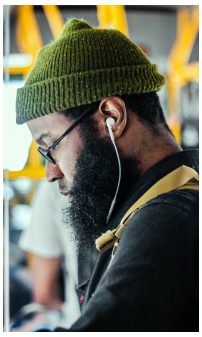
Test-retest evaluation

JacotiHearingCenter






Evaluation of Jacoti hearing diagnostics based on the user data

Jacoti provides a hearing diagnostics tool, which can be used by anyone, without the need to visit a healthcare professional. It comes in a form of an iPhone application Jacoti Hearing Center [1], which carries out two types of automated test: pure tone audiometry test and screening test. Users who own an iPhone device and wired EarPods interested in assessing their hearing can download the App for free and complete the test. The tests can be performed anywhere and without any external control. Therefore, it is important to know how reliable they are in practice. Below, we describe the analysis we performed to find that out. The results show that test-retest variability of both Jacoti diagnostic tests equals that of clinical audiometry, indicating that the app is a reliable option for self-assessment of hearing thresholds.

1 Jacoti Hearing Center is separately Class II FDA listed Medical Device, classified under product code EWO in the U.S. Jacoti Hearing Center is CE registered in Europe as a Class IIa medical device according to Rule 10 of Annex IX of the Medical Devices Directive 93/42/EEC.

Test-retest analysis as a way to validate automated audiometry

Automated audiometry test designs have been around for many years and there are established ways to measure the validity of an automated audiometry. One of the ways to assess if a test is valid is to measure how consistent are the results obtained by the test. This can be done by comparing the results of the same tests, performed twice. If the difference between these results is in a range that is considered normal for typical, non-automated audiometry, the test is likely to be as reliable as the clinical audiometry.

Pros and cons of user data analysis

Although the idea to perform audiometry in an automated way is not new, the implementation of this idea in a form of a freely available smartphone application is. Analyzing audiometry data of users of the App is something relatively new and has its own nature. One of the biggest advantages is the amount of data that is available for this analysis - there are thousands of users who downloaded Jacoti Hearing Center and performed the hearing test. Such population size is rarely available in studies of audiometry data. On the other hand, the amount of data comes at the cost. Even with a requirement screen and a noise-monitoring component, which ensure proper usage of the application, the application cannot ensure ideally controllable and reproducible test conditions.

Challenges of the uncontrolled test conditions

Users of Jacoti Hearing Center come from various places all over the world and it makes sense to presume that they use the app in very different environments, with different attitudes. It is possible that they start or abort the audiometry test session at any time, perform a test very carefully, or engage in another task simultaneously.

© 2021 Jacoti BV. All rights reserved.

Additionally, the test is not performed in a soundproof booth, but at home, where some degree of ambient noise can usually be measured. Background noise is a problem for audiometry, because the ability to hear a sound can only be determined if the test tone is louder than the noise. Jacoti Hearing Center includes a Noise Monitoring Component to reject thresholds measured under excessive ambient noise, but even very quiet rooms can be too noisy for testing low hearing thresholds. Since acoustic conditions at home are not free from noise, the precision of the test for very soft sounds is limited.

Ceiling effect and its influence on test-retest comparison

Because of the intended use under noisy conditions, Jacoti Hearing Center does not test thresholds lower than 10 dB HL. No matter if the real hearing threshold is at 0, 5, or 10 dB HL, the threshold measured by the app will always be 10 dB HL. For screening tests, no threshold lower than 25 dB HL is tested. At the other end of the measurable scale, device output power limit and safety requirements come into play: the loudest sound played back by the app cannot exceed the level that would be uncomfortable for the user. Therefore, iOS device connected to Apple Headset has a limited output and cannot test thresholds above 85 dB HL.

This is what is called the bottom/ceiling effect: there is a hard limit of the range of measurable values. This effect has an influence on the test-retest differences. For example, if both in the first test, as well as in the retest the measured threshold is 10 dB HL, the computed test-retest difference is 0 dB HL. Normally, the small test-retest difference is desired and indicates a reliable test procedure, however in this case it is not known whether the 0 dB HL is a valid value or rather just a consequence of the limited measurement range. As a result, the ceiling effects might produce misleadingly low threshold deviation, which is not related to the reliability of the diagnostic test.

Even though there are many studies validating the smartphone-based hearing technologies [5], they usually use the data collected under controlled conditions. There are no established ways of treating the more problematic, sometimes incomplete data collected from users in real-world environments. The high-level goal of this analysis was to assess how reliable are the diagnostic tests provided by Jacoti Hearing Center based on the data from these real-world uses of the smartphone application.

Relation to previous study:

Similar analysis has already been done before and was presented at the VCCA conference [1].
The VCCA study measured the frequency-dependent test-retest variability. The test-retest measurements pairs were defined as two consecutive measurements at the same frequency.

As a follow-up to that study, a further aspect of test-retest variability was explored: in addition to computing the test-retest difference for measurement pairs at individual frequencies, we analysed the test-retest for *session pairs*. The approach was to find session pairs from the same ear and user, which have several frequencies in common and to compute the averaged across frequency, session-to-session threshold difference.

Additionally, the ceiling effects were taken under consideration: in the analysis the threshold differences coming from the thresholds at the limits of the measurable range were excluded.

Advantages of session-related analysis:

- Knowing that the measurements came from a similar test procedure (including same frequencies), we can provide a fairer test-retest comparison.
- The fully completed sessions are more likely to be performed by a user with a required care.
- We can quantify the influence of session on the results (variability across measurement sessions).

Step 1: Collecting information about the measurement sessions

Initially, the general information about the measurement sessions performed by the users was extracted. Jacoti data base stores audiological data of almost 10 thousand users. From them, around 3000 carried out the test at least twice, which allowed for studying test-retest reliability. Within that group, most of the users repeated a test 2-3 times, but there are also users who repeated a test more than 100 times. They might be hearing practitioners

© 2021 Jacoti BV. All rights reserved.

already using the App as an alternative tool for testing their patients.

In 50 % of cases, the session is repeated by the user within one day, in 75% the time difference is less than 7 days and in almost 90 % of the cases it's less than 50 days.

There is quite a large number of sessions in which full measurement procedure was performed (see Figure 1 and 2). Around 50 % of all audiometry sessions (12 frequencies), and around 90% of all screening sessions (4 frequencies). In the 50 % of audiometry sessions, the audiometry test was incomplete and contained (in an order of likelihood) 6, 4, 2, 10 or 8 frequencies.
A test session can be incomplete if a user decides to abort the session or if the in-build noise monitoring component excludes some of the measurements.

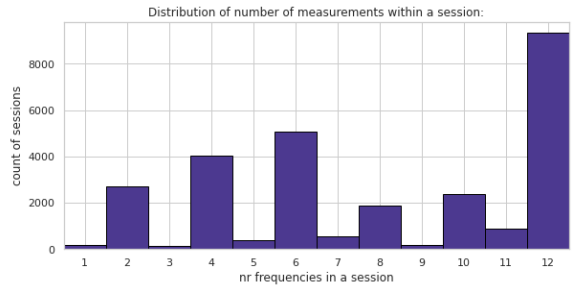


Figure 1 Histogram of the number of frequencies measured in one session of audiometry test.

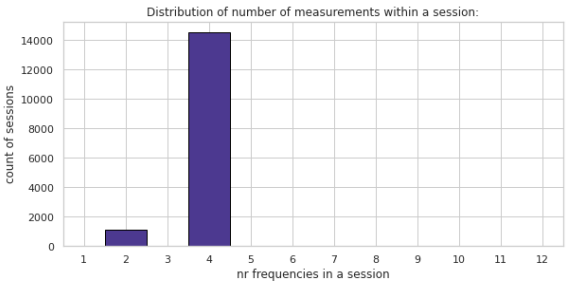


Figure 2 Histogram of the number of frequencies measured in a screening test.

Step 2: Analysing the test-retest difference of all session pairs

The test sessions coming from the same user and the same ear were arranged into *consecutive session pairs*. 12281 session pairs originating from 1823 users were considered for audiometry test and 5705 session pairs originating from 1464 users for screening test. In each pair, the difference between the measured thresholds was computed for all frequencies. The differences coming from thresholds at the bottom or ceiling of the measurable scale were excluded from the analysis.

Note: Test-retest deviation is used in many studies as measure of how reliable a diagnostic tool is. In the context of audiometry, a test-retest deviation of 5 dB HL (sometimes 10 dB HL) is considered a consistent result [2, 3, 5]. To indicate how reliable the test is in general, apart from the typical boxplots, the percentage of the test-retest differences within the 5 dB HL and 10 dB HL can be computed. In this report, to represent test-retest statistics the typical boxplots, the letter-value plots [4], the percentage of differences within 5 dB margin, and the percentage of differences within 10 dB margin are used.

➔ Test-retest threshold difference for different frequencies

Figure 3 and 4 show threshold differences for different frequencies. For audiometry test, in all tested frequencies min. 60 % of the observed threshold differences lie within the margin of 5 dB and min. 79 % within 10 dB. For screening min.62 % were within 5 dB and min. 81 % within 10 dB. Although the tests were performed in the uncontrolled conditions, the results are consistent in most of the test-retest pairs. It shows a good reliability of that test.

In both tests, a higher variance was observed in the lower frequencies (see Figure 3 and 4). It was in agreement with the previous Jacoti study measuring how much the sound level at the eardrum might vary depending on the way the earphones have been positioned in the ear. It was shown that the variation of the sound pressure level in the frequencies ranging from 125 Hz to 1000 Hz was larger than for other frequencies. Another possible explanation could be that the environmental noises, which influence the results, can typically be found in this frequency range.

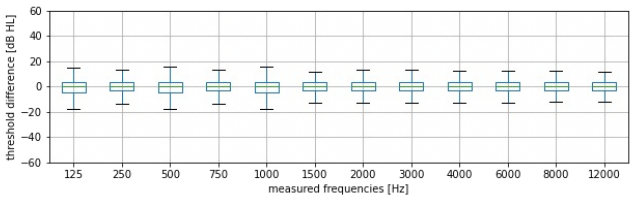


Figure 3 Statistics of threshold difference as a function of frequencies of Jacoti audiometry test.

© 2021 Jacoti BV. All rights reserved.

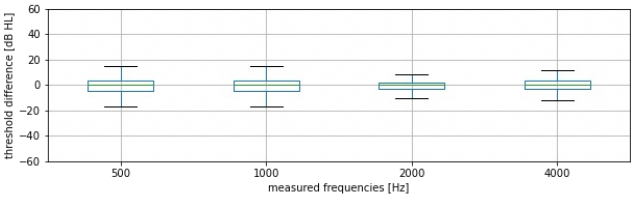


Figure 4 Statistics of threshold difference as a function of frequencies of Jacoti screening test.

➔ Average threshold difference

For each session pair the *average threshold difference* (across frequencies) was computed. It was a measure of the overal threshold deviation between two sessions.
In both tests, the values were clearly concentrated around zero, however for the screening test the deviation is slightly bigger than in audiometry test. This could be caused by the lower precision of the screening test or by the number of frequencies contributing to the average (only 4 for screening test, and 12 for audiometry test).

For the Jacoti audiometry, **73 %** of average threshold differences for lie within 5 dB and **86%** within 10 dB. For the Jacoti screening , there was **75 %** within 5 dB and **88 %** within 10 dB.

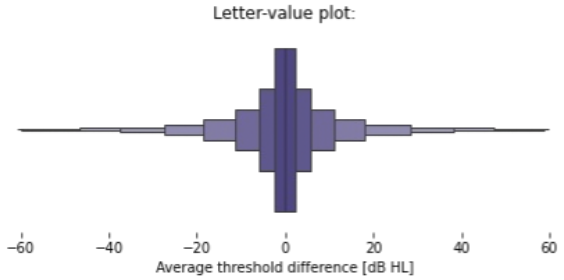


Figure 5 Distribution of average threshold deviation for Jacoti audiometry test.

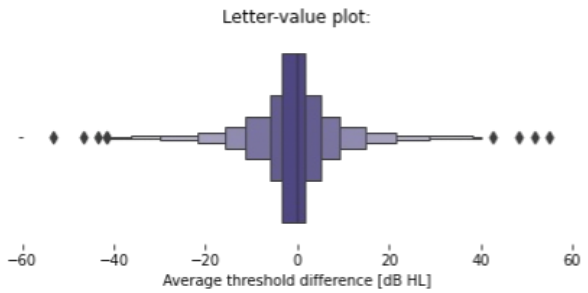


Figure 6 Distribution of average threshold deviation for Jacoti screening test.

Time is a factor that can potentially influence the threshold difference between two sessions. One could expect that the further two sessions are in time, the bigger the difference in the measured threshold would be, however no correlation between the time difference and the average threshold difference was found (see Figure 7 and 8).

Nonetheless, the analysis of average threshold difference values revealed that there is a fraction of the screening session pairs for which it is equal to 60 dB HL (see Figure 8). Because of the limits of the range of measurable values (25 dB HL - 85 dB HL), 60 dB HL is the largest possible difference which can be observed in between two measured thresholds. The 60 dB HL average threshold difference for a session pair means that in one session all the frequencies were measured at 25 dB HL and in the second session, all frequencies were measured at 85 dB HL (or the other way around). It is not very likely that the hearing threshold undergoes such a dramatic change between two test session and might be indication of the *incorrect use* of the screening test. The 60 dB HL average threshold difference can be found in 7 % of all screening session pairs.

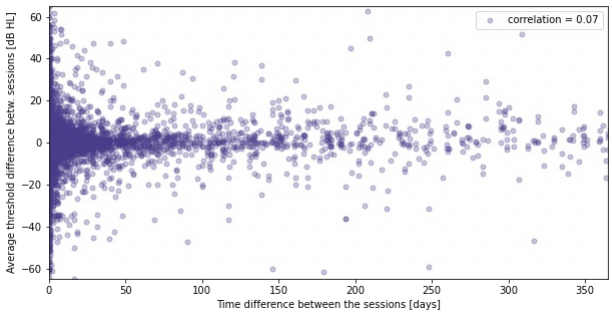


Figure 7 Average threshold difference as a function of time difference for audiometry test.

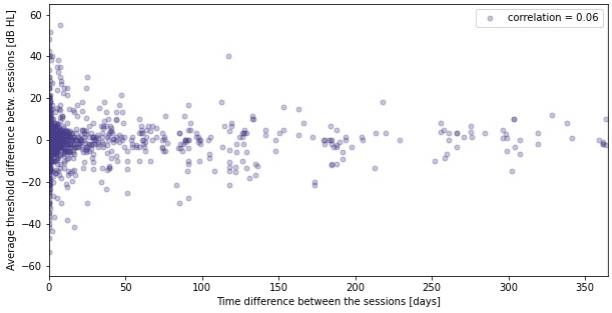


Figure 8 Average threshold difference as a function of time difference for screening test.

© 2021 Jacoti BV. All rights reserved.

➔ Test-retest deviation for different categories of hearing impairment

Each session was assigned to a hearing loss category based on the classification proposed by the WHO [6] based on the audiometric ISO value (see Table 1). To classify a session *pair*, the audiometric ISO value of two session was averaged. Test-retest average threshold deviation was analyzed in each hearing loss category.

Audiometric ISO value *(average of values at 500,1000, 2000 and 4000 Hz)*	*Hearing loss category*
<20 dB HL	normal
20 – 40 dB HL	mild
40 – 60 dB HL	moderate
60 – 80 dB HL	severe
>80 dB HL	profound

Table 1 Categories of hearing loss according to the WHO-proposed classification system.

Small differences were observed between the average threshold deviation in different hearing categories. For audiometry test the mild and moderate group has the highest test-retest deviation and for Screening test the highest test-retest deviation is observed in the severe and profound Category. A possible explanation is that session pairs, for which we observe large threshold differences are assigned to these intermediate categories. For example, if the first of the pair is in normal category and the second is in severe category, the session pair would be considered moderate category. The test-retest deviation in the mild and middle categories is more influenced by this type of session pairs.

Nevertheless, in all categories, more than **60 %** of the average threshold differences were within the 5 dB HL and minimum **80 %** in 10 dB HL.

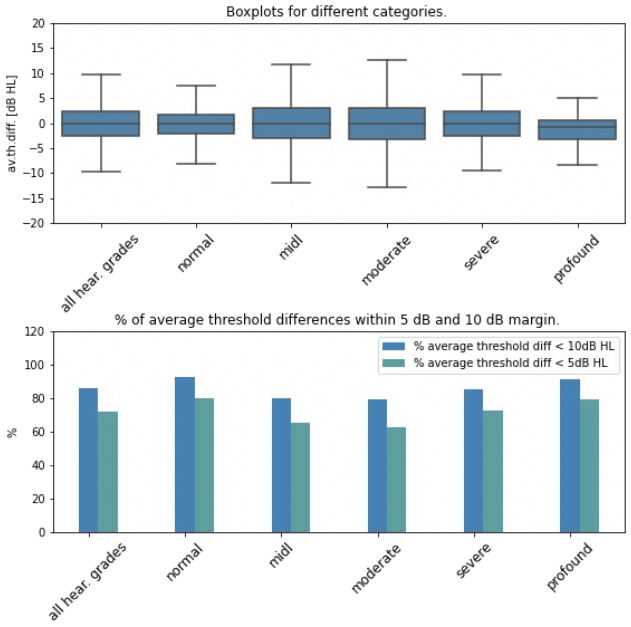


Figure 9 Average threshold deviation for different categories of hearing loss in Jacoti audiometry test.

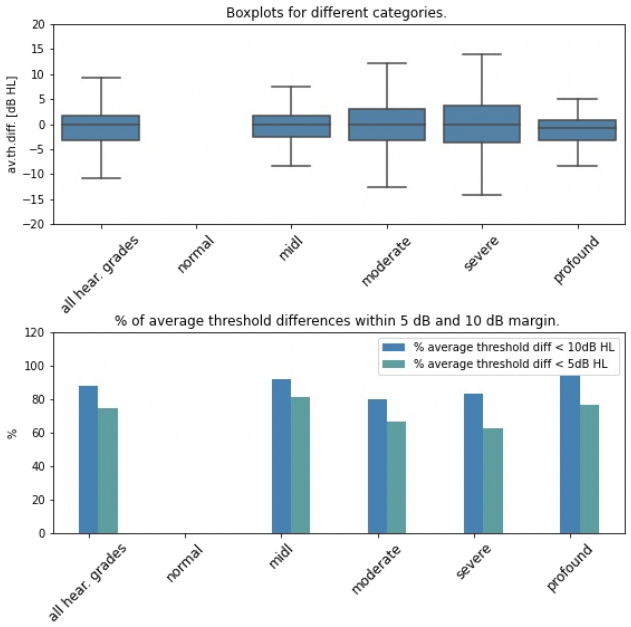


Figure 10 Average threshold deviation for different categories of hearing loss in Jacoti screening test.

➔ A closer look into excluded bottom/ceiling scores.

Originally, excluding bottom and ceiling effects was performed to remove the underestimated threshold differences which can make the overall threshold difference seem very low, even if doesn't reflect reality. However, looking closer in the excluded values, we can see that the method excudes also high threshold differences (See Figure 11).

© 2021 Jacoti BV. All rights reserved.

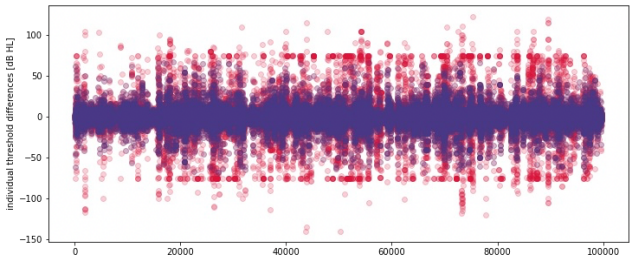


Figure 11 Data points before and after excluding bottom/ceiling threshold differences, Jacoti audiometry test.

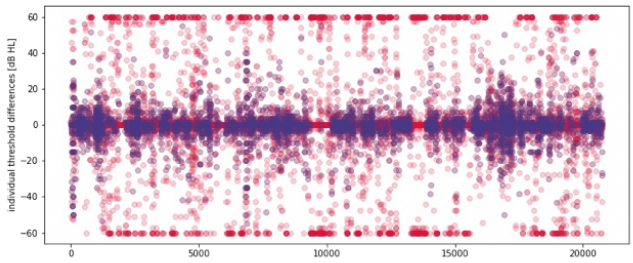


Figure 12 Data points before and after excluding bottom/ceiling threshold differences, Jacoti screening test.

Theoretically, a high threshold difference should only be seen when a user's hearing has degraded a lot since the previous measurement. However, a situation in which the hearing threshold degrades or improves by more than 40 dB HL within a few days is very unlikely. A more realistic explanation for such high threshold differences is a usability issue that might appear when interacting with a smartphone app. Another reason could be a test procedure with a large step size, which very quickly reaches either the bottom or the ceiling threshold.

The method for exluding bottom/ceiling thresholds removes most of these very high differences. This is correct, as these scores are outliers and can lead to falsen the overall results. However, they indicate that there might be some relatively frequent usability issue, which should be examined in the future.

Step 3: Analysing the test-retest difference of a subset of session pairs

An important advantage of the session-based test-retest analysis is that it allows for analyzing a subset of most reliable sessions independently. From the available session pairs, 1474 audiometry pairs from 500 users and 1240 screening pairs from 593 users were selected according to these criteria:

- the session pair had to be fully completed
- the session pair had to originate from the user who did not perform more than 10 sessions of the same type
- the time difference of the session pair had to be smaller than one year

This subset was considered the most reliable, because it did not include tests discontinued by a user or aborted by the noise monitoring system. By excluding users who performed more than 10 sessions it was assured that the test-retest pairs came from the same subject. If a retest is performed more than one year after the first test, the difference can reflect the change in the patients hearing loss, therefore these pairs were also excluded in this data subset.

The full analysis was performed for the selected subset of session pairs. According to the expectation, the test-retest reliability improved both for audiometry and for screening. Table 2 summarizes the test-retest scores for the Jacoti audiometry and screening test and compares the results for all session pairs with the results for a subset of complete session pairs.

TEST-RETEST DIFFERENCE	BASED ON ALL AVAILABLE SESSION PAIRS	BASED ON SUBSET OF MOST RELIABLE SESSION PAIRS
IN INDIVIDUAL FREQUENCIES, AUDIOMETRY	Min.60 % < 5dB Min. 80 % < 10 dB	Min. 62 % < 5 dB Min. 84 % < 10 dB
IN INDIVIDUAL FREQUENCIES, SCREENING	Min. 62 % < 5 dB Min. 81 % < 10 dB	Min.67 % < 5dB Min. 84 % <10 dB
AVERAGED ACROSS FREQUENCIES, AUDIOMETRY	72 % < 5 dB 86 % < 10 dB	77 % < 5 dB 89 % < 10 dB
AVERAGED ACROSS FREQUENCIES, SCREENING	75 % < 5dB 88 % < 10dB	76 % < 5 dB 88 % < 10 dB

Table 2 Summary of the test-retest scores for all session pairs and for the subset of most reliable (complete) session pairs.

What has been found?

The test-retest analysis of the *audiometry* and *screening* test provided by the Jacoti Hearing Center has shown that both tests can be considered reliable: the majority of all test-retest threshold differences lie within the margin of 5 dB HL. This result does not depend much, neither on the frequency, nor on the hearing loss category. For a subset of complete

© 2021 Jacoti BV. All rights reserved.

session pairs, the results are even slightly better, indicating that the test reliability benefits when users perform the test with a required care.

The proposed method for excluding invalid threshold deviations originating from ceiling or bottom of the scale removes both the abnormally low and abnormally high threshold differences, that can distort the results. The unnaturally low differences are a direct consequence of the limited measurable scale and not much can be done to avoid them at the test level. On the other hand, the very high threshold differences can be a result of a too quickly converging test procedure or some difficulty during user interaction with the app. This aspect has to be investigated in the future. One technical solution could be to detect the users incorrectly using the app by performing a session test-retest comparison after the second test. The incorrect use could be communicated to the user (similar to detecting asymmetric hearing loss).

Although test-retest analysis for the data from the real uses of the smartphone app brings various challenges, it could be performed. It is of great interest to analyse the user data in addition to studies from the lab. It might require developing new analysis methods but is an essential aspect for assessing the true reliability of the smartphone tests. It was helpful to focus only on the most complete fraction of collected audiometry data: the test-retest sessions pairs containing all frequencies of the test. Results for both tests provided by Jacoti Hearing Center show that this smartphone-based audiometry can be in general as reliable as clinical audiometry, even in the uncontrolled test conditions.

This report was created by:

Joanna Luberadzka, Jonatan Rivilla, Amaury Hazan

References:

[1] Amaury Hazan, Jonatan Rivilla, Num Méndez, Nicolas Wack, Oscar Paytuvi, Andrzej Zarowski, Erwin Offeciers, Jacques Kinsbergen. Test-retest analysis of aggregated audiometry testing data using Jacoti Hearing Center self- testing application. VCCA Conference 2019

[2] Svensson, U. Peter, Olav Kvaløy, and Tone Berg. "A comparison of test–retest variability and time efficiency of auditory thresholds measured with pure tone audiometry and new early warning test." Applied Acoustics 90 (2015): 153-159

[3] Schmuziger, Nicolas, Rudolf Probst, and Jacek Smurzynski. "Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones." Ear and hearing 25.2 (2004): 127-132)

[4] Hofmann, H., Wickham, H., & Kafadar, K. (2017). Letter Value plots: Boxplots for large data. Journal of Computational and Graphical Statistics, 26(3), 469-477

[5] Sandström, Josefin, et al. "Accuracy and reliability of smartphone self-test audiometry in community clinics in low income settings: a comparative study." Annals of Otology, Rhinology & Laryngology 129.6 (2020): 578-584

[6] Olusanya, Bolajoko O., Adrian C. Davis, and Howard J. Hoffman. "Hearing loss grades and the International classification of functioning, disability and health." Bulletin of the World Health Organization 97.10 (2019): 725.

© 2021 Jacoti BV. All rights reserved.