Beyond speech intelligibility: Using response times, sound quality, and task load to evaluate the benefit of noise reduction

Christophe Lesimple[1] and Julie Tantau[1]

INTRODUCTION

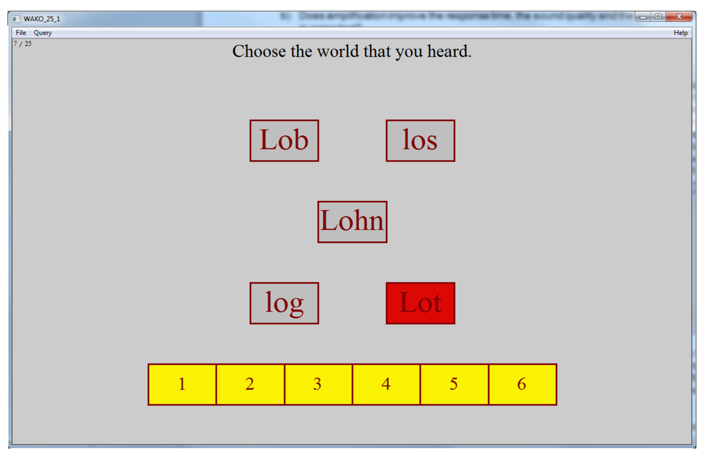
Noise reduction (NR) systems are found widely within digital hearing instruments. Objective benefits of this feature have been difficult to demonstrate using currently available speech-in-noise intelligibility tests. Alternative outcomes have been shown to be good indicators of NR benefit such as response times (van den Tillaart et al., 2013), sound quality (Alcantara, et al., 2003), and listening effort (Desjardins et al., 2014). A word-based closed-set speech recognition test based on the Wallenberg & Kollmeier Rhyme Test (WAKO) (Wallenberg and Kollmeier, 1989) combined with a modified version of the National Aeronautics and Space Administration (NASA)Task Loading Index (Hart & Staveland, 1988) were developed in order to evaluate the effect of NR on speech intelligibility (SI), response time (RT), sound quality (SQ), and task load (TL). The TL combines different aspects of perceived workload including mental demand, performance, effort, and frustration. The following research questions were addressed in this investigation:

1. Does amplification improve performance on the tested variables using a closed-set task?
2. Does the NR algorithm provide any measurable benefit with this test design?

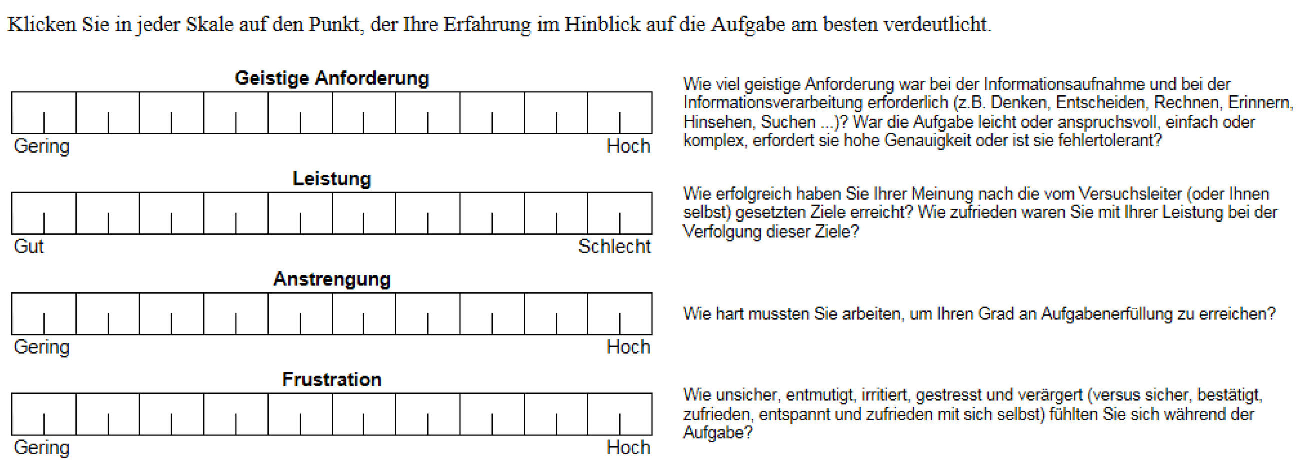
MATERIAL AND METHODS

Test setup

The WAKO consists of 18 lists each with 25 words of equal intelligibility, including a carrier sentence. Before each stimuli, five rhyming or similar alternatives were displayed on a touch screen. The subject was asked to indicate the answer most similar to what was heard and then to indicate the sound quality for each word on a 6 point scale, as in the Swiss school grading system where 1 is rated as being poor and 6 is good. A test interface using PRAAT software (version 5.3.77: Boersma & Weenink, 2014) was used to collect the chosen answer, RT, and SQ.



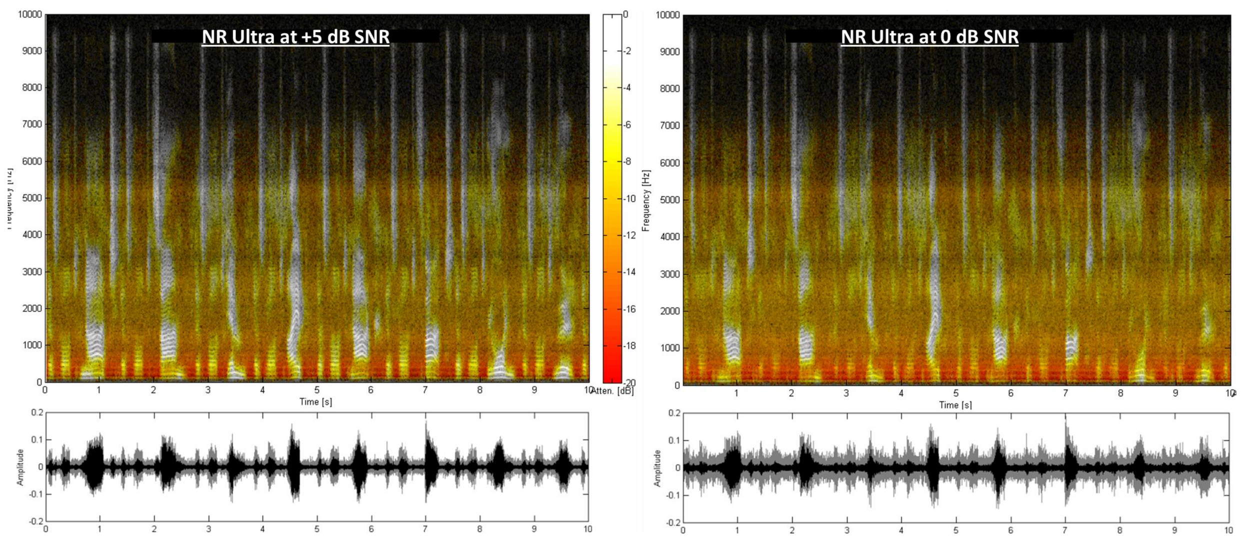
An indirect measurement of subjective TL was assessed with a modified version of the NASA Task Load Index (TLx) (Hart & Staveland, 1988). After each test list, the subject was asked to rate his perceived workload using a 100-point scale for the following subscales: mental demand, performance, effort, and frustration.



Signal and noise were presented at +5 dB SNR and 0 dB SNR via a two channel audiometer (GSI 61) and a single active loudspeaker at 0° (Fostex 6301B). The speech level was fixed at 65 dB A. The test condition order was assigned with a Latin square design: unaided, aided, and aided with NR.

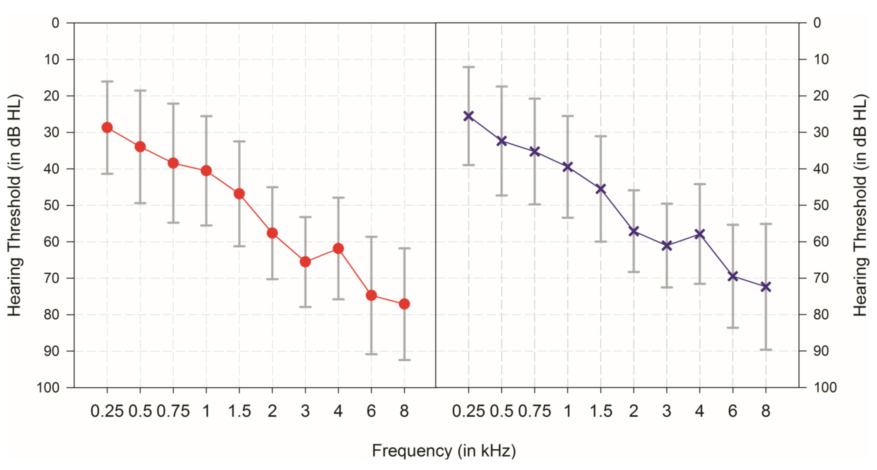
Test Hearing Instrument

Receiver-In-The-Ear hearing instruments were used for this test . Automatic features such as expansion, transient noise reduction, and reverberation reduction were disabled. The microphone mode was set to omnidirectional. The NAL-NL2 fitting rational target match was verified with real ear measurements at 65 dB input (Audioscan Verifit v. 3.10.71). The tested noise reduction is commercially available in a Bernafon HI and is a multichannel modulation based system. The efficiency of the feature was verified with differential spectrograms and by measuring the output SNR with the test material at +5 and 0 dB SNR : flat linear 20 dB insertion gain with NR on and off.



Test subjects

- 19 experienced (> 6 months) HI users
- Average age = 66.9 years old (min = 45, max =84)
- Hearing loss:
  - 4 mild (26-40 dB HL)
  - 11 moderate (41-55 dB HL)
  - 4 moderate to severe (56-70 dB HL)



HI output SNR measured using the inversion technique as described by Hagerman and Olofsson (2004):

Input SNR		Output SNR
at 0 dB SNR	NR off	0.3 dB
	NR on	5.4 dB
	Effect of NR	+5.10 dB SNR
at 5 dB SNR	NR off	5.33 dB
	NR on	10.3 dB
	Effect of NR	+4.97 dB SNR

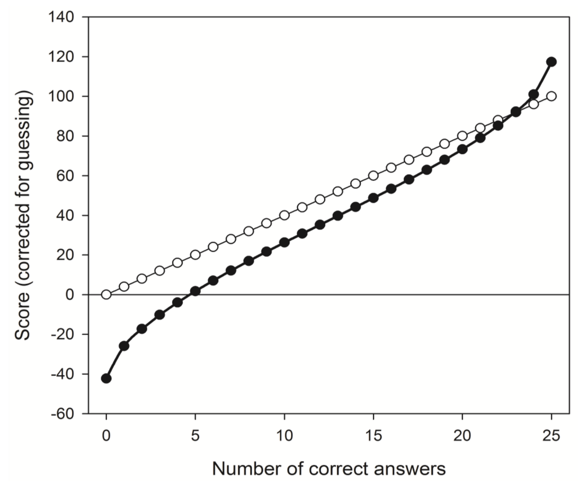
DATA ANALYSIS

Speech Intelligibility: correction for variance and guessing

Speech scores were transformed according to recommendations by Sherbecoe and Studebaker (2004) for closed-set tests:

- An arcsine transform for variance correction,
- A correction for guessing (i.e. 20%)

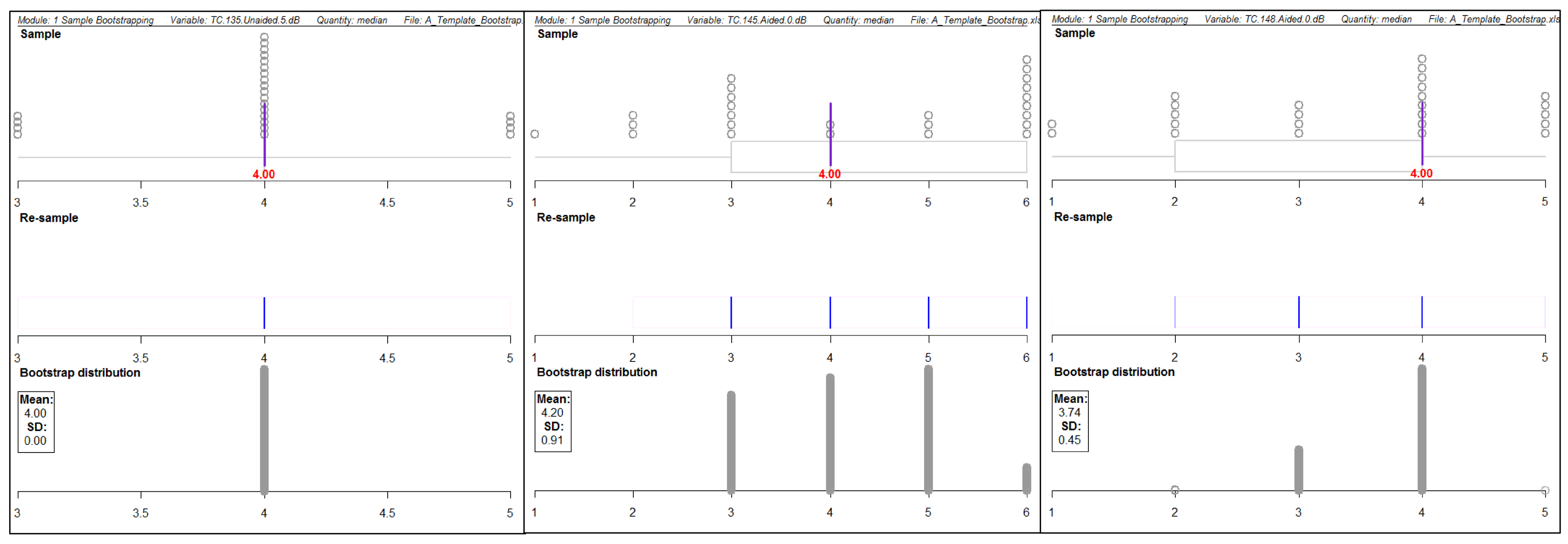
The graph shows the correspondence between the raw scores (white circles) and the corrected scores (black circles).



Bootstrapping: correction for ordinal data

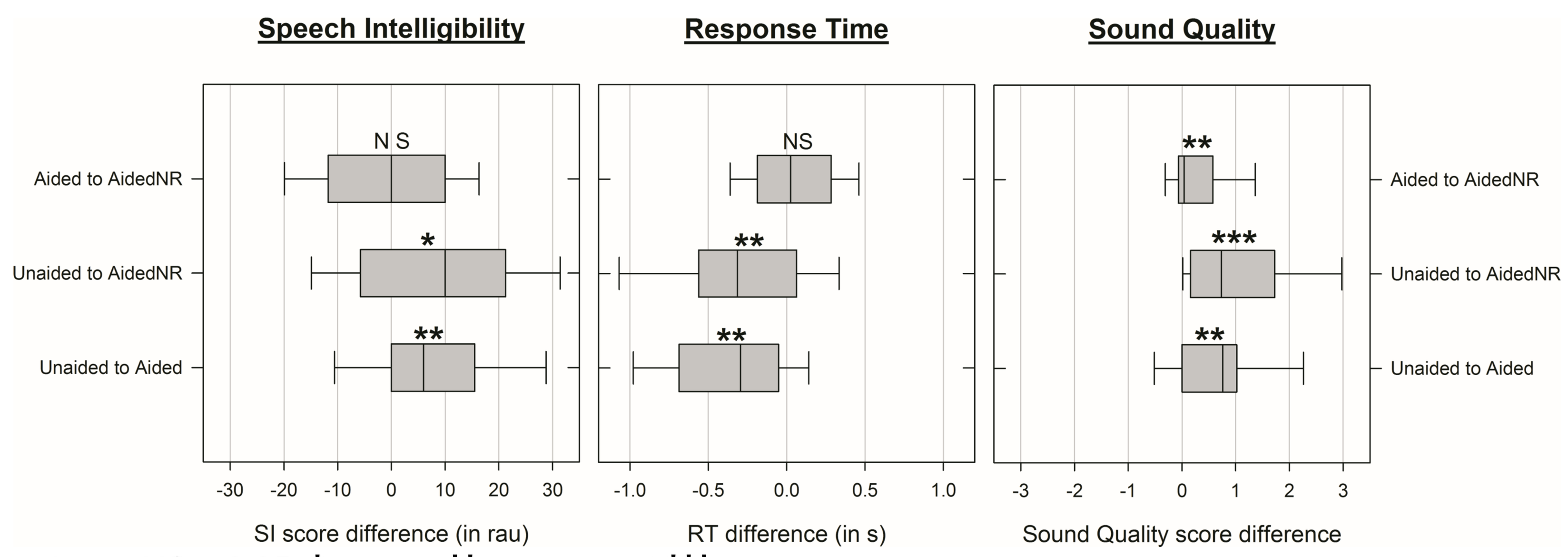
Ordinal data from the SQ rating and the TL index indicate an ordered structure but not the magnitude of the difference (Svensson, 2003). Re-sampling techniques can improve the precision of the parameter estimation. The figures (below) illustrate three individual cases from the SQ rating that have the same median but different distributions, along with the result from the bootstrapping procedure:

- Symmetrical distribution, observed median = bootstrapped mean = 4,
- Bimodal distribution, observed median ≠ bootstrapped mean = 4.2,
- Asymmetrical distribution (left skewed), observed median ≠ bootstrapped mean = 3.74.



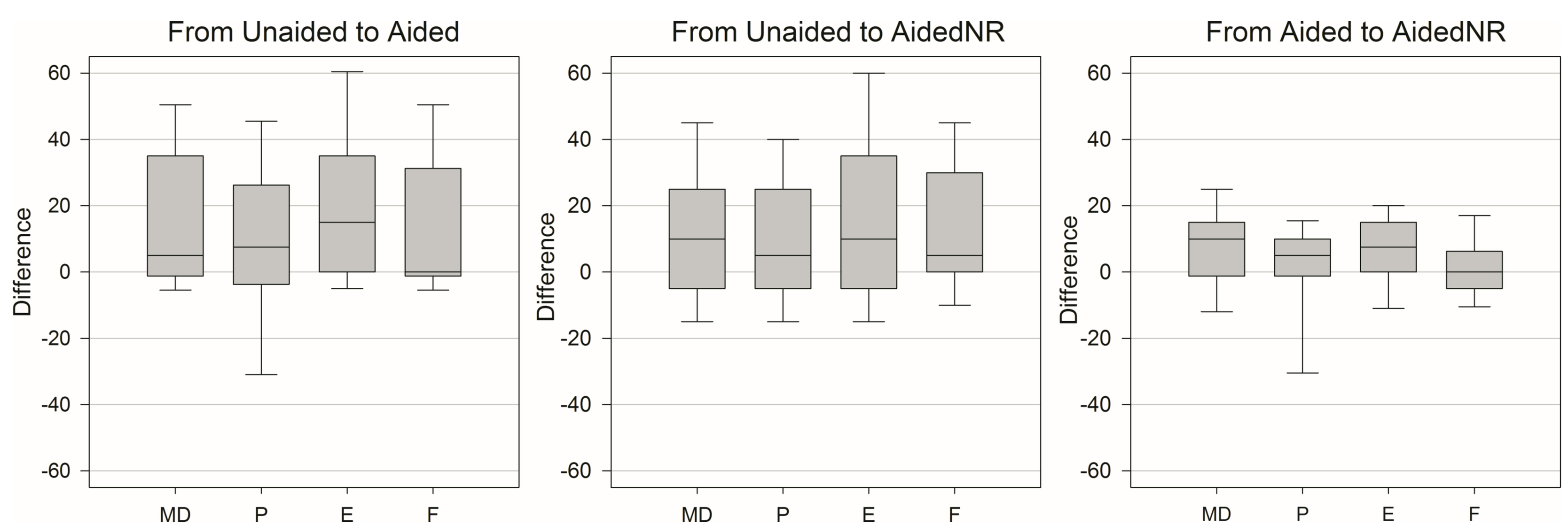
RESULTS

The measured outcomes were subjected to a two-way repeated measures analysis of variance (ANOVA) with repetition of both factors i.e. test condition (unaided, aided, and aided with NR) and SNR (0 and +5 dB SNR). The main SNR effect was significant (SI and SQ $p < 0.001$ and RT $p = 0.008$) as well as the main test condition effect (SI $p = 0.02$, RT $p = 0.001$, and SQ $p < 0.001$) . Results from the pairwise multiple comparison procedures (Fisher LSD) between the test conditions are summarized in the figures below.



P-values : NS ≥ 0.05, ***** < 0.05, ****** < 0.01, and ******* < 0.001.

The distribution of the TL score differences between all test conditions and the different subscales: mental demand (MD), performance (P), effort (E), and frustration (F) are shown in the figures below.



The boundary of the box closest to zero indicates the 25th percentile, the line within the box marks the median, and the boundary of the box farthest from zero indicates the 75th percentile. Error bars above and below the box indicate the 90th and 10th percentiles respectively.

Bootstrapped confidence intervals (CI) were computed for each TL score difference distribution and are seen in the table.

Only two comparisons had a CI above 0 indicating a systematic change, i.e. there was a perceived benefit with NR on the effort scale between the unaided and aided conditions.

	Bootstraped Confidence Interval		
Mental Demand	Min	Median	Max
Unaided to Aided	-5	5	20
Unaided to AidedNR	0	5	20
Aided to AidedNR	0	10	15
Performance			
Unaided to Aided	-5	5	17.5
Unaided to AidedNR	0	5	20
Aided to AidedNR	0	5	7.5
Effort			
Unaided to Aided	-5	7.5	27.5
Unaided to AidedNR	5	15	35
Aided to AidedNR	2.5	7.5	12.5
Frustration			
Unaided to Aided	0	25	45
Unaided to AidedNR	0	0	30
Aided to AidedNR	-5	0	5

CONCLUSIONS

It was found that amplification significantly improved SI, RT, and SQ. As expected there were no significant differences in SI when the aided conditions were compared. The NR significantly improved the SQ and effort scores which correspond with the measured output SNR improvement. Speech scores from speech-in-noise tests only measure one aspect of speech perception. Future research should examine the broader picture by further investigating the implications of other aspects of speech perception.

We want to thank N. Hockley for his support and review, M. Kuriger for the DSP design and technical help, and D. Bueb for software assistance on the Task Load index.

REFERENCES

Alcantara, J. I., Moore, B. C. J., Kuhnel, V., & Launer, S. (2003) . Evaluation of the noise reduction system in a commercial digital hearing aid. Int. J. Audiol. 42, 34–42.

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. Ear and Hearing. 35: 600-10.

Van den Tillaart-Haverkate, M., Brons, I., Dreschler, W.A., & Houben, R. (2013). The influence of noise reduction on the response time to digits in noise. Poster session presented at ISAAR, Nyborg, Denmark.

Hagerman, B., & Olofsson, A. (2004). A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. Acta Acustica, 90, 356–361.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press

Sherbecoe R.L. & Studebaker G.A. (2004) Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units. Int J Audiol. 43(8), 442-8.

Svensson, E. (2003). Statistical methods for repeated qualitative assessments on scales. International Journal of Audiology, 42 (s1), S13-22.

Wallenberg, E.L. & Kollmeier, B., (1989). Sprachverständlichkeitsmessungen für die Audiologie mit einem Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten. Audiologische Akustik, 38, 50-65.

For more information please contact cl@bernafon.ch

[1] **Bernafon AG Bern Switzerland**

bernafon

Your hearing · Our passion