Development of a Danish test material for assessing speech-in-noise reception in school-age children

SHNO KOIEK[1], JENS BO NIELSEN[2], LAILA KJÆRBÆK[3], MARIA BALTZER GORMSEN[4] AND TOBIAS NEHER[1,*]

[1] *Institute of Clinical Research, University of Southern Denmark, Odense, Denmark*

[2] *Hearing Systems, Technical University of Denmark, Lyngby, Denmark*

[3] *Department of Language and Communication, University of Southern Denmark, Odense, Denmark*

[4] *Department of Audiology, Odense University Hospital, Odense, Denmark*

For the audiological assessment of the speech-in-noise abilities of children with normal or impaired hearing, an appropriate test material is required. However, there is no standardized speech material for children in Denmark. The purpose of the current study was to develop a Danish sentence material suitable for school-age children. Based on the 600 test sentences from the Danish DAT corpus (Nielsen *et al*., 2014), 11 test lists comprising 20 sentences each were carefully compiled. These lists were evaluated in terms of their perceptual similarity and reliability with a group of 20 typically-developing normal-hearing children aged 6-12 yrs. Using stationary speech-shaped noise and diotic stimulus presentation, speech reception thresholds (SRTs) were measured twice per list and participant at two separate visits. The analyses showed that six test lists were perceptually equivalent. These lists are characterized by a grand average SRT of −2.5 dB SNR, a test-retest improvement of 0.4 dB, and a within-subject standard deviation of 1.1 dB SNR. The remaining test lists produced slightly higher SRTs but were generally also usable. Altogether, it is concluded that the developed test material is suited for assessing speech-in-noise reception in Danish school-age children.

INTRODUCTION

Children are often exposed to noise (e.g., in classrooms), which causes difficulties with speech understanding (e.g., Shield and Dockrell, 2003). Reliable methods for assessing speech reception in noise in school-age children are essential, specifically when difficulties in noise are suspected. In Germany, for instance, the "Oldenburger Kinder Satztest" (OlKiSa) was developed for that purpose (Neumann *et al*., 2012). OlKiSa consists of three-word pseudo-sentences, each with a numeral, an adjective, and an object noun (e.g., 'four red flowers') which is applicable for children from age four.

*Corresponding author: tneher@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

In Denmark, a number of speech materials are available for clinical and research purposes, e.g., DANTALE-I (Elberling *et al.*, 1989) or DANTALE-II (Wagener *et al*., 2003). DANTALE-I includes lists of monosyllabic words for the measurement of the discrimination score (DS) for children and younger children. DANTALE-II contains semantically unpredictable, nonsensical sentences that are difficult to memorize. However, significant learning effects have been observed (Wagener *et al*., 2003).

Nielsen and Dau (2009) developed a Danish speech intelligibility test named conversational language understanding evaluation (CLUE). This test was based on the principles and test procedure of the original Hearing in Noise Test (HINT; Nilsson *et al*., 1994). However, the speech materials of CLUE are not well-suited for children (Nielsen and Dau, 2011). In 2010, Nielsen and Dau developed a Danish version of HINT that was based on the same speech materials as CLUE with some modifications (Nielsen and Dau, 2011). Furthermore, Nielsen *et al.* have developed a Danish open-set speech corpus (DAT) containing 600 unique sentences that were systematically distributed in 30 test lists with three talkers (Nielsen *et al*., 2014). The DAT material was validated using free-field speech-on-speech measurements in normal-hearing and hearing-impaired Danish adult listeners. However, there is currently no standardized Danish speech test that is suited for testing speech reception in noise in children. An open-set speech material that simulates a real-life communication situation more than a close-set speech material is required to assess speech reception in noise abilities among children.

To summarize, there is no standardized Danish test material that is suited for children. The purpose of the current study was to address this shortcoming. In particular, the aim was to develop a set of test lists that is characterized by small training effects, high test list equivalence and low measurement uncertainty, which are suited for assessing speech reception in noise in Danish school-age children. To ensure reliable SRT measurements that are independent of the applied test list, these should result in very similar SRT measurements. In the current study, the reliability of the results was examined by a retest 5-15 days after the initial test.

MATERIALS AND METHODS

Generation of test lists

For the compilation of the test lists, the 600 test sentences from the Danish DAT corpus (Nielsen *et al*., 2014) were used. The DAT corpus is an open-set, low-context, multi-talker speech corpus. All sentences in this corpus have a fixed, simple structure. That is, they start with a name [Dagmar (D), Asta (A) or Tine (T)] and contain two short keywords (nouns), e.g., “Dagmar tænkte på en teske og en næse i går” (“Dagmar thought of a teaspoon and a nose yesterday”). In terms of their semantic properties, the noun pairs are not related, which makes them difficult to predict. For each name, there are 200 test sentences uttered by one of three professional female talkers with similar voice characteristics. For the current study, 220 sentences suitable for children with keywords judged to belong to the vocabulary of a typical 6-year-old were selected. For this selection, two audiologists and one psycholinguist (three of the

authors of the current study) individually went through all 600 sentences of DAT corpus. They each individually decided whether which sentences are suitable for a 6 years old child. Finally, they selected those 220 sentences that they all agreed being suitable for children and belong to the vocabulary of a typical 6 years old child. These sentences were combined into 11 lists containing 20 sentences each. All sentences in a given list are uttered by the same talker and start therefore with the same name. Specifically, four D-lists, three A-lists and four T-lists were created. The intelligibility was defined as the average SNR at which both keywords could be correctly identified by the 16 participating adults (Nielsen *et al*., 2014). In the present project, these intelligibilities were assumed to be valid for the 220 child-friendly sentences. In the children's lists, the sentences with relatively high and low intelligibilities were counterbalanced at the beginning of each test list, while the sentences with approximately equal intelligibility were put in towards the end of each list.

Participants

Twenty typically-developing, normal-hearing children (13 female) participated in the study. They were aged 6-12 yrs (mean: 8.7 yrs). All participants fulfilled the following inclusion criteria: (i) normal middle-ear function, (ii) pure-tone hearing thresholds ≤ 25 dB HL at all standard audiometric frequencies from 250 to 8000 Hz, (iii) native Danish speakers, (iv) normal language development, and (v) normal cognitive function. Middle-ear function and hearing thresholds were assessed using standard tympanometry and audiometry. Language development of the children was assessed using the Peabody Picture Vocabulary Test. Cognitive development was assessed based on parental reports.

Apparatus and procedures

All measurements were conducted in a soundproof booth. To evaluate the 11 created test lists in terms of their perceptual similarity and reliability, SRT measurements were made. The speech stimuli were presented diotically in stationary speech-shaped noise via supra-aural headphones (Sennheiser HDA200). The order of the test lists was balanced across the participants. The starting level of the speech signal was 67 dB SPL. The level of the noise was fixed at 60 dB SPL. The SRTs were measured using the adaptive procedure from the standard HINT (Nielsen *et al*., 2014). Before the start of the actual measurements, the participants were verbally instructed to repeat the two keys words in each sentence. In case of any doubts, they were encouraged to guess. Reponses were scored as correct if both keywords were repeated accurately. In this case, the level of the target speech was decreased by 2 dB. Otherwise, the level of target speech was increased by 1 dB. To familiarize them with the procedure and the speech material, all participants performed one SRT measurement in quiet and two SRT measurements in noise. The lists used for these purposes were training lists from the original DAT material (Nielsen *et al*., 2014). A short break was included after the first five SRT measurements and whenever a participant felt tired. A set of retest measurements was made on average 10 days (range: 5-19 days) after the first set of measurements.

Statistical analysis

The collected data were analyzed using SPSS version 25. To begin with, test-retest reliability was assessed, resulting in the data of one child being excluded from all subsequent analyses because of large inconsistencies. To assess the influence of the talker, visit and test list repeated-measures analyses of variance (ANOVAs) were performed on the SRTs. In all cases, a significance level of 5% was used.

RESULTS

The grand average SRT across all test lists and participants was −2.0 dB SNR with a standard deviation (SD) of 1.3 dB SNR. For the test measurements only, the mean SRT was −1.7 dB SNR with an SD of 1.5 dB SNR; for the retest measurements, the corresponding values were −2.4 and 1.4 dB SNR. The within-subject SD for all 11 test lists was 1.2 dB SNR. Figure 1 shows the mean list SRTs for the two visits.

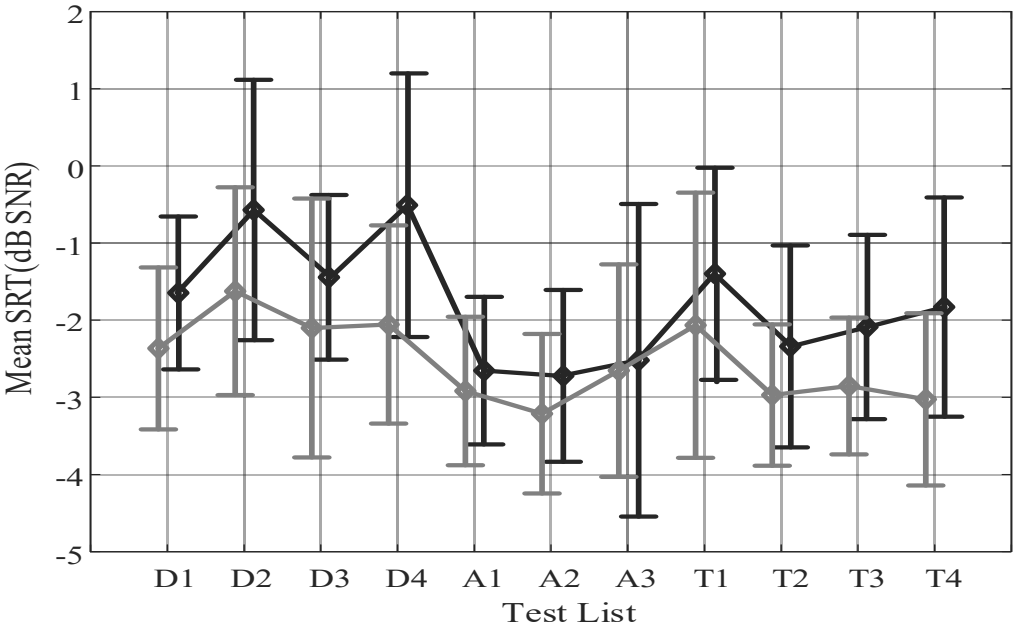

Fig. 1: Mean list SRTs for the first (solid line) and second (dashed line) visit.

Recall that the sentences of the D-, A- and T-lists were uttered by three different talkers. The overall mean SRTs of the three talkers were −1.4 dB SNR (D), −2.6 dB SNR (A), and −2.2 dB SNR (T), respectively. A one-way ANOVA comparing the mean SRTs of the three talkers showed a significant effect [$F_{(2, 206)} = 19.2, p < 0.001$]. Post hoc comparisons using Tukey's test revealed that the mean SRT of talker D was significantly higher than those of talkers A and T, whereas the mean SRTs of talkers A and T did not differ from each other (see Table 1).

Talker 1	Talker 2	Mean difference (dB)	*p*-value
T	A	0.4	0.126
D	T	0.8	< 0.001
	A	1.2	< 0.001

Table 1: Results of post hoc tests comparing the mean SRTs of talkers D, A and T.

To investigate the perceptual similarity of the seven test lists of talkers A and T, a two-way repeated-measures ANOVA with the within-subject factors visit and test list was carried out. This showed statistically significant effects of test list [$F_{(6, 108)} = 3.6$, $p = 0.002$] and visit [$F_{(1,18)} = 7.9$, $p = 0.012$]. Post-hoc comparisons using Tukey's test showed that the T1-list differed significantly from T2, T3 and A lists (all $p < 0.05$). T2, T3, T4, A1, A2, and A3 lists, on the other hand, did not differ from one another (all $p > 0.05$).

Figure 2 shows the mean SRTs of the test and retest of the 11 test lists. For eight of these test lists (T1, T2, T3, T4, A1, A2, A3, and D1), the mean SRTs were within 1 dB of each other. For the six lists that were found to be perceptually equivalent, the grand average SRT was – 2.5 dB SNR, the average test-retest improvement was 0.4 dB, and the within-subject SD was 1.1 dB SNR.

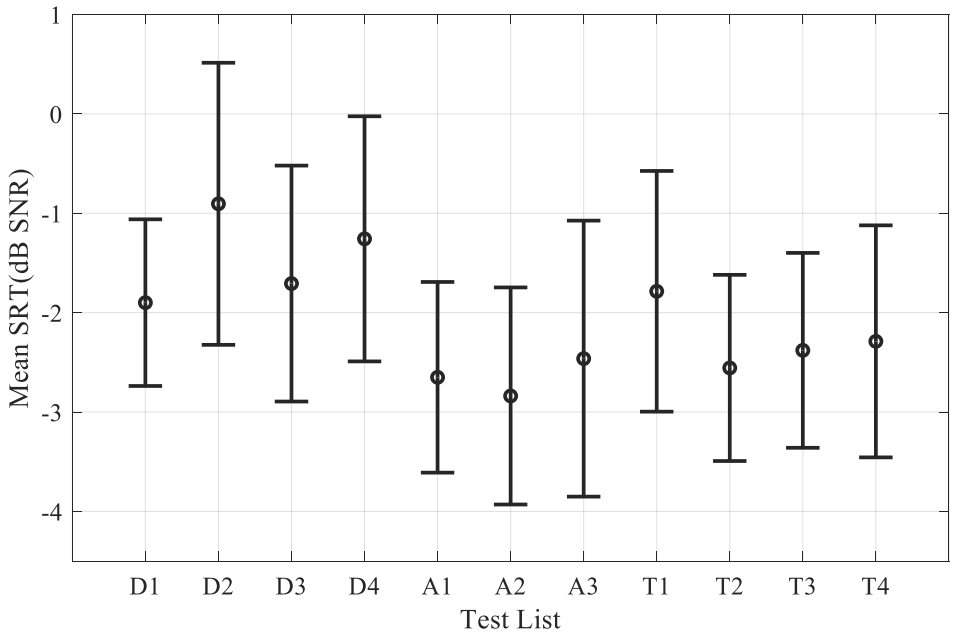

Fig. 2: A graph of mean SRTs and ±1 standard deviation for the 11 test lists.

Given that the children who participated in the current study covered a rather wide age span (6-12 yrs), we also investigated the effect of age on the SRT results. Figure 4 shows a scatter plot of age against the mean SRT. As expected, older children achieved lower (better) SRTs compared to younger children. The relationship between age and mean SRT was statistically significant ($r(19) = -0.53$, $p < 0.05$).

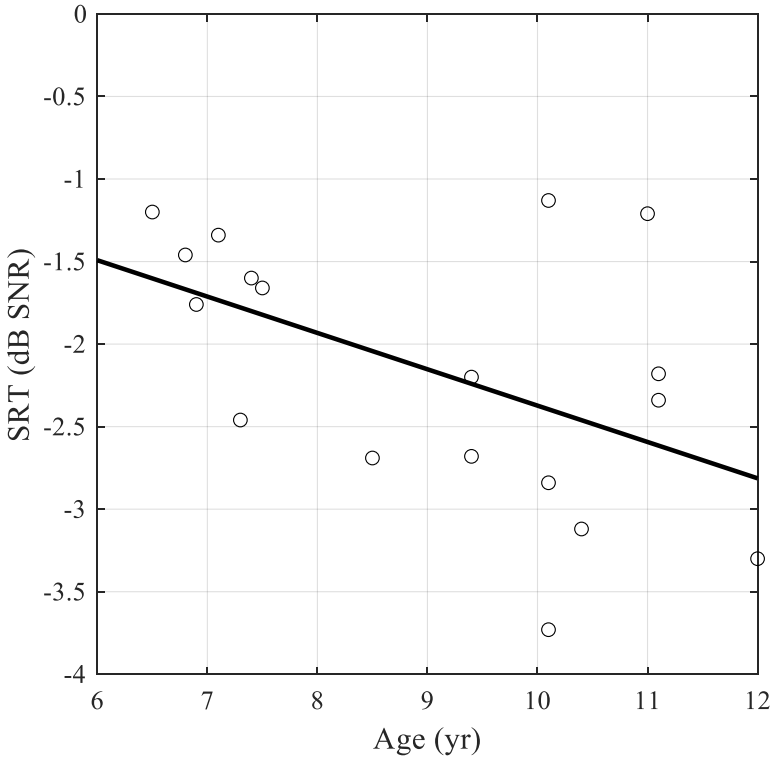

Fig. 3: Scatter plot of mean SRT versus age. The solid line shows a linear regression line that shows the average trend.

DISCUSSION

The aim of the current study was to develop a Danish test material, which is suitable for assessing speech reception in noise in school-age children. More specifically, the objective was to develop a set of test lists with small within-subject and between-list variation for performing SRT measurements with 6-12 yrs olds. Eleven test lists comprising 20 sentences each were created, and their equivalence was examined with the help of 20 typically-developing, normal-hearing native Danish children.

We verified our results by comparing them to the study by Nielsen and Dau with Danish HINT (Nielsen and Dau, 2011). However, Danish HINT has been studied among adult listeners. But in both studies, SRTs have been assessed using sentence reception in stationary speech-shaped noise. The overall mean SRT of the 11 lists in the current was −2.0 dB SNR, which is slightly higher than the mean SRT of the Danish HINT obtained with 16 normal-hearing adults (−2.5 dB SNR; Nielsen and

Dau, 2011). The within-subject SD for these lists was 1.2 dB SNR, which is somewhat larger than that of the Danish HINT (0.9 dB SNR). The relatively higher mean SRT and within-subject SD might be due to differences in the speech material used in the two studies. DAT sentences are without context whereas the HINT sentences are everyday sentences with contextual information. Another explanation could be the large age difference between the participants of the two studies (children vs. adults).

Since the sentences of the D-, A- and T-lists were uttered by three different talkers, we considered the influence of talker on our results. We found that the D-lists resulted in significantly higher mean SRTs than the A- and T-lists. Ideally, the test lists of a given speech material should result in very similar SRT measurements, so they can be used interchangeably. Based on our results, lists T2, T3, T4, A1, A2 and A3 are equivalent in terms of mean SRT. These test lists can be used in actual measurements in future speech-based research in 6-12 years old children in Denmark. The rest of the test lists including D1, D3, D4, and T1 were found with equivalent mean SRTs. They can be used as training lists in measurements. They can also be used in tests with less conditions of measurements. The mean test-retest improvement for these lists was 0.4 dB, corresponding to that observed for the Danish HINT (Nielsen and Dau, 2011). The within-subject SD across six test lists was 1.1 dB, which is slightly larger than the within-subject SD found by Nielsen and Dau (2011) with normal-hearing Danish adults (0.9 dB SNR). This can also be explained by speech material differences (without context vs with context) as well as age difference (children vs adults).

In addition to the within-subject SD, we also calculated the list-SRT SDs for lists T2, T3, T4, A1, A2 and A3. The result was 0.2 dB SNR. This result is very similar to the list-SRT SD of the Danish HINT (0.3 dB SNR). Furthermore, the maximum deviation from the overall mean SRT of 0.3 dB SNR was observed for list A2. It is smaller than the maximum deviation from the overall mean SRT that Nielsen and Dau (2011) found (0.6 dB SNR). This indicates a high equivalence of these six test lists with respect to the measured SRT.

CONCLUSION

Eleven test lists compiled from the Danish DAT corpus (Nielsen *et al*., 2014) were evaluated in terms of their perceptual similarity and reliability with 20 native Danish 6-12-year-old children. Six of these test lists (T2, T3, T4, A1, A2 and A3) were found to be suitable for speech-based studies among Danish 6-12-year-olds. These lists produced a grand average SRT of −2.5 dB SNR. The observed test-retest improvement of 0.4 dB, which suggests that reusing the lists after about 10 days is possible. The A- and T-lists produced mean SRTs that were within 1 dB of each other. The D-lists resulted in mean SRTs that were on average 1 dB higher than the other lists. They may be used for training purposes.

REFERENCES

Elberling, C., Ludvigsen, C., and Lyregaard, P. E. (**1989**). "DANTALE: A new Danish speech material," Scand. Audiol., **18**(3), 169-175. doi: 10.3109/01050398909070742

Nielsen, J., Dau, T., and Neher, T. (**2014**). "A Danish open-set speech corpus for competing-speech studies," J. Acoust. Soc. Am., **135** (1), 407-420. doi: 10.1121/1.4835935

Nielsen, J. B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**(10), 729-741. doi: 10.1080/14992020903019312

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol, **50**(3), 202-208. doi: 10.3109/14992027.2010.524254

Nilsson, M., Soli, S.D., and Sullivan, J.A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., **95** (2), 1085-1099. doi: 10.1121/1.408469

Neumann, K., Baumeister, N., Baumann, U., Sick, U., Euler, H. A., and Weißgerber, T. (**2012**). "Speech audiometry in quiet with the Oldenburg Sentence Test for Children," Int. J. Audiol., **51** (3), 157-163. doi: 10.3109/14992027.2011.633935

Shield, B. M., and Dockrell, J. E. (**2003**). "The effects of noise on children at school: a review," Build. Acoust., **10**(2), 97-116. doi: 10.1260/135101003768965960

Wagener, K., Josvassen, J.L., and Ardenkjær, R. (**2003**) "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol., **42** (1),10-17. doi: 10.3109/14992020309056080