LICOS: TEST-RETEST RELIABILITY AND ADAPTIVE PROCEDURES

Objectives. The Linguistic Controlled Sentences (LiCoS; Coene et al, 2018) were further explored on reliability and applicability in an adaptive procedure. **Design, Study and Sample**. Test and retest data from the LiCoS in silence and noise were collected of thirteen normal hearing adults via fixed and adaptive procedures. Different measures of reliability were performed on the SRT's of the obtained psychometric functions. **Results.** The SRT's of the LiCoS in silence (25.79 ± 2.34dB (test) and 26.44 ± 2.63dB (retest)) did not differ significantly and showed excellent reliability based on the interclass correlation coefficient (ICC=0.93) and the within-subject standard deviation (0.89dB). In speech noise the SRT's of the LiCoS were significantly lower in retest (-3.67 ± 0.91dB) than in test condition (-2.99 ± 0 .74dB) with an ICC of 0.45 and a small within-subject standard deviation (0.79dB). The SRT's of the adaptive procedure did not differ significantly from the fixed procedure. **Conclusion.** The reliability of the LiCoS in silence and noise is in line with other speech audiometric test materials. The adaptive method is an accurate and quick procedure to obtain SRT's. Within-subject standard deviations of 1-2dB should be taken into account when comparing results from different moments in time.

Keywords: Speech audiometry; adaptive procedure; test retest reliability; psychoacoustics; Dutch sentences

Introduction

Recently two new sets of speech materials were developed for the evaluation of speech understanding in Dutch (Flanders, Belgium and the Netherlands): the Linguistically Controlled Sentences (LiCoS; Coene et al., 2018). The LiCoS sentence test consists of 12 lists of 30 sentences. It differs from the current existing sentence tests for the Dutch speaking area (the 'Leuven Intelligibility Sentence Test' (LIST) (van Wieringen and Wouters, 2008) and the 'VU sentence test' (Plomp and Mimpen, 1979)) in two important aspects. First, the sentences are controlled for both acoustic and linguistic parameters. Second, the test tries to estimate the hearing performance in everyday listening situations by using linguistically complex sentences at conversational rates. This latter aspect is frequently referred to as ecological validity, which has been the area of interest in the development of new hearing test batteries for the clinical audiological practice.

For the same reason, an increasing number of hearing test batteries include speech-in-noise tests (Nilsson et al., 1994, Taylor, 2003). Speech-in-noise tests provide meaningful information typically related to real life hearing challenges. In real life, background noise may adversely affect speech intelligibility. This adverse effect is related to several factors: the hearing status of the listener (i.e. audibility and spectro-temporal auditory resolution), the ability to benefit from perceptual closure (i.e. the ability to form linguistic wholes from perceived fragments), and other non-auditory factors (i.e. language proficiency and cognitive reserves). The LiCoS in noise takes the interplay of these factors into account and will therefore attempt to accurately represent speech understanding of the listener in daily situations.

Measures of speech intelligibility are commonly plotted in psychometric functions. The points of interest of the function are the slope of the function and the speech reception threshold (Strasburger, 2001). In audiometric practice, the slope of the function is commonly described as the increment of proportion correct per unit of stimulus increase (1 dB). Typical slope values range from 0.05 to 0.2 dB^{-1} (Brand and Kollmeier, 2002). The speech reception threshold (SRT) is defined as the point of inflection of the function, which corresponds with the point of maximum slope (Strasburger, 2001). In the audiometric practice this point refers to the speech level corresponding with 50% speech understanding in silence (in dB_{SPL}) or in noise (in dB SNR). Normative SRT values for the LIST are 27.1 dB_{SPL} (SD = 0.9 dB_{SPL}) for the LIST in silence and -7.8 dB SNR (SD = 0.2 dB) for the LIST in noise (van Wieringen and Wouters, 2008). Normative SRT values for the LiCoS Dutch and LiCoS Flemish are respectively 26.8 dB_{SPL} (SD = 2.3 dB_{SPL}) and 25.8 dB_{SPL} (SD = 2.2 dB_{SPL}) for the LiCoS in silence, with a mean slope of respectively 9.9 (SD = 1.6) and 10.2 (SD = 2.3) $\%dB^{-1}$. For the LiCoS in noise SRT's of -5.1 dB SNT (SD = 0.7 dB) and -2.8 dB SNR (SD = 0.7 dB) were found and slopes of 9.9 (SD = 1.0) and 12.9 (SD = 3.3) $\%dB^{-1}$ (Coene et al., submitted, this journal). The difference in SRT's in noise between the LIST and the LiCoS is explained by the higher speech rate and the linguistic complexity of the LiCoS. In previous research our group has shown that this delicate balance between bottom-up and top-down processing may explain why the SRT's in noise obtained through LiCoS are less favorable than those obtained through other test batteries. When listeners are not able to rely on linguistic knowledge, they are highly dependent on the peripheral auditory system to process the incoming speech signals. As such, the LiCoS outcomes may be taken to represent auditory performance to the highest possible degree (see Coene, Krijger, et al. 2016).

In the current study we report the results of a reliability and efficacy analysis of the LiCoS before implementing the new test in clinical practice. In clinical practice speech understanding tests are frequently performed to compare two or more conditions (e.g. different type of hearing aids, different fitting adjustments) on subsequent moments

in time. For this purpose the speech tests require to have reliable results with low within-subject variability. However speech reception scores are subject to different sources of variation. The variation can originate from the speech material, the scoring method or patient related factors (Boothroyd, 1968). For this reason, it is important to perform measures of variability. In one of our previous studies we have shown how variation across the 12 LiCoS sentence lists is controlled for (internal validation, see Coene, Krijger et al, submitted, this journal), but not for measures across subsequent moments in time (reliability). In this paper, we report the results of a study regarding the efficacy of the test procedure and we describe the development of an adaptive test procedure for the LiCoS sentences. One of the main reasons for the development of such an adaptive test procedure is that it can reduce the administration time of the speech test significantly. Moreover, via an adaptive procedure one SRT is obtained which is not subject to ceiling and floor effects. These two advantages are both interesting for clinical practice.

The goal of the present study is to further explore the LiCoS in silence and in noise on its test retest reliability and the efficacy and accuracy of the LiCoS in an adaptive test procedure.

Method

Subjects

Thirty normally hearing Flemish subjects participated in the initial normative study of the LiCoS (Coene et al., 2018). Of these subjects 13 returned for a second retest evaluation after 1 or 2 weeks. The retest was administered in the same test center as the inital study: the Eargroup (Antwerp, Belgium). All subjects were between 19 and 31 years of age (mean = 25.7 ; SD = 2.5 years) and were screened for normal hearing (< 20 dBHL for octave frequencies 250-8000 Hz).

Procedure: fixed and adaptive

The tests were performed in a sound treated room with one speaker positioned in front of the subject (0° Azimuth, 1m distance). LiCoS sentences were played from the A§E software (Otoconsult nv, Antwerp, Belgium) at intensities ranging from 16 to 48 dBSPL and signal to noise ratios from -12 to 12 dB SNR (see below for detailed procedure). The intelligibility was assessed based on two key words in each sentence.

The LiCoS sentences were administered in silence, in noise (speech weighted noise, SWN) and in three different adaptive procedures (silence, noise and babble noise). The type of presentation (silence, noise, adaptive) was randomized as well as the selection of the 12 different lists of the LiCoS. The listeners were instructed to repeat the sentences and were encouraged to guess if they had missed a word or part of a sentence.

Fixed procedure

The presentation levels for the LiCoS sentences were alternated between 40 and 42 dBSPL for the LiCoS in silence and between 8 and 10 dB SNR for the LiCoS in noise in the initial study (Coene et al., submitted, this journal). If the score at the initial presentation level was less than 100%, the initial level was increased in 4 dB increments until 100% score was obtained and a stepwise decrement of 4 dB started from that level downwards until a score of 0% was obtained. Of these data points a 50% SRT point was calculated using a non linear regression fit to a logistic function (see Coene et al, submitted, this journal). The alternating presentation levels resulted in psychometric functions of 2 dB resolution. At the second test moment (T2) a second psychometric

response curve was established by performing the procedure at a same presentation level than the intitial one (T1).

Adaptive procedure

Adaptive procedures allow a quick estimation of a predetermined point on the psychometric function. The procedure gives an estimation of the intensity of the sentence at which the subject can obtain a particular proportion of correct responses (e.g. 50% SRT). This technique differs from the standard fixed procedure because the intensity of the signal varries in function of the response on the previous sentence. By adapting the signal intensity with decreasing stepsizes, the 50% SRT is achieved. The adaptive procedure used in the current study followed a simple staircase paradigm with a varying stepsize. The starting intensity was fixed at $70dB_{SPL}$. The initial step size was 10dB and decreased during the remainder of the test with factor $2^{reversals}$. As stop criterium 8 reversals was chosen and the 50% threshold estimation was based on the average of the last 6 reversals.

Slope estimation of the function

The steepness of the slope was calculated from individual psychometric curves of each listener based on the regression fits to the logisctic model function. The steepness of the slope is expressed as % per dB.

Test Retest Reliability

Three measures of test-retest variability were calculated for each of the LiCoS in silence, the LiCoS in noise and for the three adaptive procedures (in silence, in speech noise and in babble noise):

(i) the standard error from multiple measures (Lafon, 1965).

(ii) the within-subject standard deviation of repeated measures (Plomp and Mimpen, 1979). The within-subject SD was defined by calculating the root mean square of the differences between the two SRT's of each participant in each condition divided by √2. This method was adapted by Vaillancourt et al (2005) who defined the within-subject standard deviation with an unambiguous formula:

$$\sigma_\omega = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{i,j}-\mu_i)^2}{n(k-1)}} \qquad (1)$$

Where $x_{i,j}$ is the j^{th} treshold of the i^{th} subject, μ_i stand for the mean of the tresholds of i^{th} subject, n is the number of subjects and k the nummer of trials. Furthermore Vaillancourt untangled this within-subject standard deviation in two sources of errors: the constant error due to systematic errors (such as learning effects) and the variable error (due to e.g. list choise, attention,…). The constant error was estimated by the mean difference between T1 and T2 whereas the variable error was calculated by substracting the estimated constant error from the total error (Vaillancourt et al., 2005).

(iii) the interclass correlation coëfficient (ICC), which compares the between subject variance with the within-subject variance. The ICC for absolute agreement (ICC_{agree}) was calculated via a 2 way mixed model for both average as single measures. The calculated SRT's from the fixed procedure represent an average of two measures ($ICC_{average}$), whereas the SRTs obtained from the adaptive procedure are single measures (ICC_{single}). ICC values below 0.4 stand for a poor reliability, ICC values between 0.4 and 0.75 represent a fair to good reliability and ICC values higher than 0.75 represent an excellent

reliability (Jansen et al., 2014; Vanspauwen, Wuyts, Krijger et al., 2017). Furthermore, paired student's t-tests were performed to compare mean SRT's from T1 and T2.

Accuracy and efficacy adaptive procedure

The accuracy of the adaptive procedure was analysed by comparing the calculated SRT's from the fixed procedure to the SRT's from the adaptive procedure on T1 (n=30) by performing paired student's t-tests.
Furthermore, the test efficacy was addressed by registering the test time and the amount of sentences used to obtain the SRT.

Statistical Analyses

Statistic analyses were performed with IBM SPSS statistics version 22.0 (SPSS Inc., Chicago, Illinois). The data were checked for a normal distribution by means of QQ plots and Shapiro Wilk test. Results were considered to be statistically significant for p values less than 0.05.

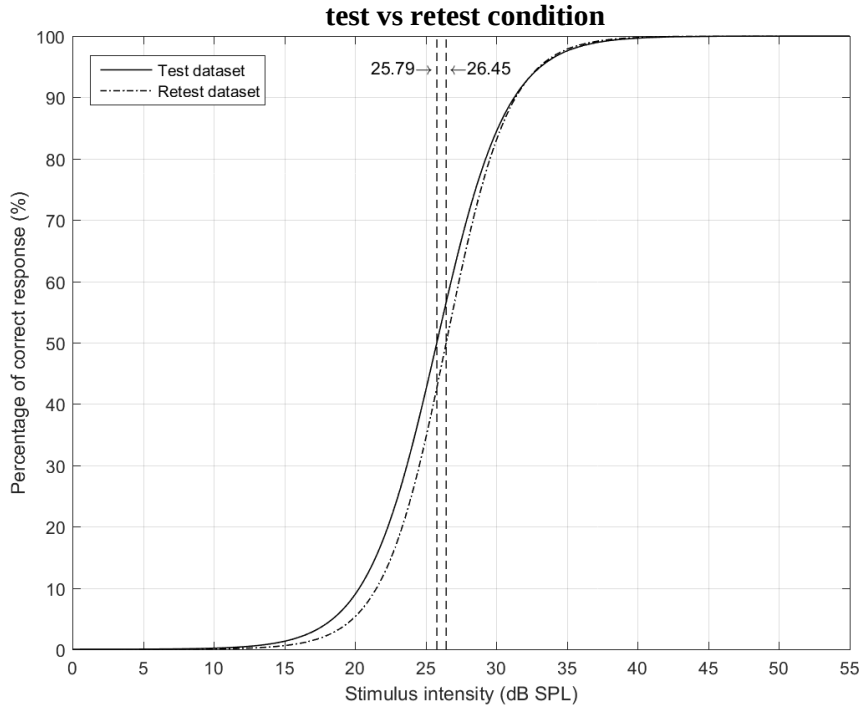
Results

Test retest reliability

Fixed procedure

Data of T1 and T2 of the fixed procedure from the LiCoS in silence and noise were plotted in Figure 1.

a. Non-linear curve fitting of Speech Reception Scores of LiCoS silence: test vs retest condition



b. Non-linear curve fitting of Speech Reception Scores of LiCoS noise: test vs retest condition

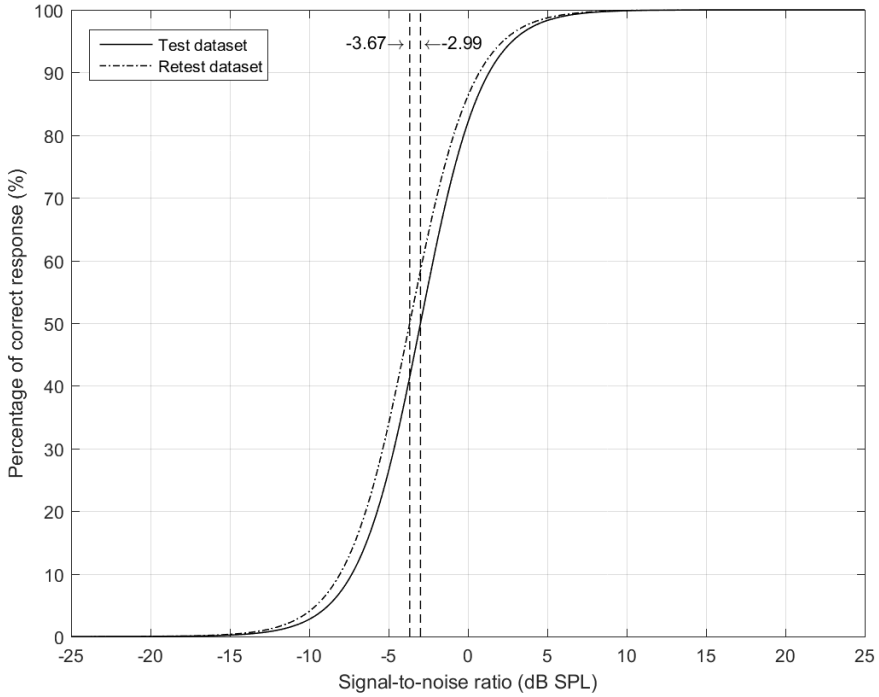


Figure 1. Speech receptions scores for the LiCoS in silence (a) and the LiCoS in noise (b) obtained at test and retest condition.

Of these data 50% SRT's were obtained via a non lineair regression fit. The interpolated SRT points for T1 and T2 are summarized in Table 1, together with the standard deviations. The slope at the 50% SRT was calculated based on the individual data resulting in a mean slope of 10.02 (SD = 2.45) $\%dB^{-1}$ at T1 and 11.17 (SD = 1.42) $\%dB^{-1}$ at T2 for LiCoS in silence and 12.73 (SD = 2.40; T1) $\%dB^{-1}$ and 12.53 (SD = 2.47; T2) $\%dB^{-1}$ for the LiCoS in noise.

	Test	**Mean SRT (dB)**	**SD (dB)**	**Within-subject SD of RM**	**SEM of RM**	**ICC**
LiCoS silence	T1	25.79	2.34	0.89	0.32	0.93
LiCoS silence	T2	26.45	2.63			
LiCoS noise	T1	-2.99	0.74	0.79	0.26	0.45
LiCoS noise	T2	-3.67	0.91			

Table 1. Mean SRT's and standard deviations (SD) from the LiCoS in silence and the LiCoS in noise at T1 and T2 (n=12). Additionally the following measures of reliability are shown: the mean absolute difference between T1 and T2, the within subject standard deviations (SD) of repeated measures (RM), the standard error (SEM) of repeated measures (RM) and the interclass correlation coefficient (ICC) based on absolute agreement of average measures.

The interpolated SRT's were both normally distributed for T1 and T2. A paired t test comparison showed no significant difference between the SRT's of the LiCoS in silence obtained at T1 and T2 ($t(12) = -2.088$; $p = .059$). For the LiCoS in noise the SRT's of T1 and T2 differed significantly ($t(12) = 2.650$; $p = .021$) with better SRT scores at T2 suggesting a learning effect when administrating the LiCoS in noise.

Additionally the ICCs (measures of absolute agreement based on average measures) were calculated via a 2 way mixed model ANOVA which confirmed the results

above, by demonstrating an excellent reliability for the LiCoS in silence (ICC = 0.93) and a borderline fair reliabilty (ICC = 0.45) for the LiCoS in noise.

The mean differences between T1 and T2 were calculated for the LiCoS in silence and the LiCoS in noise, resulting in mean absolute differences of 1.00 (SD = 0.83) and 1.01 (SD = 0.51) dB. The within-subject standard deviations (Vaillancourt 2005) of individual measurements were 0.89 dB and 0.79 dB for respectively LiCoS in silence and in noise.

Based on the suggestions of Vaillancourt (2005) these within-subject standard deviations could be further classified in constant errors (= mean difference) of -0.66 (SD = 1.14) and -0.68 (SD = 0.92) dB and estimated variable errors of 0.23 and 0.11 for LiCoS in silence en LiCoS in noise respectively.

Adaptive procedure

SRT's for LiCoS in silence, LiCoS in noise and LiCoS in babble noise were obtained via the adaptive procedure on T1 and T2 (n = 13). An overview of the mean SRT's, and standard deviations can be found in Table 2. In addition the measures of reliability (the standard error, the within-subject standard deviation of repeated measues and the ICC on absolute agreement) were added in the adjacent columns.

	Test	**Mean SRT (dB)**	**SD (dB)**	**Within-subject SD of RM**	**SEM of RM**	**ICC_{agree}**
LiCoS silence	T1	25.98	3.42	1.79	0.72	0.72
LiCoS silence	T2	25.68	3.38			
LiCoS noise	T1	-1.91	1.70	1.52	0.59	0.40
LiCoS noise	T2	-2.62	2.15			
LiCoS babble noise	T1	-4.06	2.87	1.83	0.74	0.47
LiCoS babble noise	T2	-4.36	2.21			

Table 2. SRT's obtained with the adaptive procedure at T1 and T2 for LiCoS in silence, LiCoS in noise and LiCoS in babble noise (n=13). Standard deviations and the mean absolute difference between T1 and T2 were tabulated together with the within-subject standard deviation of repeated measures (RM), standard error for repeated measures (SEM) and the interclass correlation coefficient based on absolute agreement of single measures.

The paired t test revealed no significant differences between T1 and T2 for LiCoS in silence, noise and babble noise (resp. t(12) = .416, p = .685; t(12) = 1.217, p = .247; t(12) = 0.402, p=.695).

The SRT scores differed on average 0.30 (SD = 2.61) and 0.71 (SD = 2.11) dB and 0.30 (SD = 2.68) for the LiCoS in silence, the LiCoS in noise and the LiCoS in babble noise. These values correspond to the constant error. By extracting these values from the wihtin-subject standard deviations variable errors of 1.49, 0.81 and 1.53 were estimated for the LiCoS in silence, LiCoS in noise and LiCoS in babble noise.

The absolute SRT differences of T1 and T2 were on average 2.09 (SD = 1.49), 1.71 (SD = 1.36) and 1.76 (SD = 1.98) dB for LiCoS in silence, LiCoS in noise and LiCoS in babble noise.

Accuracy adaptive procedures

The accuracy of thresholds obtained using the adaptive methods (silence and noise) versus the fixed standard procedure was determined by the intertest differences of the obtained SRT's at T1 (n=30).

Paired t tests on the data obtained at T1 showed no significant difference between the adaptive procedure compared to the fixed procedure for both LiCoS in silence and LiCoS in noise (resp. $t(29) = .460$, $p = .649$; $t(29) = 1.010$, $p = .321$). The mean difference between the two tests was 0.16 (SD = 1.90) for LiCoS in silence and 0.29 (SD = 1.58) for LiCoS in noise. The mean within-subject differences were higher when considering the absolute values, which was 1.60 (SD = 0.98) for the LiCoS in silence and 1.28 (SD = 0.94) for the LiCoS in noise.

The within-subject standard deviation was 1.32 (silence) and 1.12 (noise) and a standard error of resp. 0.35 and 0.29 was calculated for the repeated measures.

Efficacy adaptive procedure versus fixed procedure

The total average administration time was registered for the sentences of the LiCoS in silence and the LiCoS in noise (Table 3). In the fixed procedure 5 to 8 lists of 30 sentences were used to obtain results from 0% to 100% speech intelligibility with a 4dB resolution. In the adaptive procedure less than one list (<1) was necessary to obtain a SRT with a stop criterium of 8 reversals. Since the difference in administration time between the two methods is evident, the administration time was calculated for one list based on the total administration time, which allows for further comparison.

	Sentences	**Lists**	**Total administration time and standard deviations (minutes:seconds)**	**Estimated administration time calculated for 1 list (minutes:seconds)**
LiCoS fixed silence	201.0 ± 27.5	6.7	20:02 ± 00:25	2:59
LiCoS fixed noise	181.7 ± 45.7	6.1	28:57 ± 1:07	4:46
LiCoS adaptive silence	20.4 ± 4.6	<1	2:50 ± 1:25	2:05
LiCoS adaptive noise	18.6 ± 4.1	<1	2:29 ± 0:33	4:00
LiCoS adaptive babble	19.8 ± 5.3	< 1	3:00 ± 1:09	4:33

Table 3. Efficacy of the administration of the LiCoS sentences in silence and the LiCoS in noise in adaptive and fixed procedure. For each measurement the amount of sentences and lists are shown, together with the total administration time and estimated administration time for 1 list.

Discussion

The aim of the current study was to further explore the recently developed speech lists of the LiCoS in silence and noise. The standard fixed procedure of the test was investigated on its test retest reliability. In addition an adaptive procedure was developed for LiCoS in silence, noise and babble noise. This procedure was evaluated on its accuracy, efficacy and test retest reliability compared to the standard fixed procedure.

Data obtained from the LiCoS in silence from 13 normal hearing adults on two subsequent moments in time were analysed and two SRT's were calculated via a non linear regression fit (25.79 ± 2.34 dB at T1 and 26.44 ± 2.63 dB at T2). The SRT's from T1 and T2 did not differ significantly and showed excellent reliability based on the calculated interclass correlation coëfficient (ICC = 0.93). This high reliability was also reflected in the small within-subject standard deviation of 0.9 dB, which is in line with other speech lists: 0.9 for the LIST sentences in silence in fixed procedure (van Wieringen and Wouters, 2008) and 1.1 for the Plomp sentences (Plomp and Mimpen, 1979). The within-subject standard deviation of the LiCoS in silence is smaller than the within-subject standard deviation found for the American English HINT (1.39; (Nilsson et al., 1994) and for the Canadian French HINT (2.2; (Vaillancourt et al., 2005)).

The differences in the within-subject standard deviations between the tests reported in literature can be explained by several factors: the difference in number of participants, the used test protocol, the time interval between the test and retest condition, the use of a practice list and the content of the speech material itself (see Appendix 1 for an overview). In literature a within-subject standard deviation of 1 or 2 dB is considered small, but in clinical practice such a small difference is often incorrectly considered to be significant. For this reason the use of confidence intervals is recommended for the interpretation of results of all tests of speech audiometry, especially when comparing results of different moments in time.

The SRT's from the LiCoS in noise differed significantly when administered on two different moments in time. The calculated SRT's obtained at T2 were lower (-3.67 ± 0.91 dB SNR) than the SRT's at T1 (-2.99 ± 0.74), This upward trend was also reported in the Swedish HINT wherein a systematic increase of 0.77 dB was found when the retest was performed in a one week time interval. This could suggest that there might be a learning effect (i.e. when the listeners recall the sentences) or habituation effect (i.e. when the listeners get more acquainted with the test) (Hallgren et al., 2006). A learning effect is less apparent in this test procedure because the significant increase only appears in the noisy condition. Noisy conditions tax the cognitive processes, which reduce the cognitive reserves needed to achieve learning in se. Moreover, the sentences have low redundancy and were developed to have a low learning effect in all conditions. These explanations suggest rather a habituation effect than a learning effect. A habituation effect occurs when the listener learns to supress the background noise. The fact that such habituation effect does not seem to occur in existing test batteries such as LIST may be related to the fact that the latter used a double retest procedure: first listeners received a practice list through which they were able to tune in to the noisy listening condition, followed by the test-retest procedure itself. In our test procedure, the test-retest analysis is based on two lists without prior listening-in-noise practice moment. In clinical practice the administration of a practice list might be encouraged before administrating the LiCoS in noise.

Consistently, the LiCoS in noise demonstrated only a fair reliability (ICC = 0.45). The standard deviations and standard errors of the LiCoS in noise were nonetheless smaller than the LiCoS in silence, which expresses less variability. Data with low variability and a small range of results, will inevitably lead to low ICC values and might therefore not be eligible as measure to compare test retest data of speech in noise tests.

The low within-subject standard deviation (0.8) is, however, comparable with the data published for several speech in noise tests: the Plomp sentences in noise (0.9; (Plomp and Mimpen, 1979)), and its revised version (1.07; VU sentences in noise (Versfeld et al., 2000)), the Canadian French HINT (1.1; (Vaillancourt et al., 2005) and the American English HINT (1.13; (Nilsson et al., 1994). The within-subject standard deviation of the LiCoS in noise is higher than the within-subject standard deviation of the LIST in Noise in fixed procedure (0.2; (van Wieringen and Wouters, 2008). An important reason for the difference between the within-subject standard deviation of the LIST and the LiCoS in noise could be found in the speech material. The sentences of the LIST are more redundant compared to the LiCoS which results in a steeper slope and less variation round the 50% SRT point (Lyregaard, 1987; Hammer, Coene & Govaerts, 2013).

The estimated slopes for the psychometric functions for both LiCoS in silence as LiCoS in noise are consistent with the slopes found in literature (see Appendix 1). Compared to the LIST in silence and the LIST in noise, the slopes are shallower for both the LiCoS in silence (10.02 vs. 12.5% dB^{-1}) as the LiCoS in noise (12.73 vs. 17.5 $\%dB^{-1}$). The fact that the slopes from the psychometric functions obtained from the LiCoS data are shallower than the slopes from the LIST data emphasizes once again the differences in the speech materials used in the LIST (well articulated speech material with higher

redundancy, 'clear speech') versus the LiCoS (complex linguistic sentences at a conversational speaker rate).

In addition to the exploration of the standard fixed procedure of the LiCoS test an adaptive procedure was introduced to increase the efficacy of the test procedure in clinical practice. The adaptive procedure was administered in silence, noise and babble noise in test and retest condition. The obtained SRT's revealed no significant differences between the test and retest data.

The data from the adaptive procedure for the LiCoS in silence showed an excellent reliability (ICC = 0.72), whereas the LiCoS in noise and babble noise showed borderline poor reliability. This difference is consistent with the difference found in the standard procedure and can therefore also be explained by the low subject variability and limited range of results in the speech in noise data.

The within-subject standard deviations were higher for all three conditions in the adaptive procedure compared to the fixed procedure, but were still in the same order of magnitude of reported results in literature (0.2-2.2, cfr supra). When dichotomizing the errors in a constant error and a variable error, the results showed overall equal or lower constant errors for the adaptive procedure compared to the standard procedure. The constant error, which is mainly due to a learning or habituation effect, is the lowest in the silence and babble noise condition. In the adaptive procedures less sentences (18 to 20) were used compared to the standard fixed procedure (201 for LiCoS in silence and 181.7 for LiCoS in noise), wherefore less habituation effect could be expected in the adaptive procedure. On the contrary the variable errors, originating from the list choice or attentional processes of the subject, were higher in the adaptive procedures, with the highest variable error found in the LiCoS in silence and in babble noise. As opposed to the continuous noise condition, listeners are in the babble noise condition able to benefit from glimpsing or dip listening (i.e. the benefit from acoustical cues obtained in the in momentary reductions of the fluctuating masker noise; (Howard-Jones and Rosen, 1993)). The amount of glimpsing is dependent on several factors (such as the available acoustic cues in the dips and the listeners ability to do dip listening) and will involve an extra factor of variability, hence the higher variable error. The theory of glimpsing was further substantiated by the better SRT's found in babble noise compared to the SRT's found in the noise conditions.

Overall, the reliability measures calculated from the adaptive procedures were lower than the measures from the fixed procedure. This finding was expected, since the data obtained from the adaptive procedure resides from less measure points (sentences) compared to the standard fixed procedure. Nevertheless the data obtained from the adaptive procedure score high on measures of accuracy. Data obtained from the LiCoS in silence and in noise showed no significant differences with the data obtained through the adaptive procedure, resulting in within-subject differences (1.32 dB (silence) and 1.12 (noise)). Furthermore, the administration time of the LiCoS in silence en the LiCoS in noise was less than 3 minutes, whereas a minimum of two lists are needed in the standard fixed procedure to obtain a SRT, which takes on average 6 to almost 10 minutes. These findings demonstrate that the adaptive method for the LiCoS is a valid method to obtain a quick and accurate estimation of the SRT, at least if the variable errors are taken into account.

Noteworthy, is the upward trend in the SRT's from the LiCoS in noise and the LiCoS in babble noise obtained at T1 and T2 in the adaptive procedure, which also occurred in the fixed procedure. This trend could therefore suggest an identical, but not significant, habituation effect.

In the present study different measures of reliability were suggested based on the measures used in literature. Although the within-subject standard deviation seems to be the most common method to check for reliability in the field op psychoacoustics, it seems to be less sensitive to detect significant differences when comparing mean values. Therefore, it is important to perform different measures and to know the limits of variability of each measurement. This is best expressed by 95% confidence intervals of SRT's, which should always be respected when interpreting the results of speech audiometry. The ICC seems not to be an eligible measure to investigate the reliability speech perception in noise. The ICC is the ratio of the inter-subject variability and the between subject variability and may result in low values due to low between-subject variability in the obtained data. This is the case with the speech perception thresholds in noise, for which within-subject standard deviations and difference tests are recommended.

Conclusion

The sentences of the LiCoS seem to be complementary to the existing speech materials in the Flemish area of Belgium. Next to the added value in terms of content (see Coene et al., 2018), it now has been substantiated that the reliability of the LiCoS is in line with the other speech materials and could therefore be implemented in clinical practice.
An adaptive procedure was developed for the LiCoS sentences in order to increase the test efficacy for clinical use. The SRT's obtained from the adaptive procedures in both silence as noise (SWN and babble) condition also showed high reliability and accuracy when comparing the SRT's to the standard fixed procedure. The adaptive procedure is subject to more variability due to less measure points, especially when administered in adaptive procedure with babble noise in which dip listening may occur. Although the errors for the adaptive procedure in LiCoS in silence and noise are within the same order of magnitude of other speech lists, it is definitely recommended to use of 95% confidence intervals when interpreting the obtained SRT results. Furthermore, the implementation of a practice list was advised before administrating the LiCoS in noise to let the listener habituate to the noisy test conditions.

Funding

Our research has received funding from the Research Foundation Flanders, FWO Vlaanderen, Belgium (PhD. Fellowship, first author), the Netherlands Organisation for Scientific Research, within the NWO KIEM funding scheme (second author) and the European Union's Seventh Framework Program for research, technological development and demonstration under FP7-PEOPLE-2012-IAPP project "Hearing Minds", Grant Agreement No 324401 (last author).

References

BOOTHROYD, A. 1968. “Developments in Speech Audiometry”. *International Audiology,* 7**,** 368-368.

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,* 111**,** 2801-2810.

COENE, M., KRIJGER, S., MEEUWS, M., DE CEULAER, G. & GOVAERTS, P. J. 2016. Linguistic Factors Influencing Speech Audiometric Assessment. *Biomed Res Int,* 2016**,** 7249848.

COENE, M., KRIJGER, S., VAN KNIJF, E., MEEUWS, M., DE CEULAER, G. & GOVAERTS, P. J. J. (2018). Development and validation of a norm-referenced Linguistically Controlled Sentences (LiCoS) test for Dutch and Flemish speech audiometry *Folia Phoniatr Logop* 70, 90-99.

HALLGREN, M., LARSBY, B. & ARLINGER, S. 2006. A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *International Journal of Audiology,* 45**,** 227-237.

HOWARD-JONES, P. A. & ROSEN, S. 1993. Uncomodulated glimpsing in "checkerboard" noise. *J Acoust Soc Am,* 93**,** 2915-22.

JANSEN, S., KONING R. WOUTERS, J., & VAN WIERINGEN, A. (2014). Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m). *International Audiology, 53(1)*(55-9).

WOUTERS, J. & VAN WIERINGEN, A. 2014. Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m). *International Audiology,* 53(1).

LAFON, J. P. M., A.; GAUTHIER, J. 1965. L’intervalle de confiance des mesures audiométriquesvocales. *International Audiology,* 4**,** 94-96.

NILSSON, M., SOLI, S. D. & 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**,** 1085-99.

PLOMP, R. & MIMPEN, A. M. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiology,* 18**,** 43-52.

STRASBURGER, H. 2001. Converting between measures of slope of the psychometric function. *Perception & Psychophysics,* 63**,** 1348-1355.

TAYLOR, B. 2003. Speech-in-noise tests: How and why to include them in your basic test battery. *The Hearing Journal,* 56**,** 40,42-46.

VAILLANCOURT, H., LAROCHE, C., MAYER, C., BASQUE, C., NALI, M., ERIKS-BROPHY, A., SOLI, S. D. & GIGUERE, C. 2005. Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology,* 44**,** 358-369.

VAN WIERINGEN, A. & WOUTERS, J. 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol,* 47**,** 348-55.

VANSPAUWEN, R., WUYTS, F. L., KRIJGER, S. & MAES, L. K. 2017. Comparison of Different Electrode Configurations for the oVEMP With Bone-Conducted Vibration. *Ear and Hearing,* Publish Ahead of Print.

VERSFELD, N. J., DAALDER, L., FESTEN, J. M. & HOUTGAST, T. 2000. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am,* 107**,** 1671-84.

Appendix 1.

Reference	**Speech material**	**Language**	**Speaker**	**Retest condition**			**Mean SRT and SD** **(dB; dBSPL)**	**Mean slope and SD** **(%dB^{-1})**	**Within-subject SD**
				#pp	time	practice			
Plomp & Mimpen, 1979	Plomp sentences	Dutch	m	10			S: /	S: /	S: /
					<1d	x	Na: -5.9 ± 0.9	Na: 20	Na: 0.9
Versveld et al., 2000	VU sentences	Dutch	f/m	12			S: /	S: /	S: /
					<1d	x	Na: -3.1	Na: 11.7	Na: 1.1
Wouters & Wieringen, 2008	LIST-f sentences	Flemish	f	7	<1d	x	S: 27.1	S: 12.5	S: 0.9
					<1d	x	N: -7.8 Na: -7.8	N: 17.8	N: 0.2 Na: 1.2
Jansen et al., 2014	LIST-m sentences	Flemish	m	6	<1d	x	S: 21.1 ± 2.5 Sa: 21.1 ± 2.5	S: 12.3 ± 2.2	Q: 0.2 Qa: 1.7
					<1d	x	N: -7.8 ± 0.4 Na: -7.8 ± 0.4	N: 18.7± 1.8	N: 0.2 Na: 1.1
Nilsson et al., 1994	HINT senteces	American English	m	18	<1d	x	Sa: 23.9±3.45 dBA	S: /	Sa: 1.39
					<1d	x	Na: -2.9 ± 0.78	Na: 9.7	Na: 1.13
Vaillancourt et al., 2005	HINT sentences	French Canadian	m	36	<1d	x	S: 16.4	S: /	S: 2 2
					<1d	x	N: -3.3 ± 0.5	N: 10.3	N: 1.1
Hällgren et al., 2006	HINT senetences	Swedisch	f	10			S: /	S: /	S: /
					<1d	x	Na: – 3.0 ± 1.1	Na: 17.9	Na: 0.9
					7d		Na: -3.8		Na: 1.4

Summary of different factors contributing to the variability in reported within-subject standard deviation of different speech lists. For each speech list following factors are summarized: the language, the speaker's voice (female (f), male (m) or both (f/m)), three features of the retest condition: the number of participants (#pp); the retest time interval (time; if the retest was performed on the same day (<1d)) and the use of a practice list (x), the mean SRT and SD for silence (S) and noise (N) (if available) together with the used test procedure (fixed (S,A) or adaptive (Sa,Na)), the mean slope, the within-subject standard deviation and the retest time interval (Time; if the retest was performed on the same day (<1d)).