Clinical feasibility of the test battery

Introduction

Intraclass correlation (ICC) is a way of measuring the reliability of a measurement method (Goldsmith & Stratford, 1997). A higher ICC value (ranges between 0 - 1) indicates a higher correlation between the test and retest measurement, which is an indication of higher reliability. It is important to state that there is no standard value for acceptable reliability, and a low ICC could not only reflect the low degree of measurement agreement but also relate to the lack of variability among the sampled subjects. Koo and Li (2016) suggest that an ICC value of less than 0.5 is an indication of poor reliability. Values varying between 0.5 and 0.75 indicated a mediate reliability, and an ICC value of above 0.75 indicates good reliability. An ICC value of above 0.9 indicated excellent reliability of the measurement method. Both the Pearson's r and the ICC can give misleading results, as they are very sensitive to the spread of data between subjects. Therefore Downham et. al. (2005) suggest to also investigate if there is a systematic bias of the data and a measurement error. In this study, the test-retest reliability of the test battery has been explored by using the intraclass correlation (Koo & Li, 2016), Pearson's correlation, systematic change in means (Downham et al., 2005) and standard error of measurements (SEM; Goldsmith & Stratford, 1997).

Methods

Participants

Test-retest measurements were performed in seven HI and three NH people for all tests of the test battery. The average PTA of the HI listeners was 31 dB HL, and all with bilateral HL. The retest session (second set of two visits) was measured 4 months after the first visit. The same type of equipment was used in both sessions, but four HI listeners were measured at a different location for the first visit. There were two different examiners for the first test session. The same person examined all the subject in the second session.

Measures of reliability

Interclass cross-correlation calculation has different forms based on different assumptions. By looking at the flowchart presented in the paper by Koo and Li (2016), a two-way mixed effect model was chosen for the analysis. The absolute agreement was chosen, as Koo and Li (2016) state that the test-retest reliability study would be meaningless if there was no agreement between repeated measures. Depending on the test, either a single measurement or the mean of k measurements was used as the type of measurement. The equation below shows the formula for the two-way mixed model, with absolute agreement.

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n(MS_C - MS_E)}} \tag{1}$$

where MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square for error; MS_C = mean square for columns; n = number of subjects and k = number of raters/measurements.

Pearson's r is another way of investigating the reliability and gives very similar results to the ICC. While the ICC is looking at the distance of the point from a straight line that is going through the origin, Pearson's r is looking at the distance from any kind of linear line (Koo & Li, 2016).

For investigating if there is a systematic bias of the data and a measurement error, both the mean difference in results between the two sessions and the 95% confidence interval is calculated for all tests. If the mean difference ($\bar{d}$) had a negative value, this indicates that the results from the first session tend to be larger than the second one. If the confidence interval is including zero, it can be concluded that there is no systematic bias between the two sessions.

The standard error of measurement (SEM) is also calculated. SEM is a way of calculating measurement error (Goldsmith & Stratford, 1997) and is a way to compare different measurement methods. Because it is in the same units as the original measure, also the SEM is calculated in percentage to compare different measurement method with different units.

$$SEM = \sigma_T \sqrt{(1 - ICC)} \qquad (2)$$

where σ_T is the total sample standard deviation, and ICC is the ICC value shown in equation 2.

$$SEM\ \% = \frac{SEM}{\bar{m}} \times 100$$

where $\bar{m}$ is the mean of all measurements.

Results and discussion

Table I: The ICC and Pearson's r values for all tests of the test battery

Test	*Condition*	*ICC*	*R*	*Systematic change*	*SEM (%)*
WRS	*10dB*	*ICC = 0.591, p = 0.001*	*0.59*	*d = 0.04, CI = [-0.04,0.11]*	*0.13 (23.05)*
	20dB	*ICC = 0.291, p = 0.096*	*0.28*	*d = -0.007, CI = [-0.06,0.04]*	*0.078 (9.84)*

Test	Condition	ICC	R	Systematic change	SEM (%)
	30dB	ICC = 0.251, p = 0.128	0.25	d = -0.005, CI = [-0.02,0.01]	0.03 (3.16)
	40dB	ICC = 0.475, p = 0.011	0.48	d = -0.009, CI = [-0.03, 0.01]	0.04 (4.29)
HINT	**SRT**	**ICC = 0.611, p = 0.001**	0.60	d = 0.17, CI = [-0.46,0.79]	1.02 (211.54)
	SS+4dB SNR	ICC = 0.574, p = 0.002	0.58	d = -2.96, CI = [-7.73,1.82]	7.94 (9.56)
STM	LF	**ICC = 0.916, p = 0.00**	0.85	d = -0.09 CI = [-0.84, 0.67]	0.93 (12.26)
	HF	ICC = 0.548, p = 0.003	0.59	d = 0.51, CI = [-0.27, 1.29]	1.31 (37.4)
ACALOS	**HTL**	**ICC = 0.946, p = 0.000**	0.95	d = -0.13, CI = [-1.27, 1.00]	4.59 (17.53)
	MCL	ICC = 0.678, p = 0.000	0.68	d = 0.53, CI = [-1.10, 2.16]	6.59 (7.86)
	Slope	**ICC = 0.821, p = 0.000**	0.82	d = -0.002, CI = [-0.02, 0.02]	0.07 (15.51)
Binaural Pitch	**Dichotic**	**ICC = 0.987**, p = 0.000	0.99	d = -2, CI = [-5.54, 1.54]	3.99 (4.91)
	Total	ICC = 0.983, p =0.000	0.99	d = -0.5, CI = [-2.61, 1.61]	2.27 (2.52)
Frequency tracking procedures	IPD_{fmax}	ICC = 0.950, p = 0.000	0.96	d = -15.84, CI = [-66.44, 34.75]	65.39 (6.37)
	eAUD-HF (FLFT)	ICC = 0.890, p = 0.000	0.89	d = 212.71, CI = [-89.7, 515.1]	495.3 ()
eAUD-B	Bo	ICC = 0.327, p = 0.101	0.41	d = 1.99, CI = [0.24, 3.74]	2.28 (3.24)
	Bp	ICC = 0.673, p = 0.007	0.70	d = 1.10, CI = [-1.62, 3.83]	3.1 (5.48)
	BMR	**ICC = 0.783, p = 0.002**	0.77	d = 0.89, CI = [-1.08, 2.86]	2.25 (16.2)
eAUD-N	LF	ICC = 0.325, p = 0.05	0.40	d = 1.27, CI = [0.09, 2.44]	2.02 (2.87)
	HF	ICC = 0.551, p = 0.005	0.54	d = 0.29, CI = [-0.99, 1.56]	2.11 (2.89)
eAUD-S	**S: LF**	**ICC = 0.851**, p = 0.00	0.85	d = -0.36, CI = [-1.45, 0.73]	1.78 (3.34)
	S: HF	ICC = 0.954, p = 0.000	0.95	d = 0.51, CI = [-0.66, 1.69]	1.92 (4.08)
	SMR:LF	ICC = 0.651, p = 0.004	0.68	d = 1.48, CI = [0.04, 2.91]	2.47 (14.24)
	SMR: HF	**ICC = 0.858, p = 0.000**	0.85	d = -0.32, CI = [-2.09, 1.46]	2.85 (11.19)

Test	***Condition***	***ICC***	***R***	***Systematic change***	***SEM (%)***
eAUD-T	*T: LF*	*ICC = 0.665, p = 0.002*	*0.77*	*d = 1.35, CI = [0.49, 2.21]*	*1.64 (2.59)*
	T: HF	*ICC = 0.875, p = 0.000*	*0.89*	*d = -0.96, CI = [-2.00, 0.09]*	*1.78 (2.88)*
	TMR: LF	*ICC = 0.192, p = 0.205*	*0.19*	*d = -0.06, CI = [-1.40, 1.27]*	*2.17 (30.24)*
	TMR: HF	*ICC = 0.668, p = 0.003*	*0.71*	*d = 1.13, CI = [-0.38, 2.63]*	*2.54 (23.96)*

The test-retest reliability of the test battery has been investigated, looking at the ICC, Pearson's R, systematic changes in the data and the SEM. Some of the tests, such as IPD, Binaural Pitch and FLFT showed a good to excellent test-retest reliability with all ICC values above 0.89. There was also no indication of a systematic bias and the SEM showed low values that were below 7% of the total mean for each test. The ACALOS outcome measures also showed good reliability with ICC ranging from 0.67 to 0.95 (ICC_{HTL} = 0.95, ICC_{MCL} = 0.67, ICC_{Slope} = 0.82). There was no indication of a systematic change in the data, and the SEM values for both, the HTL and MCL, varied around 5 dB, which was the same as the uncertainty in the one expected in pure-tone audiometry.

For the WRS test, lower ICC values were found, indicating poorer reliability. This could be a result of the participants having the alternatives visually in front of them, and even though they could not hear the word, they chose the word closest to it. The mean difference between test and retest for the 30 dB and 40 dB conditions is 4%, which is only one difference incorrect words (1/25). A mediate reliability was shown for both outcomes of the HINT measurements, with ICC varying between 0.57 and 0.61. One reason for this mediate reliability could be the choice of lists and lack of randomization in the first session. As mentioned in the main document, there was a small list effect between the two ears that can also play a role here. There was no indication of any systematic changes in the data, and the SEM values were relatively small, which indicates good reliability.

The STM measurements showed an excellent reliability for the LF condition (ICC = 0.91) and a mediate reliability for the HF condition (ICC = 0.548). In addition, the SEM values showed better reliability for the LF condition. A reason for this could be that many subjects could not simply detect any modulation for the HF condition and therefore answered randomly. A poor to mediate reliability is shown for each condition of the binaural extended audiometry (eAUD-B; ICC_{S0N0} = 0.327, ICC_{SpiN0} = 0.673, ICC_{BMR} = 0.783). Diotic condition (S_0N_0) showed the lowest ICC value and the lowest spread of the data. However, there was also a shift in the data, with higher values for the first session. For the $S_{pi}N_0$ condition and the BMR, there was no shift in the data. The S_0N_0 was used for calculating the BMR, and reliability can be questionable.

The TiN condition of the extended audiometry (eAUD-N) showed a poor to mediate reliability (ICC_{LF} = 0.325, ICC_{HF} = 0.551). The results of the LF condition also showed a shift towards higher values for the first session. The TiN part of the extended audiometry was used for calculating both the SMR and the TMR which is crucial to understand the following results. The temporal condition showed mediate reliability, and a systematic change showing higher values for the first session. The spectral condition of the extended audiometry (eAUD-S) showed results that are promising for its implementation in the clinics with a good to excellent reliability for both conditions (ICC_{LF} = 0.851, ICC_{HF} = 0.954), however, the reliability of the spectral masking release was lower. Moreover, all conditions of the extended audiometry show somehow the same standard error of measurements (SEM), around 2 dB, which is the same as the minimum step size.

Time efficiency of the test battery

The examiners kept track of the time used by each of the participants in completing the test battery. In the case of some events, for example additional repetitions needed, annotated the events cautiously for later investigation. Regarding the test procedure in Appendix A, additional repetitions of the threshold estimations were need if: 1) a repetition was considered as an outlier if a given threshold was greater than three scaled median absolute deviations of the three repetitions; or 2) the responses of the listeners during the tracking procedure were inconsistent or reached the maximum or minimum possible values. In that case, the measurement was considered an invalid or "missing" data point.

The results of the timing are shown in Figure 5, probability and mean number of extra repetitions per listener are shown in Table 2.

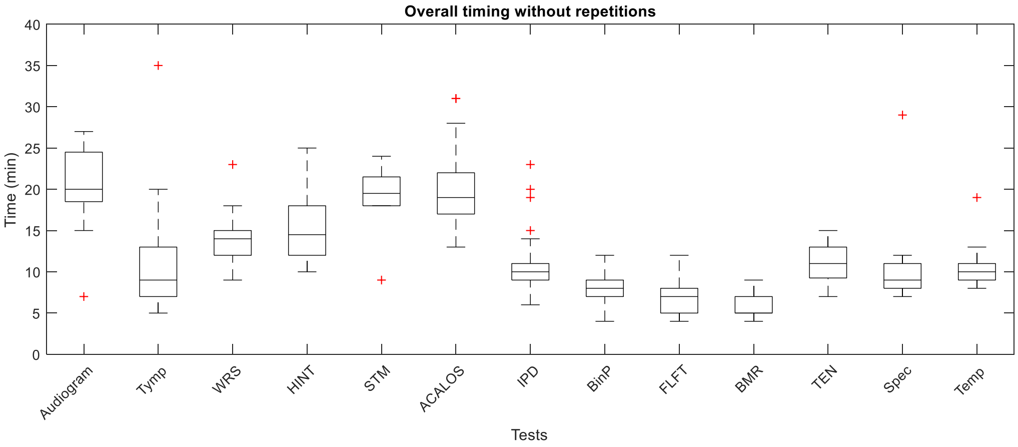

Figure 1: The overall time of the different tests in the test battery.

Table 2: Table with the probability of needing repetitions, and the probability of having missing values (unreliable data). The rows with no values in the probability of missing values column is due to the test not producing this missing values. The mean number of extra repetitions is only averaging over the number of extra repetitions.

Test	*Probability of extra repetitions (%)*	*Probability of missing values (%)*	*Total probability of having to repeat (%)*	*Mean number of extra repetitions*
HINT	*1.32*	-	*1.32*	*4 (only one subject)*
STM	*42.86*	*90.79*	*88.16*	*4.32*
IPD	*10.77*	*10.97*	*20.55*	*1.87*
Binaural Pitch	*8.11*	-	*8.11*	*0.167*
FLFT	*5.63*	*4.05*	*9.46*	*1.85*
S_0N_0	*42.59*	*27.03*	*58.11*	*2*
$S_{pi}N_0$	*20.59*	*9.11*	*27.03*	*1.85*
eAUD-N	*66.67*	*46.58*	*82.19*	*3*
eAUD-S	*48.57*	*52.70*	*75.68*	*3.07*
eAUD-T	*53.85*	*46.58*	*75.34*	*3.27*

Tests such as WRS, HINT, ACALOS, Binaural Pitch, and FLFT showed a low number of total repetitions needed. This can be partly explained by the procedure used. Looking at the timing of these tests, ACALOS and HINT had a larger spread, meaning that the estimated test-time was more subject-dependent than in the case of other tests.

The timing data of the STM tests includes also the STM screening in the beginning of the test. The probability of having to repeat the measurement was 42.86 %. Looking at the missing values, 90.79% of the subjects had missing values giving 88.16 % probability of having to repeat at least one condition. Many listeners had substantial difficulties for detecting the STM at HF condition what was unexpected and the examiners performed several extra repetitions before giving up. For the last 30 listeners, if the threshold was “missing” after more than 2 extra repetitions, the result was consider “missing”. Furthermore, STM was measured using the SIAM tracking procedure which has catch trials and a conservative criterion for giving a certain threshold as valid. If the listener missed many catch trials or if the responses are done too quickly, there is a high risk of having a no valid (missing) estimate. For many listeners that was the case for the HF condition but not the LF condition. The number of repetitions needed was largely spread for the STM test, ranging from 1 - 14. The average number of extra repetitions were around four. This supports the idea of modifying the tracking procedure of this test for further investigations.

The IPD_{fmax} showed a low variation of the timing data with some outlier. This can be due to extra systematic training provided before the test. The probability of repetition was in total 20.55% meaning that one in five subjects had to repeat either as a result of the two repetitions not being close enough or invalid threshold estimates (missing values).

The binaural part of the extended audiometry are a relatively short measurement with a total time around 5 min. Looking at the probability of repetition, both parts of the binaural extended audiometry shows a large probability of repetitions. The S_0N_0 part showed 58% total probability of repetitions, meaning that more than half the subjects needed to repeat. The duration of the eAUD-N, eAUD-S and eAUD-T, without any extra repetitions, was around 10 minutes each. However, looking at the probability of needing repetitions, these were more than 75% for all three subtests. This suggests that three out of four subjects needed to repeat the at least one of the conditions, as a result of either missing value or the two repetitions being too far from each other. It can also be seen that the mean extra repetitions was three.