Validating the Iowa Test of Consonant Perception in a large cohort of cochlear implant users

Francis X. Smith[1,a], Joel I. Berger[2], Phillip E. Gander[2], Adam T. Schwalje[3], Timothy D. Griffiths[4], Bob McMurray[1,5,6], and Inyong Choi[1,6,7]

[1] *Department of Otolaryngology – Head and Neck Surgery, University of Iowa Hospitals and Clinics, Iowa City, Iowa 52242, USA;*

[2] *Department of Neurosurgery, University of Iowa Hospitals and Clinics, Iowa City, Iowa 52242, USA;*

[3] *Otolaryngology Clinics, Unity Point Health, Moline, Illinois 61265, USA;*

[4] *Biosciences Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom;*

[5] *Department of Psychological and Brain Sciences, University of Iowa, Iowa City, Iowa 52242, USA;*

[6] *Department of Communication Sciences and Disorders, University of Iowa, Iowa City, Iowa 52242, USA;*

[7] *Department of Intelligence and Information, Seoul National University, Seoul, South Korea*

francis-smith@uiowa.edu, joel-berger@uiowa.edu, phillip-gander@uiowa.edu, adam.schwalje@gmail.com, tim.griffiths@newcastle.ac.uk, bob-mcmurray@uiowa.edu, inyong-choi@uiowa.edu

The Iowa Test of Consonant Perception (ITCP) was designed to test word-initial phoneme perception by uniformly sampling frequently used phonemes as well as balancing feature overlap of response competitors. However, the task has only been validated in normal hearing listeners. In this study, a large cohort of cochlear implant users completed the ITCP and two commonly used clinical measures of speech recognition (AzBio sentences and CNC words). At two different signal-to-noise ratios, the ITCP showed strong convergent validity with other speech recognition tasks and good test-retest reliability. The ITCP is a useful tool for both clinicians and experimental researchers.

[a] Author to whom correspondence should be addressed.

1. Introduction

Improving speech intelligibility is a critical outcome of hearing interventions, like cochlear implants (CIs). However, the substantial variability in speech intelligibility between listeners after the intervention demands accurate and efficient methods of assessment to optimize device tuning to the listener's need, and to have an accurate understanding of how people are likely to experience the intervention. Many types of tests are used to assess different aspects of speech intelligibility in CI users. Two speech recognition assessments widely used for CI users are the AzBio sentence intelligibility test (Spahr et al., 2012) and the consonant-nucleus-consonant (CNC: Luxford, 2001) word recognition test. Both tests are open set tasks in which a participant is asked to repeat a sentence or word.

Open set tasks have gained popularity as they offer a good picture of overall speech recognition ability and may resemble the demands of real-world language processing closely. Critically, as chance performance is near zero in open set tasks, they are less susceptible to ceiling effects that may limit the diagnostic ability of the test at higher levels of performance. But open set tasks have several challenges. Successful performance on the task also requires lexical retrieval, decision making and speech production. Consequently, scores may reflect more than an individual's ability to correctly perceive speech, particularly in the context of older adults or people with variable cognitive or speech-motor abilities. Moreover, open set tasks must be scored manually and thus are generally reported only with overall accuracy scores. In contrast, while closed set tasks are not as common they can allow for automated scoring and item-level analysis (Owens & Schubert, 1977), while minimizing decision and speech production demands.

Similarly, sentence tasks like AzBio are more ecologically valid. However, as tests of hearing ability, they may be less precise, as variation in performance will be in part a product of variation in things like language skill, vocabulary, working memory and speech production. These are skills that are important for real-world function, but quite out of the influence of audiological treatments. Thus, single-word tests (particularly if they are well balanced across the sounds of the language) may offer an important complement that can isolate hearing ability.

In that vein, the Iowa Test of Consonant Perception (ITCP: Geller et al., 2020) was developed as a new task to assess single word recognition using a closed-set, four-alternative forced-choice task. This type of task removes the need for the participant to produce speech in order to respond, removing speech production errors as a confounding source of variability in performance. The ITCP is implemented as a relatively quick task that can be completed by patients autonomously, which may be useful for both researchers and clinicians. Because the ITCP broadly, and relatively evenly, samples the phonetic space it has diagnostic strength when evaluating speech recognition in specific contexts. For example, the ITCP can be used to quickly identify if a particular phoneme is difficult for a listener as well as assess what types of errors occur most frequently. This could inform device tuning at a more fine-grained level than is available with coarse accuracy measures.

The ITCP was validated using normal hearing listeners who were tested remotely using online presentation of stimuli via headphones (Geller et al., 2021). Participants performed the ITCP task, the CNC words task, and the AzBio sentences task. Despite the considerable difference in task demands, the ITCP showed good convergent validity with both CNC and AzBio scores. This suggests that the ITCP is tapping into similar speech recognition abilities between both the single word and sentence tasks. The ITCP was shown to have good test-retest reliability in normal hearing listeners as well. The reliability and validity of the ITCP as a measure of speech recognition suggests that it could be useful for clinical purposes.

However, the ITCP has not yet been evaluated for use as clinical speech recognition assessment with cochlear implant (CI) users. It is important to verify that it still has test-retest reliability within a clinical population, as well as convergent validity with other speech recognition measures. In the present study we address these questions by measuring a large cohort of CI users. We measure these participants on their performance on the ITCP, as well as their performance on the CNC and AzBio tasks as administered by their audiologists during their annual visit for device tuning. As in the original ITCP validation with normal hearing listeners, we assess convergent validity using CNC and AzBio in nearly all participants, as well as test-retest reliability among a further subset of participants who returned for an additional visit.

2. Methods

2.1 Participants

One hundred twenty experienced CI patients (greater than one year device experience), between 20 and 83 years of age (*M* = 63.8, *SD* = 12.3; 62 female), were recruited for this study. Participants were recruited through a patient registry maintained by the University of Iowa Department of Otolaryngology, and most were tested during their annual clinical visit for audiological examination and device tuning. Participants were required to be over 18 years of age, have no known neurological disorders, and not have single-sided deafness. Both pre- and post-lingually deafened participants were included in this study. A full breakdown of patients' device configuration/type, demographic, and audiological factors can be found in the Supplementary Materials. Out of these 120 participants, 44 were tested on the ITCP during an additional visit to assess test-retest reliability. These visits ranged from one month to three years apart from their initial test date (*M* = 13.2 months, *SD* = 7.7 months). Patients provided written consent to participate in the ITCP and all study procedures were approved by the University of Iowa Institutional Review Board.

This sample size was chosen on the basis of convenience (we tested all patients who came through the lab in a fixed period), not power. A post-hoc sensitivity analysis (assuming α=0.05, 1-β=0.80) found a minimum detectable effect of |r|>0.249 for the full sample of 120 and |r|>0.389 for the test/retest sample – effects far lower than we expected given our prior work.

2.2 Materials and Procedures

All participants performed the ITCP during a research session that included several other tasks. The ITCP was always the first task. In a subset of CI users, we obtained performance on two common clinical tests of speech recognition: CNC word recognition (in quiet; n = 111) and AzBio sentence recognition (in noise, +10 dB SNR from the front; n = 84). These were administered by a trained audiologist in a separate clinical session as part of a routine audiological examination. If the CNC and AzBio tests were administered on a different day than the ITCP we accepted tests performed within +/- 90 days of the ITCP testing date. Participants were excluded from analyses if we did not have complete data for all measures included in that analysis. All tasks were performed in a sound field with the individual CI users' common listening configuration to replicate their daily listening conditions. The presentation level was 70 dB SPL.

2.3 ITCP

The ITCP was implemented using MATLAB (2022a, Mathworks) scripts utilizing Psychtoolbox 3 (Brainard, 1997; Pelli, 1997). The experiment was conducted in an acoustically treated booth using a single loudspeaker (model LOFT40, JBL) positioned at 0° azimuth at a distance of 1.2 m. Visual stimuli were presented via a computer monitor located 0.5 m in front of the participant at eye level. Sound levels were the same across participants and were calibrated to present two distinct signal-to-noise ratios (SNRs) as described later.

Each trial began with a white fixation cross appearing on a black background. The fixation cross remained on the screen throughout the presentation of auditory stimuli. After 500 ms, multi-talker babble began. The babble always persisted for 2000 ms and a target word would onset after 1000 ms of the babble had played. One hundred ms after the multi-talker babble offset, four words appeared on the screen (labeled one through four) and participants were instructed to choose the word that they heard during the trial, using a keypad to select the number corresponding to one of the four options.

The ITCP consists of 120 target words organized into 30 item sets, in which a given target and its three foils differ only in their initial consonant (for full details on the development of the ITCP item sets see Geller et al., 2021). Each of these target words occurred in both a high SNR (+15 dB) and low SNR (+7.5 dB) condition with eight-talker multi-talker babble. The noise level was manipulated to create the two SNR conditions while target word presentation level remained constant. Two distinct talkers (one male, one female) were used for this test and each speaker was used for all 120 target words. A given talker's production of a target was randomly assigned to one of the two noise conditions, and this assignment was counterbalanced across participants. This led to 240 trials in total per participant with 120 trials in the high-SNR condition and 120 trials in the low-SNR condition. Conditions were randomly interleaved, and the order of trials was randomized for each participant.

3. Results

We begin by reporting descriptive statistics for accuracy on the ITCP as well as accuracy for CNC and AzBio tasks. We then assess the convergent validity of the ITCP with correlational analyses between the ITCP, CNC, and AzBio scores as well as assessing the test-retest reliability for a subset of our participants.

In the high SNR condition, participants' mean proportion correct had a range of 0.31 to 0.97 (*M* = 0.71, *SD* = 0.14). In the low SNR condition, participants' mean proportion correct had a range of 0.23 to 0.90 (*M* = 0.57, *SD* = 0.13). The mean of the high and low SNR conditions for the ITCP were significantly different, $t(119) = 25.4$, $p < .001$, confirming that the noise manipulation impacted performance on the task. In the CNC words task in quiet, participants' mean proportion correct had a range of 0.16 to 1.00 (*M* = 0.76, *SD* = 0.17). In the AzBio sentence task in noise, participants' mean proportion ranged from 0.00 to 1.00 (*M* = 0.46, *SD* = 0.28).

3.1 Convergent Validity

We assess convergent validity of the ITCP in our CI population by correlating performance on both the high and low SNR conditions of the ITCP with performance on CNC words (in quiet) and AzBio sentences (in noise, +10 dB SNR from the front). The Pearson correlation coefficient was used for all validity comparisons. We found a significant positive correlation between ITCP performance in this high SNR condition and CNC word scores, $r = 0.63$, $t(109) = 8.48$, $p < .001$ (Figure 1A). We also found a significant positive correlation between ITCP performance in the high SNR condition and AzBio sentence scores, $r = 0.73$, $t(82) = 9.65$, $p < .001$ (Figure 1B). Correlations between ITCP performance in the low SNR condition were somewhat smaller (though still large in absolute terms), but still showed significant positive correlations with both CNC word scores, $r = 0.55$, $t(109) = 7.04$, $p < .001$ (Figure 1C), and AzBio sentence scores, $r = 0.64$, $t(82) = 7.47$, $p < .001$ (Figure 1D). The correlation between CNC words scores and AzBio sentence scores was also significant, r = 0.77, t(82) = 10.85, p < .001 (figure not shown for brevity).

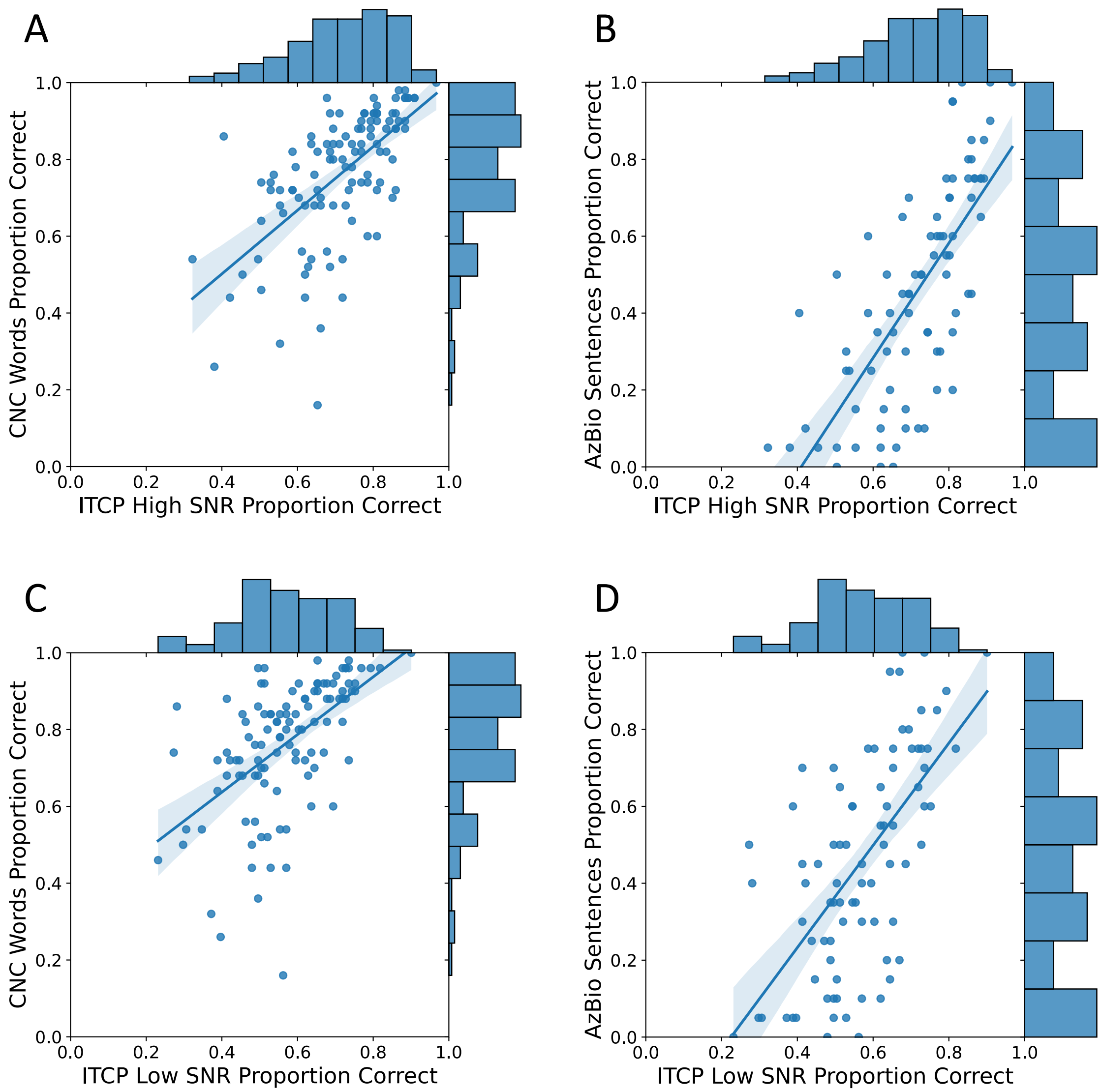

Figure 1. Scatter plots and histograms of the ITCP (high- and low-SNR conditions), CNC, and AzBio scores. A) Scatter plot of high SNR ITCP and CNC Words scores along with distribution of data (histograms). B) Scatter plot of high SNR ITCP and AzBio Sentences scores along with distribution of data (histograms). C) Scatter plot of low SNR ITCP and CNC Words scores along with distribution of data (histograms). D) Scatter plot of low SNR ITCP and AzBio Sentences scores along with distribution of data (histograms).

3.2 Test-Retest Reliability

To assess the test-retest reliability of the ITCP, we utilized a subset of CI participants who participated in the ITCP two times on separate visits (with at least one year of device experience at each visit). Among these 44 participants who were measured twice, we calculated an estimate of reliability using the intraclass correlation coefficient (Koo & Li, 2016) using the irr package in R (version 0.84.1; Gamer et al., 2019). Using this metric allowed us to assess absolute agreement between measurements at first and second visit rather than just the predictability of one score from the other. We calculated this agreement separately for the high and low SNR conditions. We used an average-score, two-way random effects model of absolute agreement and found good agreement within the high SNR condition across the two sessions, ICC(A,2) = 0.887, $F(43,44) = 8.82$, $p < .001$ (Figure 2A), and good agreement within the low SNR condition across the two sessions, ICC(A,2) = 0.843, $F(43,43.9) = 6.43$, $p < .001$ (Figure 2B). To verify that there was not a significant learning effect between the first and second session of the ITCP, we compared mean accuracy at both time points for high and low SNR conditions. For the high SNR condition, accuracy at visit one (M = .76, SD = .12) and visit two (M = .77, SD = .11) did not significantly differ, $t(43) = -0.92$, $p = .358$. For the low SNR condition, accuracy at visit one (M = .60, SD = .14) and visit two (M = .62, SD = .12) did not significantly differ, $t(43) = -1.18$, $p = .244$. To evaluate the effect of the amount of time between session one and session two, we ran a multiple regression predicting participants' second session scores from both their first session scores and the number of days between session one and session two. For both the high SNR and low SNR conditions, number of days between sessions did not reach significance (t(43) = -0.292, p = .771 and t(43) = -0.967, p = .339, respectively).

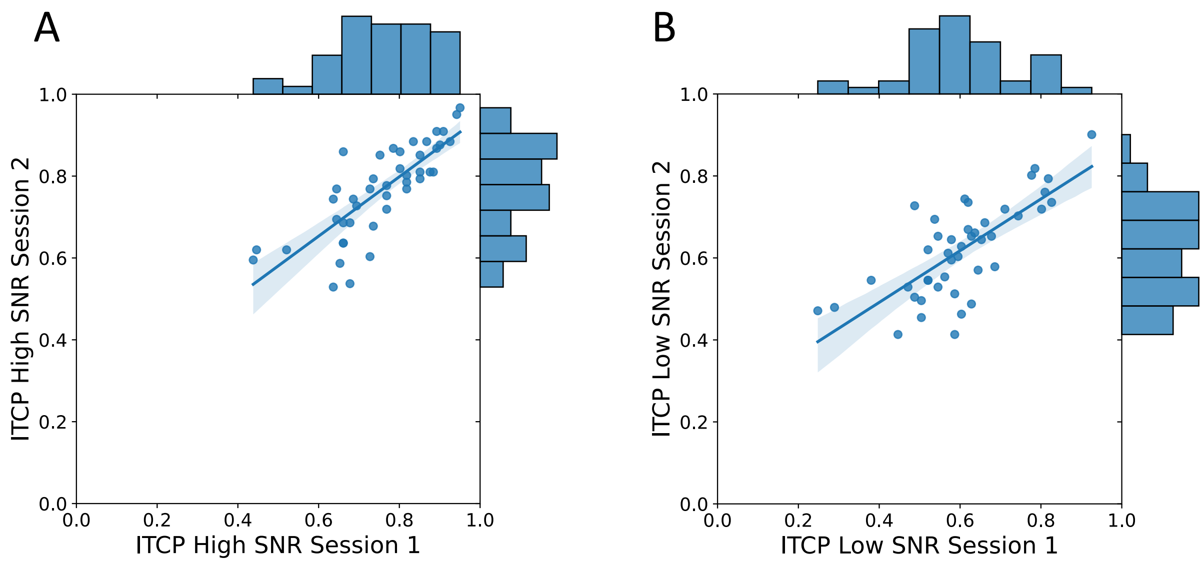

Figure 2. Scatter plots and histograms for ITCP test-retest reliability in both high and low SNR conditions. A) Scatter plot of the high SNR ITCP from session 1 and session 2 along with distribution of the data (histograms). B) Scatter plot of the low SNR ITCP from session 1 and session 2 along with distrubtion of the data (histograms).

4. Discussion

Our results show the ITCP continues to show high convergent validity with common clinical speech recognition tasks and high test-retest reliability in CI users. Our sample included both pre- and post-lingually deafened CI users with a wide variety of device configurations. The fact that we still find good convergent validity is encouraging – participants who did well in our ITCP task tended to do well in common clinical measures. Additionally, the high test-retest reliability in both high and low SNR conditions is noteworthy given the large range of difference in times between data collection sessions. The high SNR condition showed higher test-retest reliability than the low SNR condition, which might reflect the fact that some CI users may still be learning to adapt to more challenging speech-in-noise scenarios even if their performance in more ideal speech recognition scenarios has stabilized.

The ITCP has multiple advantages in assessing CI users' speech recognition abilities and can be used to track improvements over time. While closed set single word tasks are not currently favored in evaluating hearing outcomes (e.g., Yu & Schlauch, 2019; Zeitler et al., 2024), we feel that the results presented here make a strong case for the use of the ITCP both in experimental research and clinical evaluation. The relatively balanced nature of both target words and their competitor response options gives the ITCP good diagnostic strength when evaluating speech perception abilities. Combine this with the ease of administration and scoring for the task and it provides a unique opportunity to assess a listener's global speech perception abilities as well as follow up and look for specific phonemic contrasts that prove challenging for the listener. Current experimental scripts could be expanded to include summary scores for each initial consonant or various phoneme categories. This could help guide clinicians when deciding on adjustments to CI users' device programming or inform decisions about how to best help a patient train their auditory system to process their new inputs (Glennon et al., 2020). Critically, from the perspective of feasibility, the ITCP can also be run quickly. Our task that repeated every target word twice was about fifteen minutes long.

While the timing of our presentation of auditory stimuli and response options did not allow us to conduct a response time analyses, another advantage of the ITCP is that response times can be collected and analyzed. There may be patients who perform well on the task in terms of accuracy but still self-report difficulties with speech in challenging listening scenarios – within the framework of the speed-accuracy tradeoff, this may be reflected in longer response times (Heitz, 2014; Vaden et al., 2022). Future work may investigate this more directly by incorporating ITCP materials into pupillometric or dual-task measures of listening effort.

In both normal hearing listeners and cochlear implant patients the ITCP can be presented at various noise levels to avoid ceiling effects (a common concern with closed-set tasks; Sommers et al., 1997) even when only four response options are available. With the ITCP, we show strong convergent validity with two commonly accepted open set clinical tasks which suggests that, despite its closed set nature, the task is still tapping into some shared speech perception processes that underlie all these tasks. However, it does so without the heavy demands on working memory, language processing and speech production of open-set sentence tasks, demands that may be problematic when testing children, people with mild cognitive impairment, or people with other atypical profiles. Therefore, regardless of the intuitive appeal of an open set task, properly designed closed set tasks may function just as well to assess speech perception abilities.

Using the ITCP, future studies could investigate if certain device configurations or other demographic factors can be used to predict what types of errors listeners are likely to make. Indeed, recent studies of cochlear implant users have highlighted various factors that are related speech perception in noise by utilizing other tests, including duration of device use (Holder et al., 2020), duration of deafness (Kitterick & Lucas, 2016), age (Berger et al., 2023, for AzBio only), spectral and temporal resolution (Aldag & Nogueira, 2024; Choi et al., 2023), and auditory cortical responses (Aldag & Nogueira, 2024; Berger et al., 2023), though as indicated in the introduction, some of these factors may be confounded by the specific cognitive requirements of particular tests. The closed set nature of the task also lends itself to being used for EEG and/or pupillometry studies, as participants will not be preparing a motor response as they hear the word and response options can be withheld until after the onset of the auditory stimulus. This will allow for isolating auditory processes and avoiding the potential confound of preparatory motor responses (e.g. Zagha et al., 2022).

Supplementary Materials

See supplementary material at https://osf.io/9km8u for full demographic details as well as ITCP, CNC, and AzBio scores for each participant.

Acknowledgments

This work was supported by NIDCD P50 (DC000242 37) awarded to Choi, Griffiths, McMurray, and Gantz, MRC(UK) Programme grant to Griffiths (MR/T032553/1), Department of Defense Hearing Restoration and Rehabilitation Program grant awarded to Choi (W81XWH1910637 and HT9425-23-1-0912), DC008089 awarded to McMurray. This research was supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (2022H1D3A2A01092818).

Author Declarations

Conflict of Interest

The authors have no conflicts to declare.

Ethics Approval

Patients provided written consent to participate in the experiment, and all study procedures were approved by the University of Iowa Institutional Review Board.

Data Availability

The data that support the findings of this study are available within the article and its supplementary material.

References

Aldag, N., & Nogueira, W. (2024). Psychoacoustic and electroencephalographic responses to changes in amplitude modulation depth and frequency in relation to speech recognition in cochlear implantees. *Scientific Reports*, *14*(1), 1–16. https://doi.org/10.1038/s41598-024-58225-1

Berger, J. I., Gander, P. E., Kim, S., Schwalje, A. T., Woo, J., Na, Y. M., Holmes, A., Hong, J. M., Dunn, C. C., Hansen, M. R., Gantz, B. J., McMurray, B., Griffiths, T. D., & Choi, I. (2023). Neural Correlates of Individual Differences in Speech-in-Noise Performance in a Large Cohort of Cochlear Implant Users. *Ear and Hearing*, *44*(5), 1107–1120. https://doi.org/10.1097/AUD.0000000000001357

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Choi, I., Gander, P. E., Berger, J. I., Woo, J., Choy, M. H., Hong, J., Colby, S., McMurray, B., & Griffiths, T. D. (2023). Spectral Grouping of Electrically Encoded Sound Predicts Speech-in-Noise Performance in Cochlear Implantees. *JARO - Journal of the Association for Research in Otolaryngology*, *24*(6), 607–617. https://doi.org/10.1007/s10162-023-00918-x

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (0.84.1). https://cran.r-project.org/web/packages/irr/index.html

Geller, J., Holmes, A., Schwalje, A., Berger, J., Gander, P., Choi, I., & McMurray, B. (2021). Validating the Iowa Test of Consonant Perception. *The Journal of the Acoustical Society of America*, *150*(3), 2131–2153.

Geller, J., McMurray, B., Holmes, A., & Choi, I. (2020). *ITCP Iowa Test of Consonant Perception*. https://osf.io/hycdu/

Glennon, E., Svirsky, M. A., & Froemke, R. C. (2020). Auditory cortical plasticity in cochlear implant users. *Current Opinion in Neurobiology*, *60*, 108–114.

https://doi.org/10.1016/j.conb.2019.11.003

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*(8 JUN), 1–19. https://doi.org/10.3389/fnins.2014.00150

Holder, J. T., Dwyer, N. C., & Gifford, R. H. (2020). Duration of Processor Use per Day Is Significantly Correlated with Speech Recognition Abilities in Adults with Cochlear Implants. *Otology and Neurotology*, *41*(2), e227–e231. https://doi.org/10.1097/MAO.0000000000002477

Kitterick, P. T., & Lucas, L. (2016). Predicting speech perception outcomes following cochlear implantation in adults with unilateral deafness or highly asymmetric hearing loss. *Cochlear Implants International*, *17*, 51–54. https://doi.org/10.1080/14670100.2016.1155806

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Luxford, W. M. (2001). Minimum speech test battery for postlingually deafened adult cochlear implant patients. *Otolaryngology - Head and Neck Surgery*, *124*(2), 125–126. https://doi.org/10.1067/mhn.2001.113035

Mathworks. (2020). *MATLAB (R2020a)* (9.8.0.1323502). The MathWorks Inc.

Owens, E., & Schubert, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research*, *20*(3), 463–474. https://doi.org/10.1044/jshr.2003.463

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing*, *18*(2), 89–99.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and Validation of the AzBio Sentence Lists. *Ear & Hearing*, *33*(1), 112–117. https://doi.org/10.1097/AUD.0b013e31822c2549

Vaden, K. I., Teubner-Rhodes, S., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2022). Evidence for cortical adjustments to perceptual decision criteria during word recognition in noise. *NeuroImage*, *253*(March), 119042. https://doi.org/10.1016/j.neuroimage.2022.119042

Yu, T. L. J., & Schlauch, R. S. (2019). Diagnostic Precision of Open-Set Versus Closed-Set Word Recognition Testing. *Journal of Speech, Language, and Hearing Research*, *62*(6), 2035–2047. https://doi.org/10.1044/2019_JSLHR-H-18-0317

Zagha, E., Erlich, J. C., Lee, S., Lur, G., O'Connor, D. H., Steinmetz, N. A., Stringer, C., & Yang, H. (2022). The Importance of Accounting for Movement When Relating Neuronal Activity to Sensory and Cognitive Processes. *Journal of Neuroscience*, *42*(8), 1375–1382. https://doi.org/10.1523/JNEUROSCI.1919-21.2021

Zeitler, D. M., Prentiss, S. M., Sydlowski, S. A., & Dunn, C. C. (2024). American Cochlear Implant Alliance Task Force: Recommendations for Determining Cochlear Implant Candidacy in Adults. *Laryngoscope*, *134*(S3), S1–S14. https://doi.org/10.1002/lary.30879