International Journal of Audiology

ISSN: 1499-2027 (Print) 1708-8186 (Online) Journal homepage: http://www.tandfonline.com/loi/iija20

International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests

Michael A. Akeroyd, Stig Arlinger, Ruth A. Bentler, Arthur Boothroyd, Norbert Dillier, Wouter A. Dreschler, Jean-Pierre Gagné, Mark Lutman, Jan Wouters, Lena Wong & Birger Kollmeier

To cite this article: Michael A. Akeroyd, Stig Arlinger, Ruth A. Bentler, Arthur Boothroyd, Norbert Dillier, Wouter A. Dreschler, Jean-Pierre Gagné, Mark Lutman, Jan Wouters, Lena Wong & Birger Kollmeier (2015): International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests, International Journal of Audiology

To link to this article: http://dx.doi.org/10.3109/14992027.2015.1030513

informa
healthcare

International Journal of Audiology 2015; Early Online: 1–6

International
Journal of
Audiology

Technical Note

International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests

ICRA Working Group on Multilingual Speech Tests

Michael A. Akeroyd*, Stig Arlinger†, Ruth A. Bentler‡, Arthur Boothroyd#, Norbert Dillier$, Wouter A. Dreschler^, Jean-Pierre Gagné+, Mark Lutman¶, Jan Wouters§, Lena Wong** & Birger Kollmeier††

*MRC Institute of Hearing Research, Nottingham, UK, †Linköping University, Department of Clinical and Experimental Medicine, Technical Audiology, Sweden, ‡Department of Communication Sciences & Disorders, The University of Iowa, Wendell Johnson Speech and Hearing Center, Iowa, USA, #San Diego State University, San Diego, CA, USA, $University of Zurich, ENT Department, Zürich, Switzerland, ^Academic Medical Centre, Amsterdam, The Netherlands, +Université de Montréal. Montréal, Québec, Canada, ¶Institute of Sound and Vibration Research, University of Southampton, Highfield, Southampton, UK, §KU Leuven, Dept. Neurosciences, ExpORL, Leuven, Belgium, **Division of Speech & Hearing Sciences, University of Hong Kong, Hong Kong and ††Cluster of Excellence Hearing4All & Medizinische Physik, Universität Oldenburg, and HörTech gGmbH, Oldenburg, Germany

Abstract

Objective: To provide guidelines for the development of two types of closed-set speech-perception tests that can be applied and interpreted in the same way across languages. The guidelines cover the digit triplet and the matrix sentence tests that are most commonly used to test speech recognition in noise. They were developed by a working group on Multilingual Speech Tests of the International Collegium of Rehabilitative Audiology (ICRA). *Design:* The recommendations are based on reviews of existing evaluations of the digit triplet and matrix tests as well as on the research experience of members of the ICRA Working Group. They represent the results of a consensus process. *Results:* The resulting recommendations deal with: Test design and word selection; Talker characteristics; Audio recording and stimulus preparation; Masking noise; Test administration; and Test validation. *Conclusions:* By following these guidelines for the development of any new test of this kind, clinicians and researchers working in any language will be able to perform tests whose results can be compared and combined in cross-language studies.

Key Words: Speech perception; behavioral measures; psychoacoustics/hearing science; instrumentation

Preamble

An important goal in audiology is to establish internationally standardized diagnostic tests for measuring speech recognition in noise or in quiet that can be applied and interpreted in the same way across languages—an endeavor with major challenges as soon as speaker and language-specific considerations come into play. A prerequisite of such multilingual speech tests is that their construction and their results are as comparable across languages as possible. In turn, this requires a published set of guidelines for the development of any new test of this kind in any new language. The International Collegium of Rehabilitative Audiology (ICRA) is ideally placed to produce such guidelines, as it is an international body of leading experts in the field. The guidelines reported here were developed by the ICRA Working Group on Multilingual Speech Tests constituted by the authors of this paper. They complement the ISO 8253-3:2012 standard 'Acoustics- Audiometric test methods - Part 3: Speech audiometry' by considering in more detail the steps necessary for developing a specific test in any language.

The test formats considered here have a closed-set design (i.e. with a limited number of response alternatives that the test subject may choose from in order to report which speech item she or he has heard), which is preferred for multilingual applications, as they have the advantage that they can be performed in an individual's native language even if the test conductor does not understand this language. The tests can be performed in quiet to test speech recognition at the individual's absolute threshold or, more commonly, in noise to test suprathreshold hearing acuity at a predefined level of a selected background noise. The latter condition is particularly

Correspondence: Birger Kollmeier, Cluster of Excellence Hearing4All & Medizinische Physik , Universität Oldenburg, Oldenburg, D-26111 Germany. E-mail: birger.kollmeier@uni-oldenburg.de

(Received 23 November 2014; accepted 12 March 2015)

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.3109/14992027.2015.1030513

important as it represents the daily communication situation of speech masked by some kind of speech-related background noise. Though many speech-in-noise tests have been developed for these tasks over at least the last 75 years, in this report we focus on two closed-set designs that have proven to be highly useful and precise measurement tools, namely the digit triplet test (being a prototype for a hearing screening test) and the matrix test (a prototype test for professional use for the evaluation of auditory rehabilitation and audiological diagnostics).

Though our recommendations specifically concern these two tests, many of the principles are of general applicabilty. Other speech tests have been developed across languages such as the closed-set Diagnostic Rhyme test in English (Voiers, 1983), Swedish (Risberg, 1976), or German (Sotscheck, 1985; vonWallenberg & Kollmeier, 1989), or open-set speech tests such as the short meaningful sentences in Dutch (Plomp & Mimpen, 1979), British English (BKB sentences, Bench et al, 1979), or German (Kollmeier & Wesselkamp, 1997), or the HINT sentences (Soli & Wong, 2008). However, the digit triplet tests and matrix tests have the advantage that the inventory of speech items used is limited and each test list usually includes the complete inventory. This helps to make the test items and test lists very homogeneous in intelligibility and comparable in construction across languages. Also, as an increasing number of comparable digit and matrix tests in different languages have been introduced within the last years, some standardization is timely.

The digit triplet test was originally designed by Smits and colleagues in Dutch (Smits et al, 2004). Versions have since been developed in about 15 languages, including British English, American English, German, French, Italian, Mandarin, Polish, Russian, and Spanish (see Zokoll et al 2012, 2013, for a comprehensive review). It was primarily designed for use as a screening test using a telephone. In the test, a sequence of three digits (e.g. 'six three four') is presented together with a background noise to the listener, who then has to press the appropriate digits on the telephone keypad as a response. The trial is marked correct if all three digits are reported, in the right order. An adaptive tracking algorithm alters the speech-to-noise ratio (SNR) according to the listener's responses in such a way that the individual's speech recognition threshold (SRT, i.e. the SNR corresponding to 50% (or in some versions 80%) speech intelligibility) is determined quickly within just a few trials.

The matrix test was originally proposed by Hagerman (1982) in Swedish. A modified version was described by Wagener et al (1999a,b,c) in German, and is now available in 14 languages, including American English, British English, Dutch, German, French, Spanish, Turkish, and Russian (see Kollmeier et al, 2015, for an extensive review). It is intended as an audiological diagnostic sentence recognition test. It consists of five-word long sentences, each of which has the same syntax but is not necessarily meaningful, i.e. the semantic content is unpredictable (e.g. 'Thomas wins eight red shoes'; 'Kathy bought two dark spoons'). This is achieved by creating 10 choices for each of the five word positions (name, verb, numeral, adjective, and object; though their order depends on the language). This 'base matrix' is thus fifty words. Each sentence is a random walk (in the mathematical sense) through this matrix. The sentences are grouped into test lists of ten sentences, each containing the fifty words exactly once. Each list is of equivalent intelligibility so lists can be interchanged (but note that sentences cannot). The individual's speech recognition threshold (either 50% or 80% speech intelligibility; see below) is determined by an adaptive tracking procedure using one, two, or even three whole test lists (in the order of decreasing variability of the threshold estimate).

In order to maximize the comparability of test results from these tests across different languages, the procedures and design principles for constructing, recording, optimizing, evaluating, and validating the test in each respective language need be as closely matched as possible. Table 1 lists our recommendations. These recommendations are based on reviews of the existing tests of the digit triplet and matrix tests as well as on the research experience from the members of the ICRA Working Group, and represent the results of a consensus process within the ICRA group. Note that not all existing tests exactly fulfill all the specifications listed in Table 1. The recommendations listed in Table 1 should be self-explanatory, especially when read in combination with the review papers by Zokoll et al (2012) and Kollmeier et al (2015), though the following remarks on each stage may help.

General construction

For each new test, the structure of the words/sentences should be as close as possible to the construction of the existing tests, though it is not necessary for every word in a new test to be a direct translation of every word in an existing test. While the string of digits for most languages considered so far does not usually pose a problem, in the matrix test the sentence syntax and possible dependencies between the words in the sentence might cause language-specific difficulties. For example, in Spanish the inflection of the adjective depends on the gender of the noun, and so only male nouns were selected to avoid changes in pronunciation of the adjective. It should be verified that combinations across word groups do not change the pronunciation of any word in a sentence, except for unavoidable coarticulation effects at word transitions.

Word selection

For the digit triplets test the number of syllables for the digits from zero to nine needs to be considered to avoid a certain digit being recognized purely by its unique number of syllables. This may reduce the number of digits to be actually used during the test. For the matrix test, the number of syllables should be balanced within each word group. Since ideally the matrix test should also be usable for children (either in its original form or in an abbreviated form with fewer words per sentence and fewer options per word group—see Wagener & Kollmeier, 2005, and Neumann et al, 2012—to be used with children aged four years and older), the words employed should be as familiar as possible to the broadest range of the public. The words and sentences should also be neutral with respect to emotions or other features that might cause non-acoustic influences on intelligibility. The phoneme distribution of the underlying language should be approximated as closely as possible by the base matrix. The number of words and phonemes in the base matrix is usually large enough to maintain this, though sometimes it is necessary to change some words in the base matrix in order to avoid major deviations in the phoneme distribution.

Speaker

The speech used for a speech recognition test is there to assess an individual listener's ability for everyday communication situations, not to specifically address hearing-impaired listeners or to act to an audience. Hence, the speaker selected to make the recordings for any new language should not necessarily be a formally trained speaker or actor with any extreme speech quality, but rather a

Table 1. Recommendations of the construction of multilingual speech tests.

	Digit triplet test	*Matrix test*
General construction	• Three digits between zero and nine in one utterance • Each test list contains each digit three times at the respective position in the triplet	• Base matrix of 50 words (10 names, 10 verbs, 10 numerals, 10 adjectives, 10 objects) • Word-order is language-specific • All combinations of words must result in grammatically correct sentences • Word pronunciation should be equal across all possible word combinations in the sentence (except for coarticulation between successive words)
Word selection	• Balanced number of syllables • Short announcement phrase (to focus attention) with increased level with respect to the triplets (up to 3 dB in order to be audible for hearing-impaired listeners)	• Balanced number of syllables within word groups • Highly frequent words (frequency dictionary), preferably familiar to children • Semantic neutrality of words and sentences • Language-specific phoneme distribution
Speaker	• Natural intonation • Standard language pronunciation • Native speaker, not necessarily formally trained • Constant vocal effort • Average speech rate depends on language, but about 200–350 syllables per minute (e.g. Russian 200 spm, Spanish 327 spm)	
Recording	• Equipment: see ISO 8253-3:2012 (Acoustics - Audiometric test methods - Part 3: Speech audiometry)	
	• Record each digit at each position of the three triplet positions to account for intonation aspects • Short pauses between digits	• 100 sentences accounting for co-articulation between words (each word recorded with each subsequent word) • All combinations of two consecutive words (10 realizations of each of the 50 words)
Cutting	• Each digit at each position in the recorded triplet, omitting the pauses	• Cut into single words, preserving co-articulation at the end of the word
Resynthesis	• Join digits into new triplets, using first-position recordings for the first, second-position for the second, and third position for the third • Add pauses (e.g. 160 ms) between successive, individually cut digits	• Join five words into new test sentences with appropriate co-articulation • Each test list contains 10 sentences, together using all 50 words of the base matrix • Each word token should occur equally often across all generated sentences, so facilitating the determination of the word-specific discrimination function • Individual overlap-times of 0 to 300 ms to smooth the coarticulation portions • Testing the naturalness of sound of resynthesized sentences on a group of native-language listeners
Masking noise	• Should have the same long-term spectrum as the speech material • Stationary noise with no added amplitude modulations • Preferred method is a randomized superposition all recorded sentences (or all digit triplets) • If instead it is generated from another noise, an appropriate spectral-shaping filter is necessary	
Optimization	• Purpose is to maximize the homogeneity of the intelligibility across the speech material • Determined by speech intelligibility measurements at fixed SNRs, covering range of 10%–90% in speech intelligibility at a noise level at 65 dB SPL (55–75 dB is acceptable) • Word scoring • At least two lists of 20 sentences as training required per subject prior to data collection at an SNR yielding a high intelligibility score for the Matrix test • Level adjustment of each word realization to reach mean SRT (50% or 80% point, depending on target). Adjustments limited to ± 2–4 dB but can be language specific if necessary. The digit triplets test may use larger values. • Words may be eliminated if the parameter SRT and slope of word-specific discrimination function cannot be obtained within reasonable limits • Separate optimization for special test purposes (e.g. telephone version, or processed speech material) may be performed	
Evaluation 1: Test list equivalence	• Measurements at fixed SNRs for each test list (2 or 3 SNRs corresponding to about 20 and 80% or 20,50,80% correct responses) • Triplet scoring	• Measurements at fixed SNRs for each test list (2 or 3 SNRs corresponding to about 20 and 80% or 20,50,80% correct responses) • Word scoring • Appropriate training of subjects prior to data collection (see optimization).

(*Continued*)

Table 1. (Continued)

	Digit triplet test	*Matrix test*

Evaluation 2: Normative data	• Adaptive 1 up, 1 down method:
	  ∘ fixed step size of 2 dB
	  ∘ An initial SNR at a high level of intelligibility should be chosen
	  ∘ SRT estimated by averaging the SNRs from 5th trial to the last trial (plus next "virtual" SNR)
	• Broadband
	• Optional: If telephone version, then:
	  ∘ Specify Distortion & Band limitation/ Codec
	  ∘ Separate validation/ normative data
	• Mean and Standard Deviation between individuals and Test/Retest should be reported
		• Adaptive procedure with word scoring according to Brand & Kollmeier, 2002* using double test list (two lists of 10 sentences)
		• Noise level fixed, speech level varies (but if SNR > 20 dB then vary noise level and fix speech level)
		• 80%-correct target (word scoring); 50% is an alternative
		• Separate normative data for open- and closed-set version
		• Extent of training effect (using adaptive procedure) should be reported
		• Mean and Standard Deviation between Individuals and Test/ Retest should be reported

Validation	• Multi-centre studies with normal-hearing and hearing-impaired listeners in comparison to the typical country-dependent reference tests	

*Specification: This is a generalization of Hagerman and Kinnefors' (1995) procedure. The level change ΔL is determined by the percentage obtained in the previous sentence *prev*, the target percentage *tar*, the slope of the discrimination function *slope*, and a convergence function *f(i)* which depends on the number *i* of the reversals:

$$\Delta L = \frac{f(i) \times (\text{prev} - \text{tar})}{\text{slope}}.$$

Hagerman & Kinnefors used these values:
$f(i) = 1$, tar = 0.4 and slope = 0.2.
However, the recommended settings as proposed by Brand & Kollmeier (2002) for 50% are:
$f(i) = 1.5 \times 1.41^{-i}$, slope = 0.15 dB^{-i}
Note that in some adaptive rules the step size can decrease with increasing number of reversals *i*. If this is done, it is recommended to restrict the speed factor *f(i)* to a minimum value of 0.1.

normally articulating, speaker with a dialect acceptable to the largest majority of the language users. The speaker must be able to control his/her vocalization effort during the recording session, which could last several hours, and also avoid any steady deterioration of his/her voice quality during the recording. An RMS measurement and equalization on the sentence level is useful to adjust for any differences in vocalization effort across the recorded material. This is important because the resynthesis procedure (see below) combines sentence portions from different parts of the recording session. The final resynthesized sentences should sound as natural as possible and should not contain any unnatural transitions in voice quality.

Some current tests (e.g. German, Polish) were recorded with a male speaker for compatibility with other speech tests in those languages. However, in most tests a female speaker has been selected as an 'acoustic compromise' between male speech and children's speech, and this is the recommendation for any new language.

Recording & resynthesis

ISO 8253-3:2012 provides an up-to-date description of the requirements for recording speech test materials. This should be adhered to.

In both the digit and matrix tests, the recorded speech elements need be segmented appropriately and joined ('resynthesized') to make the final test materials sound as natural as possible. This is necessary because the number of possible combinations of words in both forms of test is far too large for each combination to be separately recorded, and different recorded versions of the same word would cause an increased variability in intelligibility across speech items that could reduce test efficiency. Instead, a smaller number are recorded, covering at least all the words in the base matrix, and then the recordings are segmented and resynthesized.

For the digit triplet test, recording each digit separately for each of the three positions in the triplet achieves a natural prosody of the resynthesized material. The resynthesis procedure involves the removal of the pauses between successive digits and the introduction of pauses with a fixed duration of e.g. 160 ms. There is also an announcement phrase, at a slightly higher SNR, which helps to direct the subject's attention to the first digit presented (this may be of enhanced importance for low SNRs). RMS equalization across each individual recorded digit is not recommended, as the level can vary considerably across digits even if they are adjusted to be equally intelligible. Instead, the average level across several digits during the recording should be kept constant and used as reference speech level.

For the matrix test, at least 100 sentences should be recorded, including all combinations of two consecutive words to account for coarticulation effects at word transitions. This gives 10 realizations of each of the 50 words. The recorded sentences are then cut into single words, preserving coarticulation at the end of the cut word to the required consecutive word, but truncating coarticulation at the word beginning. The test sentences are resynthesized by combining words with appropriate transitions. Test lists of ten sentences each are then generated so that each test list contains all 50 words of the base matrix.

Though there is much careful work required in recording, cutting, and resynthesis, this investment in listening and quality control by

native-language experts for the target language is necessary in order to achieve high quality speech materials that are acceptable for both patients and professional audiologists and to establish an appropriate basis for the subsequent optimization efforts.

Masking noise

Since the highest efficiency in (energetic) masking of the respective speech material is obtained by a spectral match between the (average) target speech and the masker (Hochmuth et al, 2014), the recommended masking noise is a randomized superposition of all words of the test, with a random initial delay and a random delay between successive repetitions of the speech items. The resulting high number of randomly time-shifted speech items added up at each point in time results in a quasi-stationary noise that has the same long term average spectrum as the target speech.

Optimization

This is necessary to achieve the highest possible homogeneity of the intelligibility across the speech material employed. First, the word-specific intelligibility functions have to be determined for all word tokens (i.e. all recordings of the same word employed in the test) with a group of at least 10 normally-hearing, native-language subjects. The data need be taken at fixed SNRs covering a broad range of speech intelligibility for each word token. The result is a determination of which word recordings are of high intelligibility and which of low. The words are then attenuated or amplified accordingly within a limited range to shift the intelligibility functions for each item to be as close together as possible. The result is that the spread of word-specific SRT values is minimized, so increasing the slope of the discrimination function for the test lists (according to the model by Kollmeier, 1990, reviewed by Kollmeier et al, 2015; see also MacPherson & Akeroyd 2014 for a comprehensive review of slopes). Either a 50%- or 80%-word/digit scoring target for the SRT can be employed for the optimization (the 80% digit-scoring point in the triplet test is close to the 50% triplet-scoring point in the final test; cf. Smits & Houtgast, 2006). Note that the optimization measurements and adjustments should be redone for any alterations of the original speech material, for instance if they are presented over the telephone instead of headphones (cf. Figure 1 in Jansen et al, 2010).

Evaluation

An evaluation of the tests should be done with an independent set of normally-hearing listeners. This is to assess the equivalence of all the test lists generated within the optimization process described above, and also to provide normative data for a new language. Since the matrix test is known to have a significant training effect (e.g. Hagerman, 1984), this effect should be measured and reported together with the test results.

In order to obtain stable results, at least two lists (20 sentences) should be used. In order to prove the equivalence of the test lists, speech intelligibility should be measured for each test list at (at least) the two SNRs corresponding to the 20% and 80% points (the 'pair of compromise'), as these taken together allow an efficient simultaneous estimate of SRT and slope (Brand & Kollmeier, 2002). A list-specific discrimination function can also be estimated (Wichmann & Hill, 2001a,b).

Obtaining normative values for any new test language should be performed using similar methods to those used for the current tests in other languages. Most of the work using the matrix test has taken 50% word intelligibility as the target, though ISO 8523-3:2012 (and clinical experience) support the argument to use the 80%-point instead as the threshold criterion. This roughly corresponds to a 50% target for sentence scoring (i.e. counting a response only as 'correct' if all five words have been identified in a correct way). This requirement is important for test development in any new language since most of the optimization and evaluation steps depend on the definition of a threshold criterion. We therefore recommend a 80% word recognition threshold for future tests (though a 50% word recognition threshold criterion can also be used as long as this is explicitly stated).

We also recommend to report the standard errors on the SRT and slope estimates, as only then can a meaningful comparison across the psychometric curves of different tests be done. Note that the standard errors decrease and hence the accuracy increases as more trials are used (see Brand & Kollmeier, 2002), and so it may be necessary to run multiple lists in order to get a certain accuracy of SRT measurements.

Validation

Cross-validations of the new test with existing tests and with the results from other languages and laboratories is recommended, as this should help make the results across laboratories, clinics, and language regions as comparable as possible.

Acknowledgements

The work reviewed here has been supported by numerous projects (i.e. HearCom, HurDig, Hearing4all, and others) and funding institutions (EU, DFG, NIH/NIDCD and others) and has involved many individuals, including Sabine Hochmuth, Sofie Jansen, Heleen Luts, Astrid van Wieringen, Kirsten Wagener, Anna Warzybok, and Melanie Zokoll to whom the ICRA Working Group is particularly grateful. The copyrights of the tests we have developed are owned by the non-profit organization HörTech gGmbH (majority owned by Universität Oldenburg) as well as by the universities of Southampton, Leuven, Rotterdam, Linkoping, Free University Amsterdam, Academic Medical Centre Amsterdam, Karolinska Institute Stockholm, Sør-Trøndelag University College Norway, and other universities from the HearCom consortium. An increasing number of tests are also available commercially, e.g. as a medical product for modern audiometers. For research purposes, sample sentences and free trial versions of the research version of the software are also available from the copyright owners.

This set of recommendations will be made available to the scientific community via the ICRA website www.icra.nu, in conjunction with the up-to-date list of language-specific tests that fulfill the recommendations as well as the appropriate references.

Declaration of interest: Wouter Dreschler, Birger Kollmeier, Mark Lutman, and Jan Wouters are affiliated with public or non-profit institutions that own copyrights of the tests. Besides this, the authors declare no conflict of interest.

References

Bench J., Kowal Å. & Bamford J. 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Br J Audiol*, 13(3), 108–112.

Brand T. & Kollmeier B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests, *J Acoust Soc Am*, 111, 2801–2810.

Hagerman B. 1982. Sentences for testing speech intelligibility in noise. *Scand Audiol*, 11, 79–87.

Hagerman B. 1984. Clinical measurements of speech reception threshold in noise. *Scand Audiol*, 13, 57–63.

Hagerman B. & Kinnefors C. 1995. Efficient adaptive methods for measuring speech reception thresholds in quiet and in noise. *Scand Audiol*, 24, 71–77.

Hochmuth S., Jürgens T., Brand T. & Kollmeier B. 2015. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. *Int J Audiol*, submitted.

ISO 8253-3. 2012. Acoustics - Audiometric test methods - Part 3: Speech audiometry. Geneva: International Organization for Standardization.

Jansen S., Luts H., Dejonckere P., Van Wieringen A. & Wouters J. 2013. Efficient hearing screening in noise-exposed listeners using the Digit Triplet test. *Ear Hear*, 34(6), 773–8.

Jansen S., Luts H., Wagener K.C., Frachet B. & Wouters J. 2010. The French digit triplet test: A hearing screening tool for speech intelligibility in noise. *Int J Audiol*, 49, 378–87.

Kollmeier B. 1990. Messmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache (in German). (Methodology, modeling, and improvement of speech intelligibility measurements). Habilitation, Universität Göttingen, Germany.

Kollmeier B., Warzybok A., Hochmuth S., Zokoll M., Uslar V. et al. 2015. The multilingual matrix test: Principles, applications and comparison across languages: A review. *Int J Audiol.* online first, http://dx.doi.org/10.3109/14992027.2015.1020971

Kollmeier B. & Wesselkamp M. 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J Acoust Soc Am*, 102(4), 2412–2421.

MacPherson A. & Akeroyd M.A. 2014. Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends Hear*, 18.

Neumann K., Baumeister N., Baumann U., Sick U., Euler H.A. et al. 2012. Speech audiometry in quiet with the Oldenburg Sentence Test for Children. *Int J Audiol*, 51(3), 157–163.

Plomp R. & Mimpen A.M. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiol*, 18(1), 43–52.

Risberg A. 1976. Diagnostic rhyme test for speech audiometry with severely hard of hearing and profoundly deaf children. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 2, 3, 40–55.

Smits C. & Houtgast T. 2006. Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *J Acoust Soc Am*, 120, 1608–1621.

Smits C., Kapteyn T.S. & Houtgast T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *Int J Audiol*, 43(1), 15–28.

Soli S.D. & Wong L.L.N. 2008. Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int J Audiol*, 47, 356–361.

Sotscheck J. 1985. Sprachverständlichkeit bei additiven Störungen. *Acta Acustica united with Acustica*, 57(4–5), 257–267.

Voiers W.D. 1983. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4), 30–39.

von Wallenberg E.L. & Kollmeier B. 1989. Sprachverständlichkeitsmessungen für die Audiologie mit einem Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten. *Audiol Akustik*, 28, 50–65.

Wagener K. & Kollmeier B. 2005. Evaluation des Oldenburger Satztests mit Kindern und Oldenburger Kinder-Satztest. *Z Audiol*, 44(3), 134–143.

Wagener K., Brand T. & Kollmeier B. 1999a. Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil II: Optimierung des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test - Part II: Optimization of the Oldenburg sentence tests). *Z Audiol*, 38, 44–56.

Wagener K., Brand T. & Kollmeier B. 1999b. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test - Part III: Evaluation of the Oldenburg sentence test). *Z Audiol*, 38, 86–95.

Wagener K., Kühnel V. & Kollmeier B. 1999c. Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test - Part I: Design of the Oldenburg sentence test). *Z Audiol*, 38, 4–15.

Wichmann F.A. & Hill N.J. 2001a. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys*, 63, 1293–313.

Wichmann F.A. & Hill N.J. 2001b. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Percept Psychophys*, 63, 1314–29.

Zokoll M., Wagener K.C., Brand T., Buschermöhle M. & Kollmeier B. 2012. Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype. *Int J Audiol*, 51, 697–707.

Zokoll M.A., Hochmuth S., Warzybok A., Wagener K.C., Buschermöhle M. et al. 2013. Speech-in-noise tests for multilingual hearing screening and diagnostics. *Am J Audiol*, 22(1), 175–178.