Abstract

Manual pure tone audiometry has been in consistent use for a long period of time, and is considered to be the 'gold standard' for the assessment of hearing thresholds. Increasing legislative requirements, and a significant global cost impact of noise induced hearing loss, means that a significant amount of reliance is placed on this tool for diagnosis.

There are a number of questions regarding the degree of accuracy of pure tone audiometry when undertaken in field conditions, particularly relating to the difference in conditions between laboratory and clinical environments.

This study assessed the test-retest variability of a number of commercially available screening pure tone audiometers, all with recent calibration. Repeated clinical tests were made in accordance with ISO 8253-1:2010 of the hearing thresholds of 12 volunteers, including subjects with both normal and impaired hearing, and were compared with results from a previous study in which testing was undertaken in a laboratory setting using a Bruël and Kjaer Head and Torso simulator.

The results of both laboratory and clinical studies showed a high level of test-retest variability, with maximum between test variation of up to 35 dB. Considerable variation occurred at all frequencies, with a particularly high level of variation at 6-8kHz. Levels of variation measured in this study suggests a high potential for diagnostic error when using screening pure tone audiometry, particularly in higher frequencies.

Keywords: Audiometry, Calibration, Variance, Hearing Threshold, Audiometer

1. Introduction

In the field of audiology, manually operated hearing assessment using pure tones according to ISO 8253-1:2010 [1] is considered the 'gold standard' for the assessment of hearing thresholds by airborne conduction [2]. It is for this reason that it is vitally important to make efforts to continually monitor the accuracy and reliability of the practice, in order to see that it meets modern requirements.

Exposure to high levels of noise has long been recognized as a health hazard, with the long-term result of noise-induced hearing disorder in the majority of people. Several sources [3,4,5] state that long-term exposure to sound pressure levels as low as 80 dBA poses some risk of noise induced hearing disorder, while exposure to levels regularly above 85 dB LAeq poses a risk of 'mild' hearing damage to most people, with the risk of more severe damage increasing with both length and level of exposure [5, 6].

Hearing loss presents a significant global cost impact, including the costs to productivity as well as the cost of long-term healthcare [7]. With an ageing population and legislative requirements to screen and protect workers from occupational hearing damage, the traditional method of audiometric screening in which tests are run on a 1:1 basis with a qualified audiometrist is expensive [8]. The reliance on pure tone audiometry means that it is important to continually assess the test procedure and equipment used for repeatability and accuracy.

2. Background

The fundamental methodology and equipment for audiometry have stayed very similar for a long period of time. Pure tone audiometry was originally developed from "tuning fork" tests of the early 20th century, and has now been in use for over 90 years [9], with early audiometers by and Western Electric available as early as 1923. More advanced features such as bone conduction audiometry and masking noise were already available on audiometers from Sonotone as early as 1928 [10].

The basic structure of an air conduction audiometric test involves an audiometrist, an audiometer and a patient response system. The audiometrist manipulates the audiometer to deliver pure (sine wave) tones to the patient at known amplitudes. The patient then responds to which sounds are heard through the patient response system [11]. The audiometrist uses the pattern of patient responses to determine the threshold of hearing for that patient at the various audiometric frequencies. If a person is hard of hearing in a single ear, masking noise may be used – a broad frequency spectrum sound that masks one ear from hearing loud tone presentations from a very insensitive ear.

The systems used for screening have remained fundamentally unchanged for several decades. Particular transducers (notably the Telephonics ® TDH-39 and TDH-49 supra-aural headphones) have been at the core of audiometric screening since the 1960s. Despite having been conceived during WWII, [12] and being in general usage since at least the 1960s [13] these are still commonly used, as they are some of the only transducers which have been standardized [11, 14, 15].

Although there have been studies of the reliability of different types of screening (automated, computer controlled, manual), and examination of the need for traceable calibration, there has been relatively little research into the degree of variation in performance of calibrated audiometers in clinical situations.

There are a number of aspects which suggest that calibrated audiometers may not be as reliable as generally thought. One issue is that the specifications do not currently require accreditation of the calibrating organization, as the guidelines simply state that the calibration should be performed by a 'competent' laboratory [1].

This leaves the standards open to interpretation, and many audiometer manufacturers recommend annual calibration to take place in their own facilities, which may or may not be accredited. This has the potential for errors in the accurate production of tones in audiometer systems, with no centralized standardizing authority to supervise.

The standard for the reference level of tones [15] also has a high level of 'acceptable' variation for any given tone presentation, with an acceptable variation of +/- 3 dB for test signals from 125 Hz to 4 kHz, and +/- 5 dB for test signals above 4 kHz. This means that testing the same tone and presentation level with different meters could result in an at ear difference of 6-10 dB with the maximum allowable variation.

Another important issue is that of the level of uncertainty in 'field' testing compared to laboratory conditions. The acoustic coupler (artificial ear) defined in IEC 60318-1:2009 used to assess particular headphones is a regular shape, standardized to particular dimensions [16]. The headphone is coupled to the artificial ear with a static force of 4.5 N (+/- 0.5 N) from either a mass or calibrated jig [15], rather than using the tension from the headphone band.

While this method allows for a high level of standardization in the testing of the transducer and tone generator in the system, it assumes that there is a minimal effect on the sound pressure level presented at the ear from the asymmetry of a human head and ear, as well as from different shapes and sizes of ears and heads, and variations in force of coupling. A previous study [8] found significant variation in the sound pressure level at the ear of an artificial head, in which the physiology of the head (head size and shape, size and structure of pinnae) are representative of a generic human head. As the study in [8] was undertaken under laboratory conditions using an artificial head, it was hypothesised that under clinical situations with a range of subjects there would be a further increase in measurement uncertainty.

This study aimed to assess the level of variation between measured hearing threshold level (HTL) of subjects under normal occupational screening conditions, using a variety of different manual audiometers.

3. Method

1.1. Equipment

The current study used three of the same screening audiometers as used in [8]. The fourth audiometer used in the laboratory study was not available. The audiometers were chosen to represent the whole market range of typical screening audiometers, ranging in cost from £995 GBP for the least expensive up to £4500 GBP for the most expensive.

The method was designed to give a representative sample of the performance of typical audiometers under real clinical conditions to compare to the simulated conditions in the laboratory study. Each of the audiometers had recently undergone certified traceable calibration by its recommended laboratory, meaning that the tone presentation from all these pieces of equipment should theoretically be identical. Although the standard does allow for some variance, the ideal would be for each audiometer to perform at its reference value.

1.2. Participants

Participants were made up of people who were concerned about their hearing who responded to a University-wide email. All respondents were offered the opportunity to participate in the study. The subjects were therefore self-selecting, and were not selected with any bias, other than those imposed by the demographic of the University staff and students. In total 12 subjects participated in the study, with 7 male and 5 female. The cohort included subjects with normal hearing and also subjects with a registered hearing loss and the age range was between 18 and 62 years.

Ethical approval was sought and obtained for the study by the university Research Ethics Panel, and all participants participated under full consent, having had the task and purpose fully explained to them, and with the right to withdraw from the study at any point.

1.3. Experimental procedure

The test schedule was run over 3 weeks, with each subject taking a test at the same time of the same day each week. The tests were held at a constant time of the day in order to control for varying levels of concentration experienced during the day, which might affect the test. The order in which audiometer units were used for each subject was randomized over the 3 weeks in order to reduce the impact of learning effect in which subjects become better at the tests through repetition. Testing took place in the same hemi-anechoic

chamber used for the laboratory testing. This has a NR rating of 18 and an average background noise level of 16 dBA, so the noise floor is significantly lower than the minimum requirements for audiometry [1].

Each subject was tested with each of 3 calibrated audiometers using an identical method, by the same qualified audiometrist. Each of the tests were performed to British Society of Audiology procedural guidelines [11], which are based on, and stand alongside the ISO Standard procedure [1]. Otoscopy was performed on all patients before each test in order to ascertain that excessive cerumen or other otological issues did not affect the test.

Subjects were measured over the normal audiometric frequencies of 500 Hz, 1 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz, and 8 kHz [11] and their hearing threshold level (HTL) for each ear was recorded.

4. Results

The measured thresholds of each subject for each test were analysed with a particular focus on variation in results between the hearing threshold level (HTL) for each frequency for the same subject at the same ear over repeated tests.

Table 1 shows the maximum variation in hearing threshold level between repeated tests for each ear of each subject. The between test variation for the same ear of the same subject ranges from a minimum of 0 dB (where the subject responded with no variation in each test at that frequency) to a maximum of 35 dB for the worst-case scenario.

Table 1: Maximum and mean variation in hearing threshold level between same-ear retests for individual subjects.

Subject	Ear	MaxVar	Mean	StDev
1	Left	15	8.6	3.8
	Right	15	6.4	4.8
2	Left	35	10.0	12.2
	Right	10	5.7	1.9
3	Left	15	7.1	3.9
	Right	15	7.1	4.9
4	Left	15	10.0	4.1
	Right	15	5.7	5.3
5	Left	15	7.1	3.9
	Right	15	9.3	3.5
6	Left	20	7.1	6.4
	Right	10	2.1	3.9
7	Left	15	7.9	4.9
	Right	15	8.6	4.8
8	Left	15	9.1	4.3
	Right	30	15.7	11.7
9	Left	15	8.6	5.6
	Right	10	5.0	4.1
10	Left	5	2.1	2.7
	Right	15	4.3	5.3
11	Left	15	7.1	4.9
	Right	15	8.6	6.3
12	Left	10	4.3	3.5
	Right	15	6.4	4.8

As it demonstrates the largest potential error between separate screening tests, the overall maximum variation between retests is an important consideration when analyzing the potential for error in diagnosis. The maximum and mean variation for retests for each frequency across the whole cohort are shown in Table 2 and figure 3.

There is significant individual variation between subjects, however the majority of subjects have at least 15 dB maximum difference between the measured HTL across repeated tests, with 2 subjects having 30 or more dB variance between repeated tests in one ear.

Table 2: Maximum and mean variation in hearing threshold level between same-ear retests for the cohort as a function of frequency.

	Test-retest variation (dB HL)		
Frequency	*Max*	*Mean*	*StDev*
500 Hz	15	5.8	4.1
1000 Hz	20	5.4	5.3
2000 Hz	15	5.6	4.3
3000 Hz	15	4.6	4.1
4000 Hz	20	8.5	5.2
6000 Hz	30	10.4	6.9
8000 Hz	35	10.4	7.5
R^2	0.804	0.766	0.765

Mean variation across the cohort ranges from 4.6 dB to 10.4 dB, with a standard deviation between 4.1 and 7.5 dB, dependent on frequency.

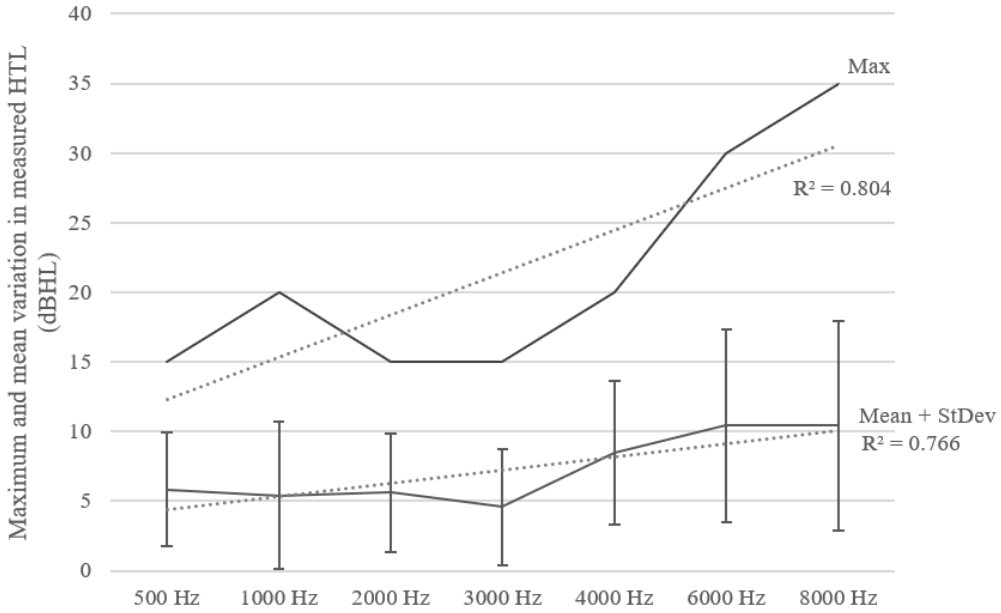

Figure 3: Maximum and mean variation in HTL for a series of audiometric tests using different audiometers.

Strong positive correlations are demonstrated between the frequency and each of the measures of Maximum variation ($R^2 = 0.804$), mean variation ($R^2 = 0.776$) (figure 6) and standard deviation ($R^2 = 0.765$) (Figure 3), indicating increasing levels of variability between the accuracy of tests with increasing frequency of the test tone.

This suggests that the use of higher frequencies could be particularly problematic when used for occupational health screening as they pose an increased risk of error.

5. Discussion

While the sample size was relatively small, the audiometers used are typical of the manufacturers and types of audiometer used in the UK and Europe. Each of the audiometers was calibrated to the appropriate standards by a competent laboratory.

A key metric is the maximum *possible* variation between presentations of the same frequency/ear, as this represents the worst-case scenario that could occur between different tests even of the same subject. The maximum variation between results of tests with different audiometers for the same subject/ear was 35 dB at the 8kHz tone, although the mean variation across the cohort for this frequency is only 10.4 dB. All frequencies had a maximum test-retest variation of at least 15 dB across the cohort.

These results can be compared to a previous study by the authors [8] which examined the variation of tone presentation at the ears of a calibrated Bruël and Kjær® Head and Torso Simulator (HATS) of type 4100, using four different audiometers in a 'simulated' clinical environment, and measuring the free field corrected LA_{EQ} of the tones presented to each ear.

Results from [8] demonstrated that the measured SPL for repeated tests on an ear simulator which replicates the shape of the human head and pinnae varies considerably. This reported a maximum variation of 21 dB between the absolute maximum and minimum values recorded in either ear across all audiometers, with a mean value of 9.8 dB and standard deviation of 5.5 dB, significantly higher than the variance allowable by the standard. It was suggested that this is due to the variability in acoustic coupling and the difficulty of accurate headphone placement on a human head.

The results shown here demonstrate a level of variation between measured hearing threshold level higher than can be attributed to the variation in sound pressure level discussed in [8]. It is reasonable to assume that the increased variation is caused by differences in the quality of acoustic coupling of the headphone to the auditory canal on each subject, particularly given individual differences in the physiologies of the different subjects. Test-retest placement of the headphones is also likely to be more difficult for the audiometrist when on a human subject rather than a completely immobile artificial head, which could add a further degree of error.

It is worth noting that variance in results is very subject dependent. From the raw data, it was observed that there was a particularly high variation on one test in the left ear at 8 kHz (subject 2), and another with a high variation in the right ear at 6 kHz (subject 8) which were not reflected in the other subjects. It can be hypothesized that particularly high levels of variability could be due to different levels of patient noise exposure, or even simply down to their mood or concentration.

As the systems were calibrated, there should theoretically be minimal variation in output sound pressure level of the transducer itself. However, calibration is undertaken on a symmetrical assembly, which has a high level of acoustic coupling between the artificial ear and the headphone, and it is suggested that a source of error is variation in the at ear level caused by factors related to the asymmetry of the human head.

Subject variability is more pronounced in the high frequency areas of the audiogram. There are two main reasons that could attribute this error. One is that directionality effects in the TDH-39 headphone design can cause headphone placement to be a factor, with some audiologists potentially placing the headphones onto the patient's ears in an off on-axis orientation, in which some occlusion is caused by the tragus, and causing a reduction in the sound pressure level at the eardrum. The other main factor could be temporary threshold shift caused by exposure to loud noise [18]. Human ears are more efficient in the 6 kHz range [19], and so this frequency area is more likely to be affected by threshold changes due to loud noise. However, this high frequency error appeared in both the clinical and the 'simulated clinical' testing under laboratory conditions. It is therefore to be suggested that this variation is more likely to be caused by placement of the headphones, causing occlusion of the sound wave or poor acoustic coupling.

Interestingly, some authors have suggested that the hearing threshold at 6 kHz is set too high, as there is a high proportion of patients who present a threshold shift at this frequency [20]. The results of this study indicate that this could be linked to variation in performance of the headphones with slight differences of placement. Other authors have suggested that artifacts from supra-aural transducers can cause errors in estimation of noise induced hearing loss, possibly linked to variance in standing wave frequency of different lengths of ear canal [21]. This is another exacerbating feature, though it would not explain the level of test-retest variation for the same subject.

Another potential source of error is the variation of tension in the headband which couples the transducers to the subject's ears. The calibration standards for audiometers [15] require a static force of 4.5N (+/-0.5N) in order to obtain a high-quality acoustic coupling between the transducer and the ear. The different

headband designs and different sizes of the heads of subjects under test are likely to cause variation in this acoustic coupling, and poor headband tension is noted by the British Society of Audiologists as a potential problem in testing. It is therefore reasonable to assume that some of the variability seen in this data could be attributed to differences in headband tension.

6. Conclusion

Results from repeated audiometric tests using screening audiometers under clinical conditions show a wide range of variability, which is increased above an already high level of variation in tests on an artificial head previously reported.

This suggests that test results from conventional pure tone screening systems need to be carefully assessed for the possibility of error or misdiagnosis, particularly where these include higher frequencies (4kHz or above). While the test itself is still considered 'fit for purpose', the potential for error is high, and other tests such as speech audiometry should always be used in conjunction, in order to reduce the possibility of diagnostic error.

The transducers commonly used in audiometric screening should be revisited, as developments in headphone technology over recent years has resulted in the availability of far higher quality transducers than those commonly used in audiometry, and it is suggested that a move should be made towards adopting a more contemporary design as a standard.

There is also potential for improving accuracy of the current methods of pure tone audiometry. Two particular areas are identified for further study in order to improve audiometer design. The first is the acoustic coupling of the transducer, in order to minimize occlusion effects at high frequencies caused by slight misplacement, which could reduce the sound pressure level at the eardrum. Acoustic coupling of the transducer could be improved by using circum-aural headphones rather than supra-aural, while at the same time improving external noise rejection, which is useful in occupational screening situations. Use of sensors or micro-cameras inside the headphone could improve positional accuracy.

The second is the relationship between headband tension and tone presentation. This may require the use of higher headband tensions or different headsets appropriate to different sizes of head. As this is potentially an important contributing factor, further research needs to be done on the exact impact of headband tension on results. An alternative approach is the use of in-ear transducers, which have been supported by other authors [21]. However there are concerns regarding infection control when using in-ear devices, and they are not usable with all patients.

Acknowledgement

This project was supported by the Technology Strategy Board, UK 2013. Grant reference: SKTP 1000821 in conjunction with Strategic Audiology Services Ltd. This material was previously presented at the Euronoise 2015 conference, Maastricht, May 2015.

Role of the Funding Source

The Technology Strategy Board's role is to provide government funding to support innovative research and design as a partnership between universities and businesses and as such have no conflict of interest. Strategic Audiology Services Ltd jointly funded the research with the aim of investigating possible areas for improvement in audiometer design. The study design, data collection, analysis and writing of the report was undertaken independently of the funder, and the authors take responsibility for the integrity and accuracy of the data collection and analysis.

References

[1] International Organisation for Standardization: ISO 8253-1:2010 Acoustics-Audiometric Test Methods. Part 1: Pure Tone and Bone Conduction Audiometry. ISO: Geneva, 2010.

[2] R.J Roeser, J. Clarke: Pure tone tests. In: RJ Roeser, M. Valente, H. Hosford-Dunn, (eds). Audiology: Diagnosis. New York: Thieme Medical, 2000. 238-260.

[3] Action on hearing loss. Action on hearing loss; 2012. Available online from: http://www.actiononhearingloss.org.uk/hearing-health/hearing-loss-and-deafness/types-and-causes/noise-induced-hearing-loss/. [Last cited on 2019 Apr 2].

[4] American Speech-Language-Hearing Association. Hearing Disorders – Noise. Available online from: https://www.asha.org/public/hearing/disorders/noise.htm. [Last cited on 2019 Apr 2].

[5] Health and Safety Executive: UK Statutory Instrument 2005 no 1643 The Control of Noise at Work Regulations 2005. HMSO: Norwich, 2005.

[6] C. Barlow, F. Castilla-Sanchez: Occupational noise exposure and regulatory adherence in music venues in the United Kingdom. Noise Health (14) 2012, 86-90.

[7] A Davis, P Smith, M Ferguson, D Stephens and I Gianopoulos: Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models, Health Technology Assessment; 11(42). 2007.

[8] C. Barlow, L. Davison. M. Ashmore and R. Weinstein. Amplitude variation in calibrated audiometer systems in clinical situations. Noise and Health 16 (72). 2014, 299-305.

[9] Feldmann H. A History of Audiology; a Comprehensive Report and Bibliography from the Earliest Beginnings to the Present. Translations of the Beltone Institute for Hearing Research. 22. (17). Chicago: The Beltone Institute 1970.

[10] D.A. Vogel, P.A. McCarthy, GW Bratt, C. Brewer. The clinical audiogram: Its history and current use. Commun Disord Rev 1(2) 2007, 83-94.

[11] British Society of Audiology. Recommended procedure. Pure-tone air-conduction and bone conduction threshold audiometry with and without masking. British Society of Audiology, Reading, 2011.

[12] Telephonics Corporation, TDH Series Audiometric Earphones Specifications, Telephonics Corporation, New York, 1939.

[13] W. G. Thomas, M. J. Preslar, R. R. Summers, and J. L. Stewart, Calibration and working condition of 100 audiometers, Public health reports, vol. 84, p. 311, 1969.

[14] International Organization for Standardization. ISO 389-1: 2017. Acoustics – Reference zero for the calibration of audiometric equipment. ISO: Geneva, 2017.

[15] British Standards Institute. BS EN ISO 60645-1: 2001 Electroacoustics – Audiological Equipment. Part 1 – Pure Tone Audiometers. British Standards Institute, 2001.

[16] International Electrotechnical Commission. IEC 60318-1:2009. Electroacoustics-Simulators of Human Head and Ear-Part 1: Ear Simulator for the Measurement of Supra-aural and Circumaural Earphones. IEC, Geneva, 2009.

[17] American National Standards Institute. ANSI S3.6-2004 Specifications for Audiometers. New York: ANSI/ASA; 2004.

[18] American Hearing Foundation. Noise induced hearing loss.

Available online from https://www.american-hearing.org/disease/common-ear-and-hearing-issues/ [last cited 01/03/2019]

[19] H Takeshima, Y. Suzuki. Equal loudness level countours for pure tones. Journal of the Acoustical Society of America, 2004. 16(2) 918-933

[20] D. McBride and S. Williams: Characteristics of the audiometric notch as a clinical sign of noise exposure. Scandinavian Audiology, 30,(2). 106-111. 2001.

[21] I.S Bhatt. Supra-aural transducer-related artifact contributes to overestimation of noise-induced hearing loss. The Journal of the Acoustical Society of America; 143, 2055-2058. 2018