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 Gagne, Mark Lutman, Jan Wouters, Lena Wong and Birger Kollmeier

Michael A. Akeroyd, Stig Arlinger, Ruth A. Bentler, Arthur Boothroyd, Norbert Dillier, Wouter A. Dreschler, Jean-Pierre Gagne, Mark Lutman, Jan Wouters, Lena Wong and Birger Kollmeier, International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests ICRA Working Group on Multilingual Speech Tests, 2015, International Journal of Audiology, (54), , 17-22.
International Journal of Audiology is available online at informaworld™:
http://dx.doi.org/10.3109/14992027.2015.1030513

ICRA recommendations for the construction of multilingual speech tests

ICRA working group on multilingual speech tests

Michael A. Akeroyd (MRC/CSO Institute of Hearing Research, Glasgow), Stig Arlinger (Linköping University), Ruth A. Bentler (Department of Communication Sciences & Disorders, The University of Iowa), Arthur Boothroyd (Distinguished Professor Emeritus, City University of New York), Norbert Dillier (Universitätsspital Zürich), Wouter A. Dreschler (Academical Medical Centre, Amsterdam), Jean-Pierre Gagné (Université de Montréal), Mark Lutman (Institute of Sound and Vibration Research, University of Southampton), Jan Wouters (Catholic University Leuven), Lena Wong (The University of Hong Kong), Birger Kollmeier (Universität Oldenburg and HörTech gGmbH, Oldenburg)

(Short communication)

An important goal in audiology is to establish internationally standardized diagnostic tests that can be applied and interpreted in the same way everywhere – an endeavor with severe limitations as soon as speaker and language-specific considerations come into play. Speech intelligibility tests with a closed-set design (i.e., a limited amount of response alternatives the test subject may choose from in order to identify which speech item she or he has heard) have the advantage that they can be performed in the native language of the respective test subject (if available) even if the test conductor does not understand this language. Such a format therefore supports speech audiometric testing in several languages which makes it attractive for a multilingual society. A prerequisite for a multilingual application, however, is that the construction and the respective results of these closed-set speech tests are as comparable across languages as possible. This calls for ICRA as an international body of leading experts in this field to produce guidelines concerned with the development of any new test of this kind in any new language. The current guidelines were therefore agreed on by an ICRA working group on multilingual speech tests constituted by the authors of this paper. The guidelines primarily concern the digit triplets test (as prototype for a screening test, see below) *and* the matrix test (as prototype for an audiologic diagnostic test for professional use, see below) but may also be used for other test designs. 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 new language.

The digit triplet test originally proposed by Smits and colleagues in Dutch (Smits et al.,2004) has meanwhile been produced in about 15 languages, such as, e.g., British English, American English, German, Spanish, Italian, Mandarin, Polish, Russian, and Spanish (see Zokoll et al 2012, 2013 for a comprehensive review). It has primarily been designed for use as a screening test using a telephone: A string of three digits is presented together with a background noise to the listener who has to press the appropriate digits on the telephone keypad as a response. An adaptive tracking algorithm allows altering the speech-to-noise ratio (SNR) according to the listener´s responses in such a way that the individual speech recognition threshold (SRT, i.e. SNR corresponding to 50 (or 80) percent speech intelligibility) is determined within a few trials. The matrix test – originally proposed by Hagerman (1982) in Swedish and in a modified version by Wagener et al (1999a,b,c) in German - is intended as an audiological diagnostic sentence recognition test consisting of syntactically fixed, but semantically unpredictable sentences like "Thomas buys eight red cars". For each of the five word positions (i.e., name, verb, numeral, adjective, and object), ten alternatives are available thus providing a "base matrix" of fifty words. Each sentence is a random walk through this matrix, and a complete test list of ten sentences contains each of these fifty words exactly once. The matrix test is by now available in

14 languages, including American English, British English, Dutch, German, French, Spanish, Turkish, and Russian (see Kollmeier at al 2014, for an extensive review).

In order to maintain maximum 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 should be as closely matched as possible. To enable the construction of such tests for any new languages in a way compatible to the existing corpus of tests, the current set of recommendations listed in Table 1 should be followed as closely as possible. These recommendations are based on the review 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. Even though not all existing tests exactly fulfill the specifications as indicated here, these recommendations are the results of a consensus process within the ICRA group. The recommendations for the digit triplet test and the matrix test listed in Table 1 should be self – explanatory in combination with the review papers by Zokoll et al (2012) and Kollmeier et al (2014), respectively. Nevertheless, the following remarks with respect to the different development steps are given:

General construction: For each new language, the construction of the digit triplet and the matrix test should be as close to the construction principle of the existing tests as possible. While the string of digits for most languages considered so far does not pose a specific problem, the sentence construction (order of the words) 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. Therefore 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 coarticulation effects at word transitions).

Word selection: For the digit triplets the number of syllables for the digits from zero to nine needs to be considered as it should be avoided that a certain digit can simply be recognized 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 the matrix test should also be usable for children (in its original form or in an abbreviated form with fewer words per sentence and fewer options per word group, see Wagener, 2005), the words employed should be as familiar as possible to the broad public, including children. The phoneme distribution of the underlying language should be approximated as closely as possible by the base matrix. However, the number of words and phonemes in the base matrix is usually large enough to maintain the approximation of the phoneme statistic of the respective language automatically. Hence, the alignment of the phoneme statistics primarily should ensure that major deviations in the phoneme distribution of the base matrix are avoided.

Speaker: A speech recognition test usually aims at assessing the individual listener's ability for everyday communication situations rather than specifically addressing hearing impaired listeners or an audience in front of a stage. Hence, the speaker to be selected for any new language should not necessarily be a formally trained speaker with any extreme speech quality, but rather a normally articulating, not necessarily formally trained, speaker with a dialect acceptable to the largest majority of the respective language users. Nevertheless, the speaker must be able to control his/her vocalization effort during the recording session of several hours. This should avoid a slow, but steady deterioration of the speaker´s voice quality during the recording. An RMS control 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 inevitably combines sentence portions from different parts of the recording session. These resynthesized sentences should eventually sound as natural as possible and should not contain any unnatural transitions in voice quality. While very few tests have been recorded with a male speaker (e.g. German, Polish) for compatibility reasons with other speech tests in the same language, in most languages a female speaker has been selected as an “acoustic compromise” between male speech and children´s speech. Hence the recommendation for any new language is to use a not specifically trained female speaker with a neutral accent who is able to keep her vocal effort constant for a recording session of a few hours.

Recording: ISO 8253-3:2012 provides an up-to-date description of the requirements for recording speech test materials which should be adhered to. In addition, it should be noted that the recorded speech elements will have to be segmented appropriately and recombined/resynthesized for producing the final test material which should sound as natural as possible.

For the digit triplets test, this involves the removal of the pauses between successive digits and the introduction of pauses with a fixed duration during the resynthesis procedure. Recording each digit separately for each of the three positions in the triplet achieves a natural prosody of the resynthesized material. Moreover, an announcement phrase at a slightly higher SNR helps to direct the subjects attention to the first digit presented (which may be of enhanced importance for low SNRs). Some language-specific versions may deviate from these recommendations. RMS equalization across recorded digits is not useful since the level varies considerably across digits even if they are adjusted to be equally intelligible.

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. That means that each of the 50 words has 10 realizations. The recorded sentences are cut into single words preserving coarticulation at the end of the cut word to the respective consecutive word but avoiding coarticulation at the word beginning. Various test lists of ten sentences are generated so that each test list contains all 50 words of the base matrix. Test sentences are resynthesised by combining words with appropriate transitions. Since the cutting and resynthesis of the recorded sentences is an effortful, language-dependent task, much work has to be invested into careful listening and quality control by native experts for the respective language. This is required in order to achieve high quality speech materials that are acceptable for both the patients and the 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), a masking noise is recommended here which is produced by a random superposition of all words of the respective corpus employing a random delay between successive repetitions of the speech items to be added. A high amount of superposition results in a quasi-stationary noise that has the same long term average spectrum as the target speech.

Optimization: The aim of the optimization is to achieve the highest possible homogeneity of the intelligibility across the speech material employed. Therefore, at first the word-specific discrimination functions have to be determined for all word realizations in the test with a group of normally hearing native subjects. Based on measurements at fixed SNRs covering a broad range of speech intelligibility for each word realization, an optimization of the speech test material can be achieved by attenuating words of high intelligibility and amplifying words of low intelligibility within a limited range. By shifting the respective discrimination function for each item as closely together as possible, the spread of word-specific SRT-values is decreased which increases the slope of the discrimination function for the test lists (according to the model by Kollmeier, 1990, reviewed by Kollmeier et al., 2014). Note that either a 50%- or 80%-definition of the SRT at word level may be employed for optimization, since, e.g., the triplet tests 80% point on a digit level is closer to the 50% point targeted for triplet-scoring in the final test (cf. Smits & Houtgast, 2006). Also note that separate optimization measurements and adjustments should be done for any alterations of the original speech material, e.g., telephone version (cf. Figure 1 in Jansen et al., 2010).

Evaluation: Evaluation of the tests should be done with an independent set of normally hearing listeners in order to assess the test list equivalence using the test lists generated within the optimization process described above, and to provide normative data for the respective language. Since the matrix test is known to have a significant training effect (e.g. Hagerman, 1984), this effect should be quantified and reported together with the test results. In order to obtain stable results, at least 2 lists of 20 sentences together 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 2 SNRs corresponding to so-called pair of compromise (i.e. 20% and 80% intelligibility), which allows an efficient simultaneous estimate of SRT and slope (Brand & Kollmeier, 2002). Fitting of a list-specific discrimination function can be also performed based on the work by Wichmann&Hill (2001a,b).

Obtaining normative values for any new language can be performed in a very similar way as for the existent tests in other languages. Note that most of the work using the matrix test has been reported on assuming fifty percent word intelligibility as the target probability. However, ISO 8523-3:2012 and clinical experience support the necessity to define the 80%-point on the discrimination function as the threshold criterion. This would roughly correspond to a 50%-criterion if sentence scoring would be applied (i.e. counting a response only as “correct” if all five words have been identified in a correct way.) This requirement is of strategic importance for the test development in any new language since most of the optimization and evaluation steps depend on the definition of a threshold criterion.

ICRA therefore recommends an 80 % word recognition threshold criterion as the preferred method, but a 50 % word recognition threshold criterion can also be used as long as this is explicitly stated. It is also recommended to report the standard errors on the SRT and slope estimates. Only then can a fair comparison across psychometric curves of different tests be done. Note that the standard errors decrease and hence the accuracy increases with increasing number of trials used for the estimate (see Brand & Kollmeier 2002).

Validation: A cross-validation with existing tests and with the results from other languages and laboratories is recommended with the aim to make the results across laboratories, clinics and language regions as comparable as possible.

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

Acknowledgement

Work supported by numerous projects (i.e., HearCom, HurDig, Hearing4all and others) and individuals (i.e. Melanie Zokoll, Anna Warzybok, Sabine Hochmuth et al.) whom the ICRA working group is highly indebted.

Table 1: Recommendations of the construction of multilingual speech tests

	Digit Triplet test	Matrix test
General construction	• Three consecutive digits between zero and nine in one utterance • Announcement phrase with increased level with respect to the triplets (up to 3 dB in order to be audible for hearing-impaired listeners) • 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 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)	• 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 (200-350 syllables per minute, depending on language, e.g. Russian 200 spm, Spanish 327 spm)	
Recording	• Equipment see standards partially implemented in the norm "Acoustics - Audiometric test methods - Part 3: Speech audiometry" (ISO 8253-3:2012)	
	• Each number at each position in the triplet to account for intonation aspects • Short pauses between digits	• 100 sentences accounting for co-articulation between words (each word recorded with each subsequent word)
Cutting	• Each digit at each position in the recorded triplet, omitting the pauses	• Preserve co-articulation at the end of the word
Resynthesis	• Composition of triplets regarding triplet position of the digit during the recording • add pauses (e.g., 160 ms) between successive, individually cut digits	• Generation of new test sentences by recombining words with appropriate co-articulation • Each test list contains 10 sentences including all 50 words of base matrix • Each word realization should occur equally often across all generated sentences, facilitating the determination of the word-specific discrimination function • Individual overlap-times of 0 to 300 ms • Testing the naturalness of sound of resynthesised sentences by native listeners

<table>
<tr><td>Masking noise</td><td colspan="2">• Preferred: Generated by multiple superposition of all sentences (or all digit triplets)
• If generation from other noise, the same long-term spectrum as the speech material has to be realized (use of appropriate filter)</td></tr>
<tr><td>Optimization</td><td colspan="2">• Determined by speech intelligibility measurements at fixed SNRs, covering range of 10% - 90% in speech intelligibility
• Noise level at 65 dB (55 – 75 dB is acceptable)
• Word scoring
• At least 2 lists of 20 sentences as training required per subject prior to data collection at an SNR of high intelligibility for Matrix test
• Level adjustment of each word realization to reach mean SRT (50% or 80% point, depending on target)
• Level corrections limited to ±2-4 dB (language specific-> requires assessment of adjustment limit with native listeners, Digit triplets test may use larger values)
• Words may be eliminated if 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</td></tr>
<tr><td>Evaluation 1:
Test list equivalence</td><td>• 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</td><td>• 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).</td></tr>
<tr><td>Evaluation 2:
Normative Data</td><td>• Adaptive 1up, 1 down method
  o fixed step size of 2 dB
  o An initial SNR at a high level of intelligibility should be chosen
  o SRT estimated by averaging the SNRs from 5th trial to the last trial (plus next “virtual“ SNR)
• Broadband
• Optional: Telephone version
  o Specification of Distortion & Band limitation/ Codec
  o Separate validation/ normative data
• Mean and Std. Deviation between Individuals and Test/Retest should be given</td><td>• Adaptive procedure with word scoring according to Brand & Kollmeier, 2002[*1)] using double test list (two lists of 10 sentences)
• Noise level fixed, Speech level varies
  o for SNR > 20 dB: change of noise level, speech level fixed
• Target probability 80% (word scoring)
• Optional 50 % target probability (word scoring)
• Separate normative data for open- and closed-set version
• Extent of training effect (using adaptive procedure) should be reported
• Mean and Std.-Deviation between Individuals and Test/Retest should be given</td></tr>
<tr><td>Validation</td><td colspan="2">• Multi-centre studies with normal-hearing and hearing-impaired listeners in comparison to reference tests (Country-dependent)</td></tr>
</table>

*1) Specification: Generalization of the procedure by Hagerman and Kinnefors (1995). 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

$$\Delta L = -\frac{f(i)\cdot(\text{prev}-\text{tar})}{\text{slope}}. \qquad (9)$$

Using the following settings, the original procedure by Hagerman & Kinnefors is obtained:

$f(i)=1,\ \text{tar}=0.4,\ \text{and}\ \text{slope}=0.2,$

The recommended settings proposed by Brand & Kollmeier (2002) for 50% are:

$f(i)=1.5\times 1.41^{-i}$, slope = $0.15\ \text{dB}^{-1}$

However, using this setting the step size can decrease too with increasing number of reversals *i*. Therefore, the authors recommend restricting the speed factor *f(i)* to a minimum value of 0.1.

References

Brand, T. & Kollmeier, B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests, Journal of the Acoustical Society of America, vol. 111, pp. 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. 2014. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. Submitted

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

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 and hearing34(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. International journal of audiology, 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 of Göttingen.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V., Brand, T. & Wagener, K.C. (under review) The multilingual matrix test: principles, applications and comparison across languages – a review. Under review in *Int J Audiol*

Smits C. & Houtgast T. 2006. Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. The Journal of the Acoustical Society of America, 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.

Wagener, C.K. 2005. Moderne Sprachverständlichkeitstests für Kinder. *8h congress of the German Society of Audiology*, Germany.

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. Perception & psychophysics, 63, 1293–313.

Wichmann F.A. & Hill N.J. 2001b. The psychometric function: II. Bootstrap-based confidence intervals and sampling. Perception & psychophysics, 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), 175-78.

Address list of the members of the ICRA working Group on multilingual speech tests (required for submission)

Michael A. Akeroyd (MRC/CSO Institute of Hearing Research- Scottish Section, New Lister Building (Level 3), Glasgow Royal Infirmary, 10-16 Alexandra Parade, Glasgow G31 2ER,United Kingdom, maa@ihr.gla.ac.uk),

Stig Arlinger (Linköping University, Department of Clinical and Experimental Medicine, Technical Audiology, S-581 85 Linköping, stig.arlinger@liu.se)

Ruth A. Bentler (Department of Communication Sciences & Disorders, The University of Iowa, Wendell Johnson Speech and Hearing Center, Iowa City, Iowa 52242, ruth-bentler@uiowa.edu)

Arthur Boothroyd (Distinguished Professor Emeritus, City University of New York, 2550 Brant Street, San Diego, CA 92101, aboothroyd@cox.net)

Norbert Dillier (Universitätsspital Zürich, Frauenklinikstr. 24. ORL-Klinik CH-8091 Zürich, norbert.dillier@usz.ch),

Wouter A. Dreschler (Academical Medical Centre, KNO-Audiologie, D2-240, Meibergdreef 9, NL- 1105 AZ Amsterdam, w.a.dreschler@amc.uva.nl),

Jean-Pierre Gagné (École d'orthophonie et d'audiologie, Centre de Recherche Institut Universitaire de Gériatrie de Montréal, Université de Montréal. Montréal, Québec, Canada; jean-pierre.gagne@umontreal.ca)

Mark Lutman (Institute of Sound and Vibration Research, University of Southampton, Highfield, Southampton, UK, mel@isvr.soton.ac.uk)

Jan Wouters (Catholic University Leuven, Research Group Experimental Oto-rhino-laryngology, O&N II Herestraat 49 - box 721, B-3000 Leuven, Jan.Wouters@med.kuleuven.be),

Lena Wong (Division of Speech & Hearing Science, University of Hong Kong, Pokfulam Road, Hong Kong, llnwong@hku.hk),

Birger Kollmeier (Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg and HörTech gGmbH, Marie-Curie-Str. 2, D-26129 Oldenburg, birger.kollmeier@uni-oldenburg.de)