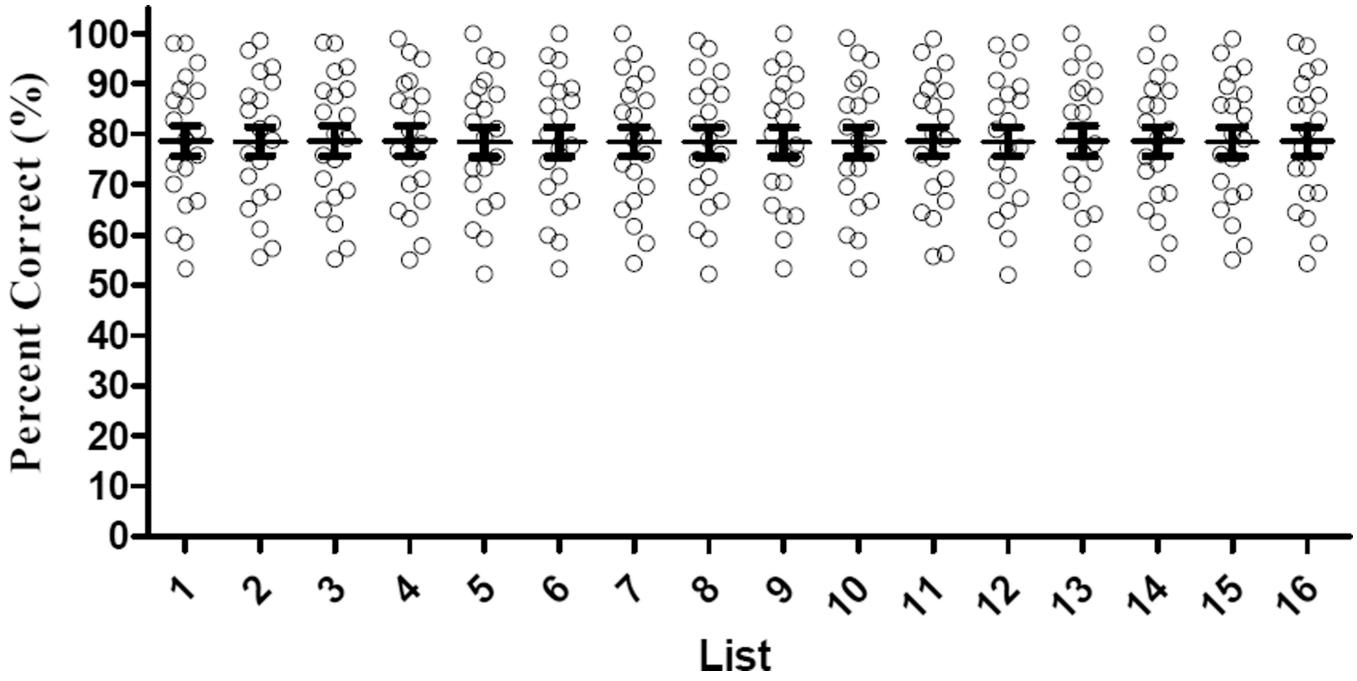

Figure 1.

Intelligibility estimates of the 16 sentence lists obtained from normal-hearing subjects listening through a vocoder. Open circles indicate the average intelligibility for individual sentences within each list. The mean percent correct score for each list is indicated by a horizontal bar with the error bars representing ±1 standard deviation.

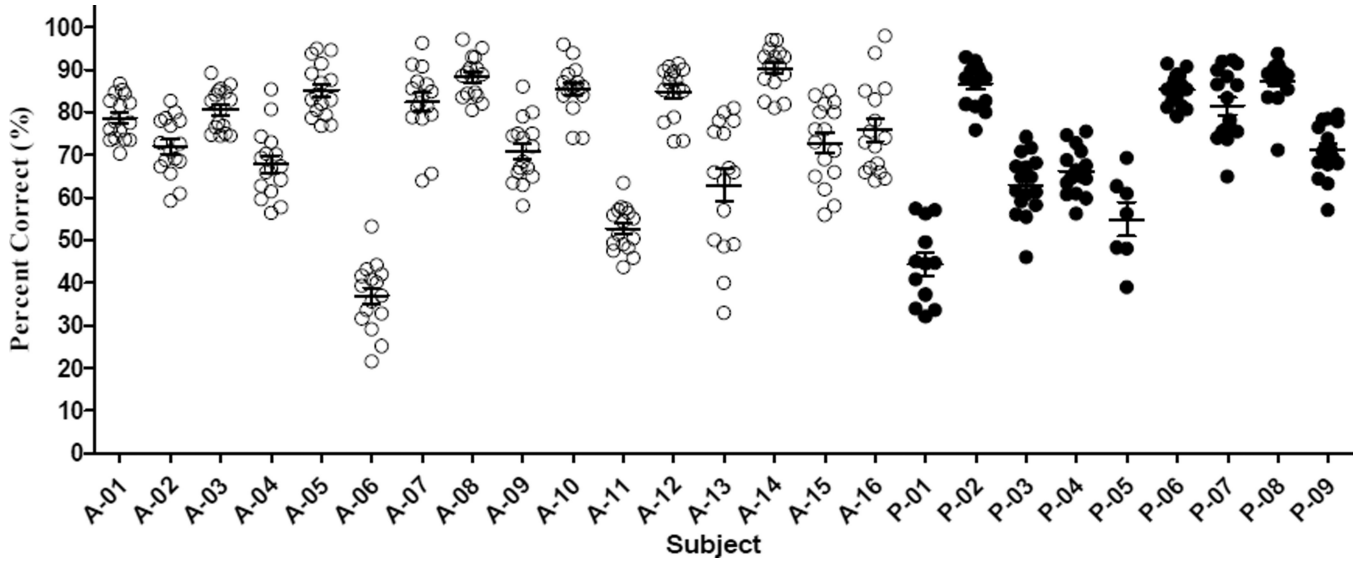

Figure 2.

List scores for 16 adult CI users (open circles) and 9 pediatric listeners (filled circles), 4 HA users (P-01, P-02, P-05, and P-09) and 5 CI users. Symbols indicate the absolute percent correct score for each of the 16 tested lists. The mean level of performance and ±1 standard deviation for each listener are indicated by horizontal lines.

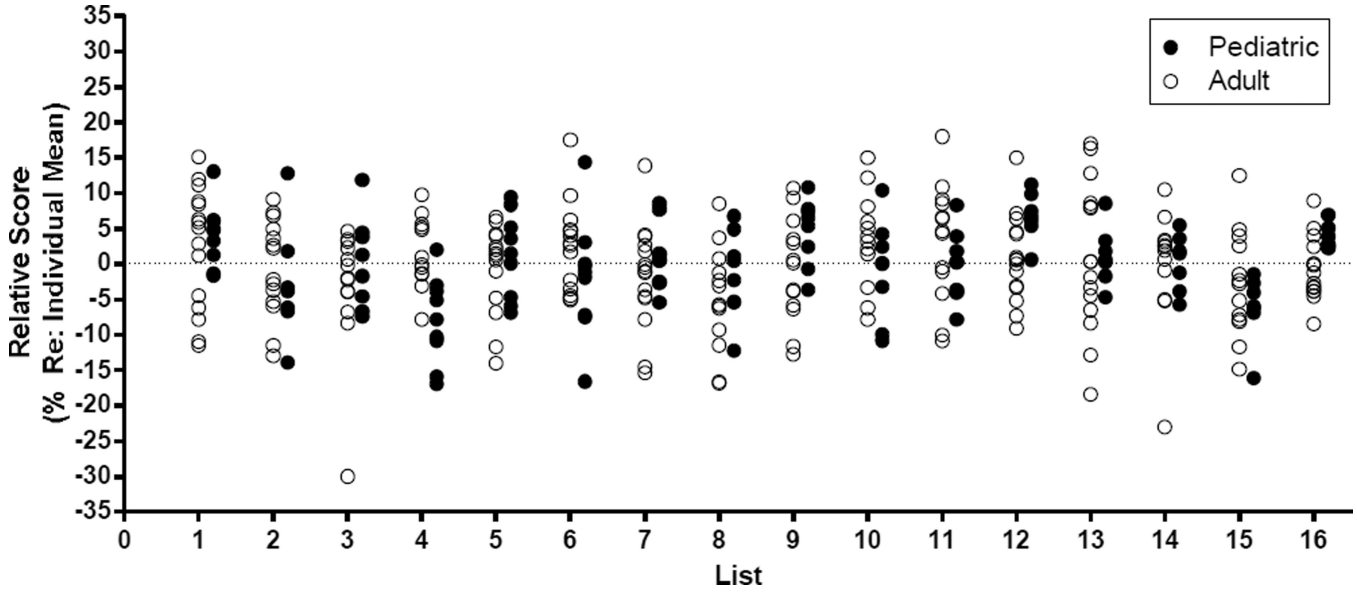

Figure 3.

Normalized scores for 16 adult CI listeners (open circle) and 9 pediatric listeners (closed circle) on all 16 sentence lists. Symbols represent an individual listener's list score relative to their overall mean level of performance. Positive values indicate better than average performance and negative values indicate below average performance. The average normalized score for each list is shown as a horizontal bar.

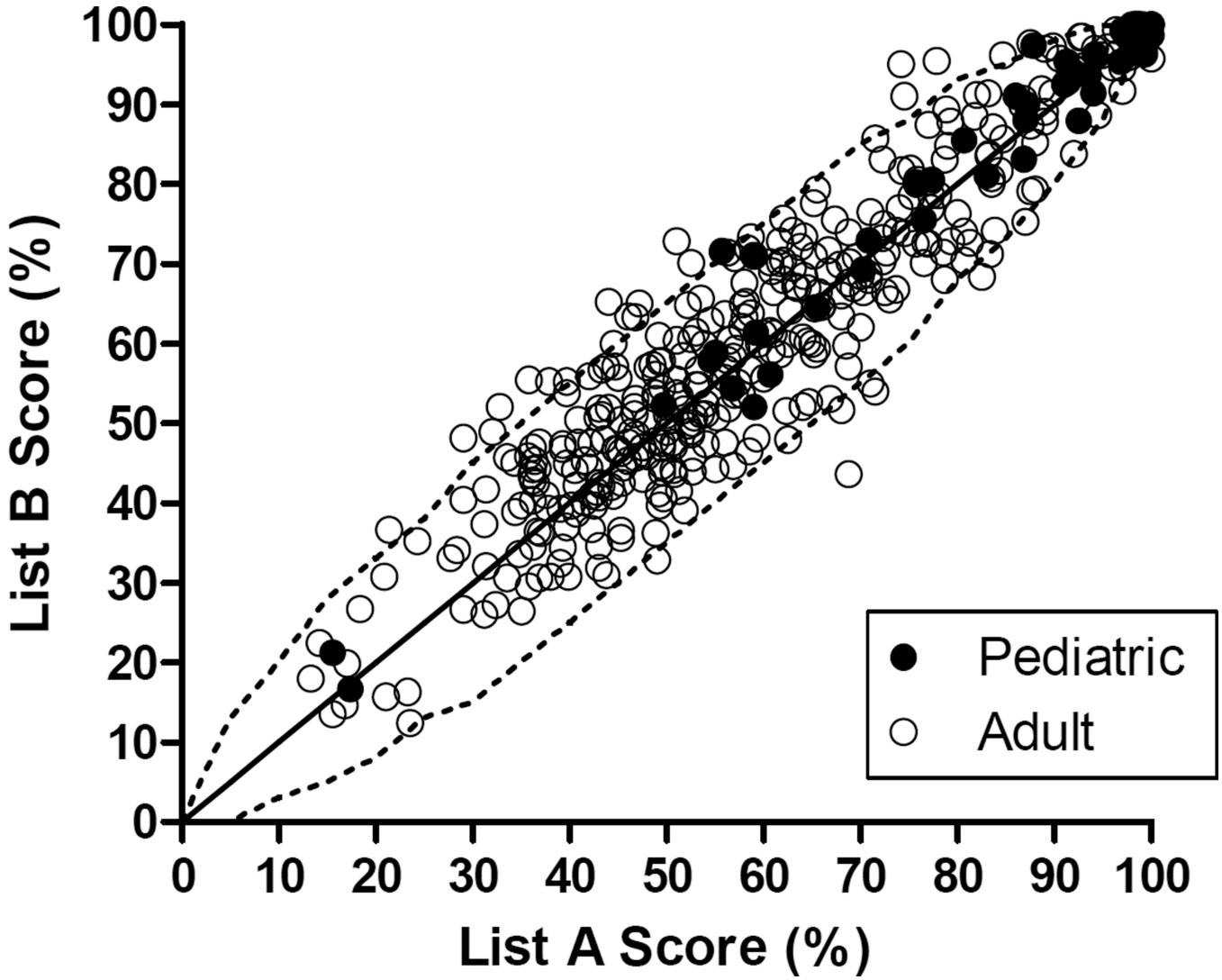

Figure 4.

Comparison of 409 instances where individual cochlear implant listeners were tested on two lists within the same listening condition. Adult scores (open circles) include 361 instances from 78 listeners and pediatric scores (closed circles) include 48 instances from 36 listeners. Solid lines represent the upper and lower 95% confidence intervals for single list comparisons. Symbols represent scores on the first (A) and second (B) list tested within the same condition. Scores falling outside of the 95% confidence interval would be incorrectly labeled as significantly different.

Table 1

Upper and lower 95% confidence intervals predicted by the binomial distribution model were constructed assuming material with 40 independent items per list. Confidence intervals for testing 1 or 2 lists per condition are shown as a function of starting level of performance (percent correct).

	1 list per condition		2 lists per condition	
Score	**Lower**	**Upper**	**Lower**	**Upper**
0	0	0	0	0
5	0	13	1	10
10	3	20	4	16
15	5	28	8	24
20	8	33	11	29
25	13	38	16	35
30	15	45	20	40
35	20	50	25	46
40	25	55	29	51
45	30	60	34	56
50	35	65	39	61
55	40	70	44	66
60	45	75	49	71
65	50	80	54	75
70	55	85	60	80
75	60	88	65	84
80	68	93	71	89
85	73	95	76	93
90	80	98	83	96
95	88	100	90	99
100	100	100	100	100