Variability of word discrimination scores in clinical practice and consequences on their sensitivity to hearing loss

Annie Moulin, André Bernard, Laurent Tordella, Judith Vergne, Annie Gisbert, Christian Martin, Céline Richard

Variability of Word Discrimination Scores in Clinical Practice and Consequences on their Sensitivity to Hearing Loss.

Annie MOULIN[1,2,6*], André BERNARD[3], Laurent TORDELLA[4], Judith VERGNE[1,2], Annie GISBERT[3], Christian MARTIN[3], Céline RICHARD[5].

1. Inserm U1028, CNRS UMR 5292, Lyon Neuroscience Research Center, Brain Dynamics and Cognition Team, University of Lyon, 69000 Lyon, France
2. Lyon Neuroscience Research Center, Brain Dynamics and Cognition Team, University of Lyon, 69000 Lyon, France
3. Otorhinolaryngology Department, Hôpital Nord, CHU of St Etienne, 42055, St Etienne Cedex, France
4. CHU of St Etienne, 42055 St Etienne Cedex, France
5. Otorhinolaryngology Department, University Hospital of Lausanne, Lausanne, Switzerland
6. DYCOG, Dynamique Cérébrale et Cognition, Centre de Recherche en Neurosciences de Lyon, CRNL, Inserm U1028, CNRS UMR5292, CH Le Vinatier, Bâtiment 452, 95 Bd Pinel, 69675 Bron Cedex, France

***Corresponding Author** : Annie Moulin

annie.moulin@cnrs.fr

This is the final version of the manuscript accepted for publication in the *European Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*.

The publisher's version is available at:
https://link.springer.com/article/10.1007/s00405-016-4439-x

Reference:

Moulin, A., Bernard, A., Tordella, L., Vergne, J., Gisbert, A., Martin, C., & Richard, C. (2017). Variability of word discrimination scores in clinical practice and consequences on their sensitivity to hearing loss. *European Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery, 274*(5), 2117-2124. https://doi.org/10.1007/s00405-016-4439-x

Variability of word discrimination scores in clinical practice and consequences on their sensitivity to hearing loss.

Annie MOULIN, André BERNARD, Laurent TORDELLA, Judith VERGNE,

Annie GISBERT, Christian MARTIN, Céline RICHARD.

Abstract

Speech perception scores are widely used to assess patient's functional hearing, yet most linguistic material used in these audiometric tests dates to before the availability of large computerized linguistic databases. In an ENT clinic population of 120 patients with median hearing loss of 43 dB HL, we quantified the variability and the sensitivity of speech perception scores to hearing loss, measured using disyllabic word lists, as a function of both the number of ten-word lists and type of scoring used (word, syllables or phonemes).

The mean word recognition scores varied significantly across lists from 54 to 68%. The median of the variability of the word recognition score ranged from 30% for 1 ten-word list down to 20% for 3 ten-word lists. Syllabic and phonemic scores showed much less variability with standard deviations decreasing by 1.15 with the use of syllabic scores and by 1.45 with phonemic scores.

The sensitivity of each list to hearing loss and distortions varied significantly. There was an increase in the minimum effect size that could be seen for syllabic scores compared to word scores, with no significant further improvement with phonemic scores. The use of at least two ten-word lists, quoted in syllables rather than in whole words, contributed to a large decrease in variability and an increase in sensitivity to hearing loss. However, those results emphasize the need of using updated linguistic material for clinical speech score assessments.

Key Words: Speech recognition scores, Disyllabic word lists, effect size, hearing-loss.

Introduction

Speech audiometry is one of the most practiced tests in audiology; its principle stems from initial research in the Bell laboratories in the 1930s, focused on the intelligibility of speech transmitted through phone lines[1]. Indeed, by assessing the ability of a patient to understand a spoken message, it remains part of the gold standard tests to evaluate a patient's functional hearing as well as the benefits of hearing rehabilitation[2, 3]. Indeed, word recognition scores allows the exploration of distortions in the auditory information, distortions that cannot be explored by pure tone audiometry alone. Those distortions are due to several neurophysiological deficits that accompany hearing loss (such as loss of frequency selectivity, jitter in precise timing, reduction of the dynamic range and loss of binaural spectral cues) and play a major part in communication difficulties experienced by hearing impaired patients.

Speech audiometry relies on two routine measures: the speech reception threshold (SRT) and the speech discrimination score. The SRT is classically defined as the decibel level required for 50% disyllabic word recognition, whereas the speech discrimination score is usually performed at a stimulus level intensity 35 dB above the SRT using monosyllabic words[4]. Disyllabic words have been initially chosen for the SRT because of the greater homogeneity and the steepness of their psychometric functions[5], due to their sensitivity to contextual influences[6]. The spondaic (disyllabic) word lists of Fournier[7], designed in the fifties, and the monosyllabic lists of Lafon[8] are the most common lists used in the present days for speech audiometry in French clinical practice[9]. Like most word lists in different countries, they were designed well before the availability of large computerized lexical databases[10, 11] and therefore could not take into account several influential factors brought to light by psycholinguistic research such as spoken word occurrence frequency[12], phonological similarities[13] and various indices of contextual influences—factors that can contribute to substantial variability in speech recognition scores[14, 15].

The development of new hearing aid devices brought a greater demand for means of measuring word recognition in a reproducible way, in multiple conditions (monaural, binaural, with/without a hearing device, in quiet and noise, etc.) and in a time frame compatible with the patient's fatigue. This requires a large corpus of word lists with a good agreement between the scores obtained from each list[16]. The standard of fifty words per list is rarely used due to time constraints[17] and "half lists" of 25 words are preferred, hence increasing the score variability. Several authors have recommended the use of phonemes scores for monosyllabic word lists, rather than whole word scores, to increase the number of items to be quoted and hence the precision, without increasing the number of words used[17, 18]. Fournier's disyllabic word lists are the recommended lists for speech audiometry[9] and cochlear implant evaluation and follow-up[19] for which they are widely used[20–22].

The present work is therefore aimed at quantifying the degree of variability in word recognition scores obtained in a daily clinical population in an ENT department, using disyllabic word lists. The ability of such material to detect differences in speech perception due to distortions was evaluated as a function of the number of words (from 10 to 60) and the type of scoring method used (phonemes, syllables and words).

Material & Methods

Patients

Audiological data from 120 patients (58 women, 62 men) aged from 23 to 85 years seen at our ENT clinic for an audiological assessment were retrospectively and anonymously selected from the clinical database. All patients were native French speakers and had undergone otoscopic, tympanometric and audiometric evaluation. Most patients had come for presbyacusis. Degrees of hearing loss were mild (21-40 dB HL, n = 50), moderate (41-70 dB HL, n = 60) and severe (70-90 dB HL, n = 10) according to the International Bureau for Audiophonology[23].

Audiology testing:

Audiometric examinations were performed in sound-proof booths. Pure-tone baseline audiograms were obtained for 250, 500, 1000, 2000, 4000 and 8000 Hz using an Interacoustic AC33 clinical audiometer. Speech Audiometry testing was conducted using words taken from current practice French word lists: triphonemic monosyllabic words taken from Lafon lists[8] and spondaic disyllabic words taken from Fournier lists[7]. Subjects were asked to repeat the word heard after each word presentation, and each patient's response was written and quoted as correct or not by an experienced audiologist. Speech scores were recorded as a function of the stimulation intensity using the Lafon lists of monosyllabic words. We then selected the ears that had speech scores recorded using the same six Fournier[7] lists of ten disyllabic words presented at the same stimulus intensity. In addition, that intensity had to correspond to a monosyllabic word score between 40 and 80% (to avoid ceiling and floor effects). This selection process yielded a population of 120 ears from 120 patients: Only one ear per patient was chosen at random and considered in the statistical analysis to keep the scores with the same degree of independence.

Data and statistical analysis

The scores were averaged for each one of the 6 ten-word lists, giving 6 scores per patient. As those ten word lists are generally used in a combination of two or more lists, all the different mathematical combinations of two, three, four and five lists were tested (i.e., 15 different combinations for two and four lists, 20 different combinations of three lists, and 6 combinations of five lists). Hence, we obtained a total of five series of combinations per patient corresponding to a total of 6 + 15 + 20 + 15 + 6 combinations. Variability in speech scores was calculated, for each patient, by the standard deviation and the range of scores obtained across each one of the 5 series of combinations as shown for two patients (Fig. 1).

To compare the performances of the different list combinations in showing significant differences in word scores, the population was split into two groups by the median average PTA, with one group with PTA $\leq$ 43.5 dB HL (the low hearing loss group), and one group with PTA > 43.5 dB HL (the high hearing loss group). Word scores were compared between these two groups and the corresponding effect size (Cohen's *d*[24]) was calculated for each

combination, using a method similar to Schlauch et al.[18]. Cohen's *d* is an objective and standardized measure of the magnitude of an observed effect; here, the observed effect is a difference in word scores between two populations. Cohen's *d* can therefore be used for comparison between different sets of results: here, the different sets of results correspond to the effect seen by different word lists combinations and type of scoring methods (whole word scores, syllabic scores and phonemic scores). As the stimulus intensity chosen was an intensity corresponding to a monosyllabic word score between 40% and 80%, it increased significantly as hearing loss increased ($r = 0.88$, $p < 0.0001$). This compensation of audibility is shown in the difference of the average stimulus intensities used between the two groups: 65 dB HL (SD = 14.6) for the high hearing loss group and 41.6 dB (SD = 9) for the low hearing-loss group. This corresponded to a monosyllabic word score of 54.8% and 58% respectively (Student's $t = 2.2$, $p < 0.04$).

The data were Gaussian, as ascertained by the Shapiro-Wilk test. Analyses of variance (Anova) for repeated measurements (Anova-R) were performed to compare the different word lists scores across the same patients. Correlation matrices were used to correlate the word scores obtained by the different lists and list combinations using Pearson correlation coefficients. Statistical differences in correlations were measured using z-scores. Statistical analyses were performed using R° software version 3.2.2 (2015) with a level of statistical significance (*p*) less than or equal to 0.05.

Results

1. Differences and correlations between lists

Mean word recognition scores for the six disyllabic word lists ranged from 54.5% to 68% (maximum mean difference of 13.5%) with highly significant differences between them whether the scores were in words, syllables or phonemes (Fig. 2). All pairwise comparisons were statistically significant except L1L2, L1L5, L2L4 and L2L5. Two-way Anova-R showed a main effect of list ($F(5, 595) = 19.4$, $p < 0.0001$) and of type of score [$F(2, 238 = 690$, $p < 0.0001$)] with a significant interaction [$F(10, 1190) = 6.9$, $p < 0.001$]. The maximum mean differences between lists were 13.5% for words, 12% for syllables and 10% for phoneme scores. The phoneme scores were, on average, 15.7% greater than word scores. When two list combinations were used, the word scores ranged from 57.4% to 67.1% (i.e., less than 10% difference), and from 58.6 to 66.7% for three list combinations.

Correlations between monosyllabic word scores and each of the six disyllabic lists scores were highly significant (Table 1, $p < 0.0005$) and ranged between 0.33 and 0.50 (i.e., a percentage of explained variance between 11% and 25%) with significantly greater correlations with some lists than others ($z = 2.2$, $p < 0.03$). Correlations between the disyllabic lists were all highly statistically significant with r values ranging from 0.43 to 0.61, i.e. a percentage of explained variance ranging between 18% and 37% (Table 1). When combinations of two lists were considered (i.e., 20 words), correlations ranged from 0.63 to 0.74, and from $r = 0.74$ and $r = 0.81$ for combinations of three lists. No significant differences in correlations were obtained between the scores in words and the scores in phonemes and in syllables.

2. Score variability

The range of the word recognition scores throughout the six lists varied greatly across patients, from 0% to more than 50%. Box and whiskers plots of the range were calculated for each list combination (from one to five lists) and for each patient (Figure 3). These data showed that for a ten word score (i.e., one list), the median is 30% (mean = 33.5%, SD = 11.6), i.e., half of patients have more than 30% of score variability. For combinations of two lists, the median is 25% (17% for phonemic scores). As the number of lists considered increased (from one to five, i.e., 50 words), the range decreased down to 6% for five lists (50 words). However, the range for four and five list combinations are underestimated as those combinations necessarily include list repetitions between the different combinations.

The differences between word scores, syllable scores and phoneme scores variability were all statistically significant for all series of combinations. The variability of syllable scores was systematically lower than word scores, and phoneme scores showed the least variability (ANOVA-R $F(2,199) = 94$, $p < 0.0001$ for single list combinations) (Fig. 3). There was no significant correlation between variability indexes across lists (such as the standard deviation and the range), patients' characteristics (such as age), or hearing loss.

3. Variability in sensitivity to distortions.

As the stimulation intensity was adjusted individually for each ear tested, based on the monosyllabic word score, in order to avoid floor or ceiling effects, a strong correlation (r = 0.88, $p < 0.0001$) was observed between stimulus intensity and hearing loss (measured as PTA). Due to this adjustment, the speech recognition scores measure, here, more distortions than hearing loss per se. This is because loss of audibility has been largely compensated. Still, correlations between word scores and hearing loss were obtained with significant differences depending on the list considered: the strongest correlation was obtained with L3 (r = -0.32, $p <$ 0.001, n = 120), whereas no statistical significant correlations were obtained for L2 (r = -0.10, p = ns, n = 120) (z = 2.6, $p < 0.01$) (Table 1).

Using a high hearing-loss and a low hearing-loss patient groups, each list combination and each scoring method were compared in their ability to show significant differences in word scores between the two groups. Although audibility has been compensated with an average stimulus level for high hearing-loss group being more than 20 dB above the one used for low hearing-loss group, the high hearing-loss group still showed lower word scores than the low-hearing loss group, and the effect size varied greatly between the different list combinations and type of scoring method. Some lists showed a highly significant difference between both groups (t = 3, $p < 0.004$, d = 0.26 for L5) whereas others failed to show any significant difference. ANOVA-R showed a significantly greater effect size (Cohen's d) for syllabic and phonemic scores than for the word scores, without any significant difference between syllabic and phonemes scores: $F(2,14) = 6.2$, $p < 0.006$ for two list combinations, $F(2,19) = 18.4$, $p <$ 0.0001 for three list combinations and $F(2,14) = 28$, $p < 0.0001$ for four list combinations. The mean Cohen's d ranged from 0.19 to 0.55 for one list, from 0.31 to 0.57 for two list combinations and from 0.35 to 0.55 for three list combinations. The mean and median effect size increased from one list to five list combinations, and the greatest increase is observed in the minimum value of d obtained within each series of combinations (Fig. 4).

Discussion

The goal of this study was to determine speech scores' variability using disyllabic word lists and the potential improvement in effect size with syllable or phoneme scores. When developing its spondaic words lists in 1951[7], Fournier emphasized the importance of the word composition of each list, in which each word list has to be as easy/difficult as the others, based on an even spread of word scores across lists. However, the present report highlights several limitations of those spondaic word lists for an accurate and reproducible evaluation of patient's speech recognition. The mean population word recognition scores varies significantly from one ten-word list to another (mean recognition scores varying from 54.5% to 68%). This average difference of more than 13.5% for one ten-word list drops down to less than 10% for two ten-word lists. However, those differences in scores between lists are not stable across patients: Variability in scores for a single patient at a single intensity, can vary from 0% to 50% depending on the list, whereas other patients can show a variation of less than 10%.

This variability observed across lists is only slightly lower when increasing the number of lists tested at the same intensity: half of patients exhibited more than 20% of score variability on three lists (i.e. 30 words). This variability of scores depending on the list used is reflected in the correlations obtained between lists: relatively low correlations, although highly statistically significant, were obtained between the different ten-word lists. The percentage of shared variance between two lists ranged between 18 to 37% and between 40% to 55% for 2 ten-word lists combinations, reaching a maximum of 65% of shared variance for 3 ten-word lists. The use of syllable or phoneme scores did not significantly improve correlations between word lists.

As expected, because the phoneme and syllable scores take into account partial responses from patients, they are significantly greater than word scores, and they show less variability from one combination of word lists to another. Indeed, speech scores can be modeled as a binomial variable, and thus the doubling of the number of items results in a decrease of standard variation by 1.4[17, 25]. A decrease in standard deviation by 1.15 was observed between word scores and syllable scores and a decrease by 1.45 from word scores to phoneme scores: this decrease is much lower than would be expected by the number of items because each word contains at least four phonemes. This is easily explained by the fact that the phonemes of a word are not statistically independent and disyllabic words have been suggested to be perceived more as two chunks of information corresponding to two syllables rather than as a string of phonemes[15].

Syllable scores and phonemic scores showed less variability than word scores. To see if this lower variability leads to greater statistical power and better clinical significance, the difference in word, syllable and phoneme scores were tested between two groups of patients differing by hearing loss. Although the word presentation levels were adjusted for each patient, a decrease in word scores with increasing hearing loss was still obtained because adjusting levels is not sufficient to restore the same level of perception for all patients due to distortions, as each patient has a different degree and origin of hearing impairment. The different word lists showed variability in their sensitivity to hearing damage: some word lists showed indeed

decreasing scores with increasing hearing losses, whereas other failed to show any difference for the same population. This is confirmed by variability in the effect sizes, depending on the combinations of lists used. The effect size ranged from "mild" (0.18) to "medium" (0.68). The increase in the number of lists used did not lead to a strong increase of the maximum effect size, but did lead to an important increase in the minimum effect size (from 0.19 for one list to 0.31 for two ten-word lists). Increasing the number of lists beyond two did not markedly increase the effect size. However, the syllabic scores showed systematically greater effects size than word scores, but the use of phoneme scores did not result in a further increase in effect size (fig. 4).

The use of a combination of at least two lists (20 words) and a syllable count can help reduce the variability and can offer a medium effect size without increasing the duration of the test too much with disyllabic words. The use of monosyllabic word lists[26–28] is only a partial solution because monosyllabic words are still subject to contextual influences, albeit in a weaker way than disyllabic words.

Indeed, speech audiometry involves not only the audibility of the word presented, but also the entire cognitive process of reconstructing a meaningful word from partial or deteriorated acoustic input. Word occurrence frequencies, word familiarity, age of acquisition and phonological neighborhood all play an important role in this process[29]. For instance, it has been largely demonstrated that word perception scores increase for words with high occurrence frequency[6, 30] and decrease for rare words. In his work in the 1950s, Fournier himself regretted the absence of large lexical databases in French language for occurrence frequency[7]. More recently, the superiority of spoken occurrence frequency[31] versus written occurrence frequency in predicting word recognition has been shown in several languages[12, 32]. The influence of occurrence frequency on word recognition acts in conjunction with phonological neighborhood (defined as the number of different words that differ from the target word by one phoneme). Words with a high number of phonological neighbors—especially if those phonological neighbors are of high occurrence frequency—are more difficult to perceive than words with a few phonological neighbors[13, 33]. In their psycholinguistic computational evaluation of the PBK word lists largely used in the English language, Meyer and Pisoni[14] emphasized the need to consider word occurrence frequencies and phonological neighborhoods in word lists construction. This would lead to less inter-subject variability due to undesired factors such as educational level[15] and better score reproducibility[16].

Several European countries use linguistic material developed more than 50 years ago (e.g. [9, 34, 35]) when psycholinguistic databases were not developed. More recent updates of speech material are found in Russian[36, 37], Mandarin[38, 39] and Cantonese[40]. The recent availability of large multilingual linguistic databases including spoken occurrence frequencies[10, 12, 41, 42] and phonological information provides the basis for an update of word lists to obtain a better psycholinguistic equivalency between word lists.

Hearing rehabilitation strategies constantly evolve, stressing the need for reliable and accurate means to evaluate speech perception and rehabilitation benefit[43, 44]. The data presented here emphasizes the need for updating speech recognition word lists currently used

in most countries. This work can be eased thanks to new technologies such as the availability of large and regularly updated lexical databases. The recent multilingual matrix tests, that use a closed set paradigm, can lower contextual influence and is definitely a valuable approach[45, 46].

Conflict of interest statement

All authors of the present publication disclose any financial or personal relationships that may be considered as a potential conflict of interest.

Acknowledgements

This work was supported in part by the ‘‘Fondation de l’Avenir/VISAUDIO’’ VI4-001 research program, by the LABEX CELYA (ANR-11-LABX-0060), the LABEX CORTEX (ANR-11-LABX-0042) of Universite´ de Lyon, within the program ‘‘Investissements d’Avenir’’ (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

1. Wilson RH, McArdle R (2005) Speech signals used to evaluate functional status of the auditory system. J Rehabil Res Dev 42:79. doi: 10.1682/JRRD.2005.06.0096
2. Metselaar M, Maat B, Krijnen P, et al (2008) Comparison of speech intelligibility in quiet and in noise after hearing aid fitting according to a purely prescriptive and a comparative fitting procedure. Eur Arch Otorhinolaryngol 265:1113–1120. doi: 10.1007/s00405-008-0596-x
3. Moon IJ, Kim EY, Jeong JO, et al (2012) The influence of various factors on the performance of repetition tests in adults with cochlear implants. Eur Arch Otorhinolaryngol 269:739–745. doi: 10.1007/s00405-011-1699-3
4. Gelfand SA (2001) Essentials of audiology. Thieme, New York
5. Hudgins CV, Hawkins JE (1947) The development of recorded auditory tests for measuring hearing loss for speech. The Laryngoscope 57:57–89.
6. Miller GA, Heise GA, Lichten W (1951) The intelligibility of speech as a function of the context of the test materials. J Exp Psychol 41:329.
7. Fournier J-E (1951) Audiométrie vocale: les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités. Maloine, Paris, France
8. Lafon J-C (1964) Le Test phonétique et la mesure de l'audition. Dunod, Paris, France
9. Legent F, Bordure P, Calais C, et al (2011) Audiologie pratique, audiométrie. Elsevier, Masson, Paris

10. Marian V, Bartolotti J, Chabal S, Shook A (2012) CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. PLoS ONE 7:e43230. doi: 10.1371/journal.pone.0043230

11. New B, Pallier C, Brysbaert M, Ferrand L (2004) Lexique 2: a new French lexical database. Behav Res Methods Instrum Comput J Psychon Soc Inc 36:516–524.

12. Brysbaert M, New B (2009) Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods 41:977–990. doi: 10.3758/BRM.41.4.977

13. Luce PA, Pisoni DB (1998) Recognizing spoken words: The neighborhood activation model. Ear Hear 19:1.

14. Meyer TA, Pisoni DB (1999) Some computational analyses of the PBK test: effects of frequency and lexical density on spoken word recognition. Ear Hear 20:363–371.

15. Moulin A, Richard C (2015) Lexical Influences on Spoken Spondaic Word Recognition in Hearing-Impaired Patients. Front Neurosci 9:476. doi: 10.3389/fnins.2015.00476

16. Dillon H (1982) A quantitative examination of the sources of speech discrimination test score variability. Ear Hear 3:51–58.

17. Gelfand SA (1998) Optimizing the reliability of speech recognition scores. J Speech Lang Hear Res JSLHR 41:1088–1102.

18. Schlauch RS, Anderson ES, Micheyl C (2014) A demonstration of improved precision of word recognition scores. J Speech Lang Hear Res 57:543–555. doi: 10.1044/2014_JSLHR-H-13-0017

19. HAS Haute autorité de santé (2007) Traitement de la surdité par pose d'implants cochléaires ou d'implants du tronc cérébral. http://www.has-

sante.fr/portail/jcms/r_1498772/fr/traitement-de-la-surdite-par-pose-d-implants-cochleaires-ou-d-implants-du-tronc-cerebral (last accessed the 9th of September 2016).

20. Blamey PJ, Maat B, Başkent D, et al (2015) A Retrospective Multicenter Study Comparing Speech Perception Outcomes for Bilateral Implantation and Bimodal Rehabilitation. Ear Hear 36:408–416. doi: 10.1097/AUD.0000000000000150
21. Lenarz T, James C, Cuda D, et al (2013) European multi-centre study of the Nucleus Hybrid L24 cochlear implant. Int J Audiol 52:838–848. doi: 10.3109/14992027.2013.802032
22. Mosnier I, Felice A, Esquia G, et al (2013) New cochlear implant technologies improve performance in post-meningitic deaf patients. Eur Arch Otorhinolaryngol 270:53–59. doi: 10.1007/s00405-011-1918-y
23. BIAP International bureau for audiophonology Audiometric Classification of Hearing Impairments. http://www.biap.org/en/recommandation/recommendations-pdf (accessed the 9th September 2016).
24. Cohen J (1992) A power primer. Psychol Bull 112:155–159. doi: 10.1037/0033-2909.112.1.155
25. Thornton AR, Raffin MJ (1978) Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res 21:507–518.
26. Alfelasi M, Piron JP, Mathiolon C, et al (2013) The transtympanic promontory stimulation test in patients with auditory deprivation: correlations with electrical dynamics of cochlear implant and speech perception. Eur Arch Otorhinolaryngol 270:1809–1815. doi: 10.1007/s00405-012-2125-1

27. Rumeau C, Frère J, Montaut-Verient B, et al (2015) Quality of life and audiologic performance through the ability to phone of cochlear implant users. Eur Arch Otorhinolaryngol 272:3685–3692. doi: 10.1007/s00405-014-3448-x

28. Vincent C, Renard C, Blond S, Lejeune J-P (2012) [Auditory evaluation in the management of acoustic neurinoma]. Neurochirurgie 58:282–286. doi: 10.1016/j.neuchi.2012.05.008

29. Goldinger SD (1996) Auditory lexical decision. Lang Cogn Process 11:559–568.

30. Savin HB (1963) Word-Frequency Effect and Errors in the Perception of Speech. J Acoust Soc Am 35:200–206. doi: 10.1121/1.1918432

31. New B, Brysbaert M, Veronis J, Pallier C (2007) The use of film subtitles to estimate word frequencies. Appl Psycholinguist 28:661.

32. Brysbaert M, Buchmeier M, Conrad M, et al (2011) The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. Exp Psychol 58:412–424. doi: 10.1027/1618-3169/a000123

33. Dirks DD, Takayanagi S, Moshfegh A, et al (2001) Examination of the neighborhood activation theory in normal and hearing-impaired listeners. Ear Hear 22:1–13.

34. Di Berardino F, Forti S, Mattei V, et al (2010) Non-verbal visual reinforcement affects speech audiometry in the elderly. Eur Arch Oto-Rhino-Laryngol 267:1367–1370. doi: 10.1007/s00405-010-1254-7

35. Löhler J, Akcicek B, Wollenberg B, et al (2015) Results in using the Freiburger monosyllabic speech test in noise without and with hearing aids. Eur Arch Otorhinolaryngol 272:2135–2142. doi: 10.1007/s00405-014-3039-x

36. Harris RW, Nissen SL, Pola MG, et al (2007) Psychometrically equivalent Russian speech audiometry materials by male and female talkers: Materiales de logoaudiometría en

ruso psicométricamente equivalentes para hablantes masculinos y femeninos. Int J Audiol 46:47–66.

37. Iliadou V, Fourakis M, Vakalos A, et al (2006) Bi-syllabic, Modern Greek word lists for use in word recognition tests. Int J Audiol 45:74–82. doi: 10.1080/14992020500376529

38. Han D, Wang S, Zhang H, et al (2009) Development of Mandarin monosyllabic speech test materials in China. Int J Audiol 48:300–311. doi: 10.1080/14992020802607456

39. Wang S, Mannell R, Newall P, et al (2007) Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. Int J Audiol 46:719–731. doi: 10.1080/14992020701558511

40. Nissen SL, Harris RW, Channell RW, et al (2011) The development of psychometrically equivalent Cantonese speech audiometry materials. Int J Audiol 50:191–201. doi: 10.3109/14992027.2010.542491

41. van Heuven WJB, Mandera P, Keuleers E, Brysbaert M (2014) SUBTLEX-UK: a new and improved word frequency database for British English. Q J Exp Psychol 67:1176–1190. doi: 10.1080/17470218.2013.850521

42. Vega FC, Nosti MG, Gutiérrez AB, Brysbaert M (2011) SUBTLEX-ESP: Spanish word frequencies based on film subtitles. Psicológica Rev Metodol Psicol Exp 32:133–143.

43. Moulin A, Richard C (2016) Sources of variability of speech, spatial, and qualities of hearing scale (SSQ) scores in normal-hearing and hearing-impaired populations. Int J Audiol 55:101–109. doi: 10.3109/14992027.2015.1104734

44. Moulin A, Richard C (2016) Validation of a French-Language Version of the Spatial Hearing Questionnaire, Cluster Analysis and Comparison with the Speech, Spatial, and Qualities of Hearing Scale. Ear Hear 37:412-23. doi: 10.1097/AUD.0000000000000269

45. Akeroyd MA, Arlinger S, Bentler RA, et al (2015) International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests: ICRA Working Group on Multilingual Speech Tests. Int J Audiol 54 Suppl 2:17-22. doi: 10.3109/14992027.2015.1030513

46. Kollmeier B (2015) Overcoming language barriers: Matrix sentence tests with closed speech corpora. Int J Audiol 54:1–2. doi: 10.3109/14992027.2015.1074295

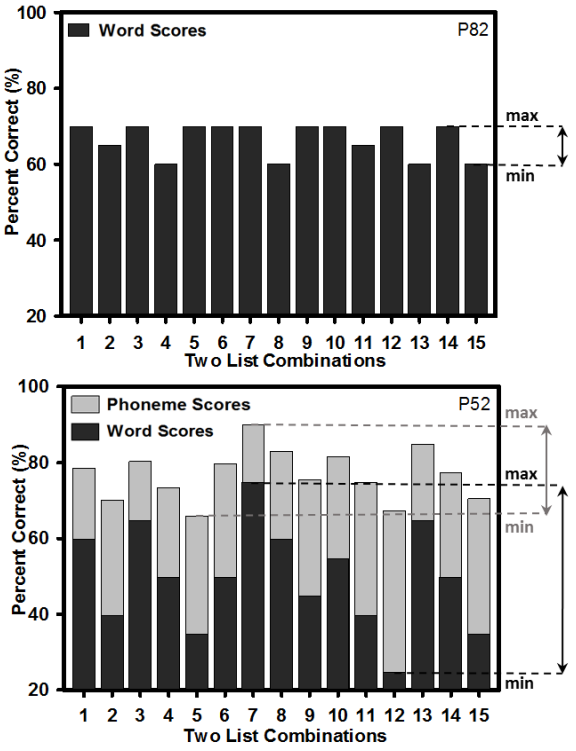

Fig. 1. Inter-list variability of word recognition scores.
Two patients (P52 and P82) word scores are represented for each combination of two lists (i.e. for 15 combinations) showing a high variability with a range of 50% for P52, and a small variability and range for P82. For P52, both word scores (black bars) and phoneme scores (grey bars) are presented illustrating the much lower variability for phonemes scores than for word scores.

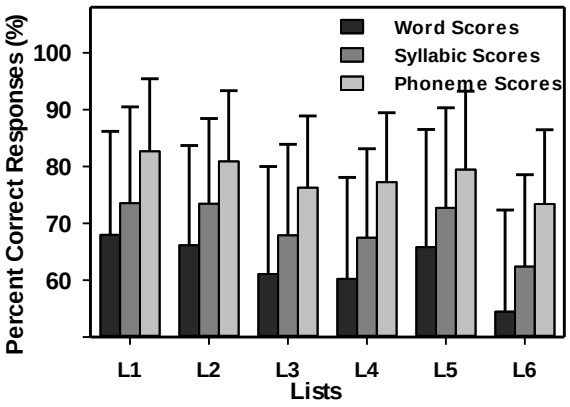

Fig. 2: Percent spoken word recognition scores (mean +/- sem) for each of the 6 disyllabic Fournier lists (L1 to L6) in whole word counts (black bars), syllabic counts (dark grey bars) and phoneme counts (grey bars). All pairwise comparisons are highly statistically significant except the comparison between L2 and L5.

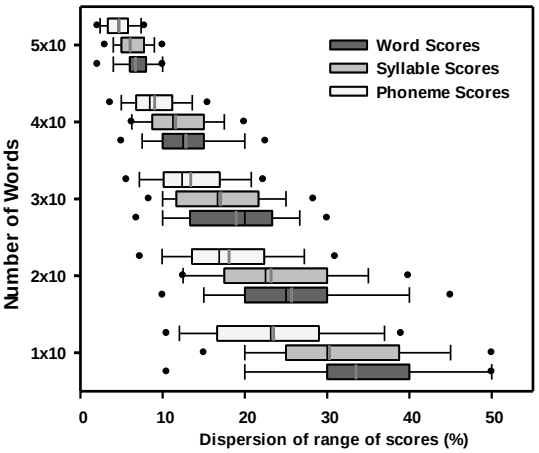

Fig. 3. The dispersion of the range of variability of scores obtained in words (dark grey bars), syllables (grey bars) and phonemes (light-grey bars) across 120 patients is shown as a function of the number of lists of 10 words considered (from combinations of 1 list to 5 lists). Box and whiskers represent the 25^{th}/75^{th} percentiles and 5th/95th percentiles, respectively. The black line within the boxes represents the median score, and the grey one the mean scores.

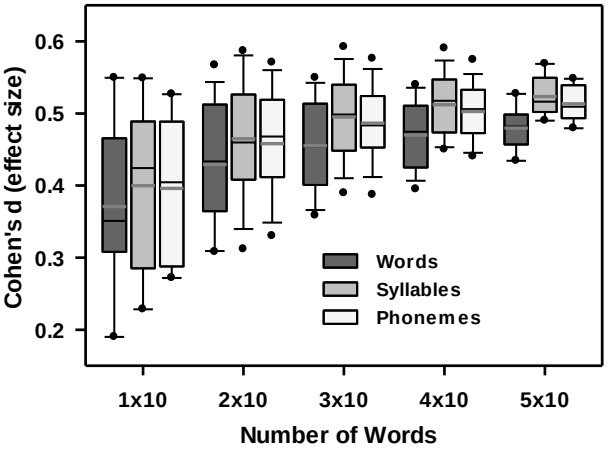

Fig. 4. Box-plots of effect size (Cohen's *d*) obtained in the comparison of two groups of patients (one group with averaged pure tone audiometry (PTA) < 43.5 dB HL, and one group with PTA > 43.5 dB HL). This was done for different combinations of lists and different scores: word (dark grey bars), syllable (grey bars) and phoneme (light-grey bars) scores. Box and whiskers represent the $25^{th}/75^{th}$ percentiles and 5th/95th percentiles, respectively. The black lines within the boxes represent the median scores and the grey lines are the mean scores.

	PTA	MW Score	Fournier Lists (Spondaic words)					
			L1	L2	L3	L4	L5	L6
Age	0.13	0.02	0.03	-0.04	-0.08	-0.02	-0.05	-0.06
PTA	1	**-0.23**	**-0.26**	-0.10	**-0.30**	**-0.28**	**-0.32**	**-0.25**
MW Score		1	**0.34**	**0.33**	**0.47**	**0.43**	**0.5**	**0.36**
L1			1	**0.52**	**0.43**	**0.44**	**0.51**	**0.51**
L2				1	**0.54**	**0.54**	**0.55**	**0.49**
L3					1	**0.54**	**0.57**	**0.47**
L4						1	**0.58**	**0.61**
L5							1	**0.56**

Table 1.

Pearson correlation coefficients between the scores obtained for different word lists (labeled from L1 to L6), the averaged pure tone audiometry (PTA in dB HL), the age (in years), and the monosyllabic word scores (MW score) obtained at the same stimulus intensity as the spondaic word lists. The correlation in bold are significant with $p < 0.01$, and highly significant correlations are in grey ($p < 0.001$).