VALIDITY OF AUTOMATED THRESHOLD AUDIOMETRY: A SYSTEMATIC REVIEW AND META-ANALYSIS

Faheema Mahomed [1], De Wet Swanepoel [1,2,3], Robert H Eikelboom [1,2,3]
Maggi Soer [1]

1. Department of Communication Pathology, University of Pretoria, Pretoria, South Africa
2. Ear Science Institute Australia, Subiaco, Australia
3. Ear Sciences Centre, School of Surgery, The University of Western Australia, Nedlands, Australia

Journal: Ear and Hearing

ABSTRACT

Objectives: A systematic literature review and meta-analysis on the validity (test–retest reliability and accuracy) of automated threshold audiometry compared with the gold standard of manual threshold audiometry was conducted.

Design: A systematic literature review was completed in peer-reviewed databases on automated compared with manual threshold audiometry. Subsequently a meta-analysis was conducted on the validity of automated audiometry. A multifaceted approach, covering several databases and using different search strategies was used to ensure comprehensive coverage and to cross-check search findings. Databases included: MEDLINE, SCOPUS, and PubMed with a secondary search strategy reviewing references from identified reports. Reports including within subject comparisons of manual and automated threshold audiometry were selected according to inclusion/exclusion criteria before data were extracted. For the meta-analysis weighted mean differences (and standard deviations) on test–retest reliability for automated compared with manual audiometry were determined to assess the validity of automated threshold audiometry.

Results: In total, 29 reports on automated audiometry (method of limits and the method of adjustment techniques) met the inclusion criteria and were included in this review. Most reports included data on adult populations using air conduction testing with limited data on children, bone conduction testing, and the effects of hearing status on automated audiometry. Meta-analysis test–retest reliability for automated audiometry was within typical test–retest variability for manual audiometry. Accuracy results on the meta-analysis indicated overall average differences between manual and automated air conduction audiometry (0.4 dB; 6.1 SD) to be comparable with test–retest differences for manual (1.3 dB; 6.1 SD) and automated (0.3 dB; 6.9 SD) audiometry. No significant differences ($p > 0.01$; summarized data analysis of variance) were seen in any of the comparisons between test–retest reliability of manual and automated audiometry compared with differences between manual and automated audiometry.

Conclusions: Automated audiometry provides an accurate measure of hearing threshold, but validation data are still limited for (a) automated bone conduction audiometry; (b) automated audiometry in children and difficult-to-test populations; and (c) different types and degrees of hearing loss.

Keywords: automated threshold audiometry, validation, test-retest reliability, accuracy.

INTRODUCTION

Automated healthcare services may include screening, diagnostic, and intervention procedures that can be conducted without the necessary healthcare professional's direct involvement. In situations where specialist healthcare personnel are limited or unavailable, this approach may ensure that services and healthcare resources are optimized (Margolis & Morgan 2008; Swanepoel et al. 2010). Automated threshold audiometry has existed for many years; however, it has not been used widely in clinical practice apart from occupational healthcare settings (Margolis & Morgan 2008).

The earliest record of automated threshold audiometry was in the seminal report of Georg von Békésy (1947). This self-recording threshold audiometer automatically increased and decreased the sound intensity while sweeping through the test-frequency range and became known as "sweep frequency Békésy audiometry." The patient is required to press a response button when the test signal is heard and release it when he or she loses perception of the signal. This method of determining the threshold is commonly known as the "method of adjustment." Subsequent systems used derivations of this technique with fixed-frequency threshold-seeking algorithms, referred to as fixed or discreet frequency Békésy audiometry, where a sweep in intensity occurs within a fixed frequency based on the patient's behavioral response relayed through a response switch (Meyer-Bisch 1996; Franks 2001).

In later years automated audiometry systems were programmed according to conventional manual audiometry procedural steps (Sparks 1972), typically using versions of the Hughson and Westlake threshold-seeking method (Hughson & Westlake 1944). The audiometer automatically makes adjustments to the intensity of the presented signal, up or downward depending on the response or lack of response. This method is known as the "method of limits." This method has also been modified in some cases to include forced-choice responses from the patient. Here the listener is required to listen and make a response that either indicates that a sound was heard or not. This can be done, for example, by pressing the appropriate “button” on a touch-screen monitor after a signal is presented (Franks 2001; Margolis & Morgan 2008).

Pure-tone threshold audiometry measures are especially suited to automation because they are based on predetermined sequenced steps (Margolis & Morgan 2008). In addition, when using a computer, results can be recorded automatically enabling all the advantages of electronic record keeping, such as reduced paperwork, transfer to other clinicians, and tracking change over time. In addition, automated testing can incorporate quality monitoring mechanisms to ensure consistent and reliable results as has recently been demonstrated (Margolis et al. 2007, 2011). Automation may also potentially improve standardization of tests protocols and procedures across clinics and even within clinics.

At present, the need for hearing healthcare services globally far outweighs the current capacity to deliver the services (Goulios & Patuzzi 2008; Fagan & Jacobs 2009; Swanepoel et al. 2010; Margolis et al. 2010, 2011). Automated audiometry has been proposed as a way to increase the reach of audiometry in underserved areas especially when conducted within asynchronous telehealth framework (Swanepoel et al. 2010; Swanepoel & Hall 2010). An automated audiometer cannot replace an audiologist, but a system that can determine pure-tone hearing thresholds with similar accuracy to that of manual audiometry may be beneficial in addressing the demand for hearing health services. Optimizing limited professional resources by incorporating automation may improve the reach of current audiological services and can improve the efficiency of current hearing healthcare resources (Margolis & Morgan 2008; Swanepoel et al. 2010).

Although automated threshold audiometry has existed for many decades, it has been used almost exclusively in industry as part of mass hearing screening and baseline monitoring and for research purposes. Clinical audiological practices, in contrast, have almost exclusively relied on conventional manual audiometry. This may partly be attributed to perceived concerns regarding the accuracy and reliability of automated air conduction (AC) and bone conduction (BC) audiometry and the availability of validation studies (Sparks 1972; Margolis & Morgan 2008). However, being a behavioral test procedure manual audiometry presents with normal variability in threshold determination (test– retest or intertester differences) due to subject factors such as fatigue and concentration as well as due to different transducers and test environments used (ANSI 1996; Margolis et al. 2007). Normal variability in audiometry has typically been quantified by test–retest reliability and occasionally by intertester reliability (Margolis et al. 2007; Ishak et al. 2011).

In the light of the potential benefits of automation in threshold audiometry, its long history, and the apparent lack of summative evidence supporting its use, the present study aimed to systematically review the current body of peer-reviewed publications on the validity (test–retest reliability and accuracy) of automated threshold audiometry. In addition the study included a meta-analysis, using results from published reports, to quantify the test–retest reliability and accuracy of automated threshold audiometry.

MATERIALS AND METHODS

Systematic Review

A systematic review of peer-reviewed literature was conducted to determine the validity, as measured by the accuracy and reliability, of automated threshold audiometry compared with manual threshold audiometry. Accuracy is defined as the indirect method of measurement between two different techniques measuring the same variable of which one is the gold standard (Bland & Altman 1999). Manual audiometry served as the gold standard and automated audiometry as the comparison method for determining auditory thresholds. Test–retest reliability refers to the ability of a test to give similar results when applied more than once on the same subjects under the same conditions (Dobie 1983).

A varied search strategy was used across several electronic databases to identify relevant research reports (excluding editorials, notes, and short surveys) from peer-reviewed literature. For inclusion reports were required to include some within-subject comparison of automated threshold audiometry to manual threshold audiometry (accuracy). Test–retest reliability information was also captured from the identified reports.

A multifaceted approach, covering several databases and using different search strategies, was used to ensure comprehensive coverage and cross-checking of search findings (White & Schmidt 2005). An initial search strategy was undertaken using the following databases and search engines: MEDLINE, SCOPUS, and PubMed. Searches were conducted on July 20, 2012 and included all relevant reports published until this date. Supplemental Digital Content 1 (see Table, links.lww.com/EANDH/A100) indicates the databases, search strategy, and search terms used.

The MEDLINE database search used a strategy of relevant key words to determine all records relating to the study aim (Supplemental Digital Content 1, Table, links.lww.com/EANDH/A100). The second database, PubMed, was searched using available Medical Subject Heading terms. SCOPUS, the third database included in the search strategy, is the world's largest abstract and citation database of peer reviewed literature also indexing MEDLINE. This served as a cross-check for reports from PubMed and MEDLINE databases.

Inclusion and exclusion criteria: only reports of a comparative nature between automated and manual threshold audiometry, written in English were included.

Descriptions of automated audiometry without these comparisons, reviews, articles, notes, and short surveys were not included.

The first author reviewed the abstracts of all reports resulting from the searches to determine whether the report complied with the inclusion criteria. If any queries arose the second author also reviewed the abstracts. Where an abstract was unavailable, the full article was reviewed (Table 1). After all duplicates and unrelated reports had been excluded, the remaining reports were reviewed in full to determine whether they met the inclusion criteria. A secondary search was used to supplement the findings of the primary search. This involved reviewing the reference lists of all reports already identified for inclusion during the primary search strategy for additional reports not identified with the primary search.

Table 1. Results from the applied search strategies

Procedural steps	Number of reports	Description
1. Database search results	1932	3 Databases (Medline, PubMed, Scopus).
2. Database results excluding duplicates	1311	621 duplicates omitted.
3. Database results excluding non-English reports	1072	223 reports omitted.
4. Database results excluding reviews, short surveys and notes omitted	971	101 reports omitted.
5. Database results related to scope of review based on abstract and title	63	971 titles and abstracts reviewed for relevance, 908 records omitted, 63 complete articles reviewed.
6. Database results within scope of review based on full article	26	37 reports omitted based on inclusion/exclusion criteria. One could not be tracked due to incorrect indexing on the journal archive.
7. Additional reports within scope of review	3	3 reports identified from secondary search strategy surveying reference lists of 26 identified reports.
8. Final reports	29	Reports utilized in systematic review.
9. Reports utilized in meta-analysis	12	Reports with data appropriate to meta-analysis aims

The reports selected for review were carefully scrutinized and categorized according to the audiological threshold-seeking method used (method of adjustment or method of limits), type of evaluation (diagnostic or screening), AC or BC thresholds, type of transducers and audiometer used, age and hearing status of participants, type of statistical analysis for accuracy, test– retest reliability, and the conclusions drawn by the article.

Meta-Analysis

A meta-analysis was conducted to combine and quantify the results of individual reports so that an overall assessment of test–retest reliability and accuracy based on existing evidence could be made for automated audiometry. To be included in the meta-analysis, reports had to meet the following criteria: (a) The report had to include data comparing manual and automated audiometry in terms of accuracy; (b) Data had to be reported in the form of mean differences (real or absolute) and standard deviations with the number of observations reported.

Mean differences and standard deviations were documented. Weighted averages, using reported real and absolute average differences and standard deviations were determined for validation (test–retest reliability and accuracy) across studies, taking into account the number of observations reported. Furthermore, a comparison of test– retest threshold differences for manual and automated threshold audiometry, indicative of normal variability, was made with the difference between automated and manual audiometry (accuracy) using an analysis of variance (ANOVA) test (http://statpages.org/anovalsm.html). A significant difference in variability was noted by a $p < 0.01$.

RESULTS

Systematic Review

The systematic review procedural outcomes are summarized in Table 1. After excluding duplicates, reviews, short surveys, notes, and non–English-language records, 971 reports remained. Sixty-three reports were identified and subsequently the full-text was reviewed. One report (Raza 2008) could not be traced because its indexing on all databases did not correspond to the actual journal listing. Despite efforts to contact the authors and the journal the report could not be sourced. A total of 26 full reports were identified, which met the inclusion/exclusion criteria.

The second stage of the search strategy, involving a review of the reference lists of identified reports, revealed three additional reports, bringing the total number to 29 reports.

The final list of reports included in the systematic review date from 1956 to 2011 (Fig. 1). Supplemental Digital Content 2 (Table, http://links.lww.com/EANDH/A101) provides a summary of all reports included according to authors, year of publication, subject descriptions, test parameters, automated threshold-seeking method (method of limits/method of adjustment), research findings (accuracy or test–retest reliability), and conclusions.

Of the 29 reports, 15 used the method of adjustment and 13 the method of limits whereas one report used both methods (Harris 1979). The majority of reports covered diagnostic audiometry whereas four reports included screening applications of automated audiometry (one for method of limits; three for method of adjustment).

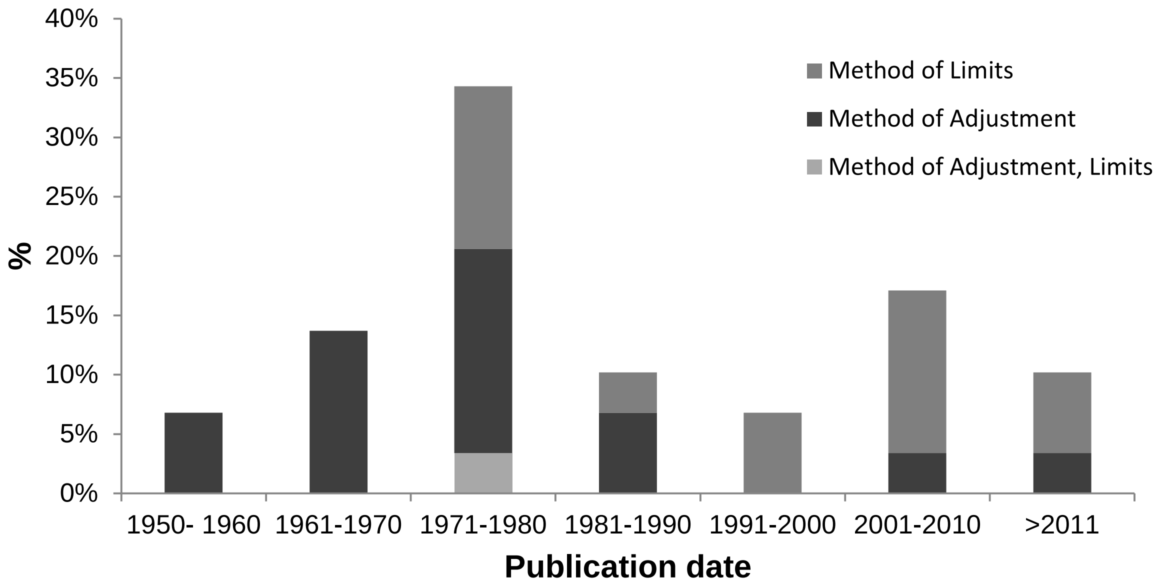

Figure 1. Distribution of reports included in systematic review (n=29) date of publication and type of automated audiometry (method of limits; method of adjustment, method of limits and adjustment).

Table 2 provides a description of data on accuracy and test– retest reliability included in the systematic review records. Test–retest reliability was included by 11 reports (7 for method of adjustment and 4 for method of limits). Ten of these included only AC audiometry, whereas one included both AC and BC audiometry. Of these 10 reports, three included participants with a hearing loss, whereas four did not indicate the hearing status of participants (Table 2).

Records obtained reported data using a variety of statistical analyses (Supplemental Digital Content 3, Table, http://links.lww.com/EANDH/A102). The most common presentation of test–retest data was presented in terms of average differences and standard deviations (n = 4) and average thresholds and standard deviations (n = 3).

All 29 reports provided information on the accuracy of automated threshold audiometry. Twenty-six records reported results for adult populations, 19 of these included AC audiometry only, whereas seven included AC and BC audiometry. Six of the 26 adult reports included persons with hearing loss only, five included persons with normal hearing, whereas six included persons with normal hearing or a hearing loss, and nine did not indicate the hearing status of their samples. Furthermore, only five of the studies reported results on children, two of which included AC and BC results.

Various techniques were used to document the accuracy, referred to as validity in records, of automated audiometry (Supplemental Digital Content 3, Table, http://links.lww.com/EANDH/A102). The most commonly used measures of accuracy were average differences between automated and manual audiometry with accompanying standard deviations (n = 11) and average thresholds and standard deviations (n = 11). Less commonly used techniques included absolute average differences and SDs (n = 6), t test (n = 4), and ANOVA analysis (n = 2).

Table 2. Distribution of air and bone conduction data for adults and children reported across studies identified in the systematic review (n=29)

	Accuracy				**Test-retest reliability**			
Type of hearing	**Normal hearing**	**Hearing loss**	**∗Both**	**Not indicated**	**Normal hearing**	**Hearing loss**	**∗Both**	**Not indicated**
Adults								
AC testing	5	3	3	8	2	3	1	4
AC and BC testing	-	3	3	1	-	-	-	1
Subtotal	**5**	**6**	**6**	**9**	**2**	**3**	**1**	**5**
Children								
AC testing	1	-	1	1	-	-	-	-
AC and BC testing	-	-	2	-	-	-	-	-
Subtotal	**1**	**-**	**3**	**1**	**0**	**0**	**0**	**0**
TOTAL	**6**	**6**	**9**	**10**	**2**	**3**	**1**	**5**

∗Indicating that both hearing and hearing loss subjects were included in the study.

AC, air conduction; BC, bone conduction.

Meta-Analysis

The meta-analysis used mean differences (real and absolute) and standard deviations at each frequency extracted from the reports, if available. In some reports the mean differences and standard deviations across all frequencies were not determined and thus were calculated when possible (i.e., if the number of observations were included). Supplemental Digital Content 4 (Table, http://links.lww.com/EANDH/A103) and Supplemental Digital Content 5 (see Table, http://links.lww.com/EANDH/A104) indicate summaries of the data obtained for test–retest reliability and accuracy across individual studies used in the meta-analysis. Weighted average calculations were subsequently obtained across these studies (Tables 3 and 4).

Only five reports provided data on test–retest reliability in the form of mean differences (real and absolute) and standard deviations for automated testing and manual testing. Test–retest variability for automated threshold audiometry indicated average differences that ranged between −1.1 and 2.2 dB with the standard deviation ranging between 6.2 and 10.4 dB for individual test frequencies, whereas the absolute average differences ranged between 2.0 and 4.9 dB with a standard deviation of 3.0 to 4.8 dB (Table 3).

Table 3. Meta-analysis weighted average test-retest reliability differences for manual and automated audiometry

Frequencies	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	3000 Hz	4000 Hz	6000 Hz	8000 Hz	All
MANUAL THRESHOLD AUDIOMETRY										
Average differences and standard deviations (3 reports)										
Average difference	-	-	2.3	2.1	1.5	2.0	-0.4	-1.7	-	1.3
N	-	-	500	500	500	40	500	40	-	532
Standard deviation	-	-	6.7	4.8	5.0	4.7	6.9	7.6	-	6.1
N	-	-	500	500	500	40	500	40	-	532
Absolute average differences and standard deviations (2 reports)										
Absolute average difference	4.8	3.4	2.9	3.2	2.7	-	2.8	-	3.0	3.2
N	60	80	80	80	80	-	80	-	80	80
Standard deviation	5	3.7	3.7	3.4	3.6	-	3.5	-	4.3	3.9
N	60	60	60	60	60	-	60	-	60	60
AUTOMATED THRESHOLD AUDIOMETRY										
Average differences and standard deviations (3 reports)										
Average difference	-	-	0.3	-1.1	0.0	2.1	0.7	1.7	-	0.3
N	-	-	500	500	500	40	500	40	-	532
Standard deviation	-	-	7.1	6.8	6.4	6.2	7.1	10.4	-	6.9
N	-	-	500	500	500	40	500	40	-	532
Absolute average differences and standard deviations (2 reports)										
Absolute average difference	4.9	3.4	2.9	2.6	2.6	-	2.3	-	2.0	2.9
N	60	80	80	80	80	-	80	-	80	80
Standard deviation	4.8	3.5	3.6	3.2	4.1	-	3.0	-	3.2	3.8
N	60	60	60	60	60	-	60	-	60	60

Table 4. Weighted average differences and standard deviations between manual and automated threshold audiometry (manual minus automated)

Frequencies	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	3000 Hz	4000 Hz	6000 Hz	8000 Hz	All
AVERAGE DIFFERENCES AND STANDARD DEVIATIONS										
Combined (10 reports)										
Average difference	-2.5	-3.5	-1.5	-1.2	-0.1	2.1	-3.6	-2.1	-5.0	0.4
N	232	360	796	796	796	428	796	556	384	820
Standard deviation	8.6	6.7	5.4	5.3	5.5	6.1	5.7	7.7	8.7	6.1
N	232	420	766	766	526	578	526	466	420	798
Method of limits (3 reports)										
Average difference	-	-0.4	-0.7	0.4	-1.3	-0.8	3.8	-1.3	-1.7	0.3
N	-	60	84	84	24	24	84	24	84	116
Standard deviation	-	5.1	4.4	5.3	5.8	-	4.9	-	7.0	5.5
N	-	60	60	60	60	-	60	-	60	92
Method of adjustment (7 reports)										
Average difference	-2.0	-2.3	0.5	0.3	2.1	1.1	0.1	-1.0	-3.1	0.8
N	232	360	796	796	796	428	796	556	384	796
Standard deviation	8.6	6.9	5.4	5.3	5.5	6.1	5.8	7.7	9.0	6.2
N	232	360	706	706	466	578	466	466	360	706
ABSOLUTE AVERAGE DIFFERENCES AND STANDARD DEVIATIONS										
Combined (4 reports)										
Absolute Average Difference	4.2	3.6	3.4	3.5	3.4	-	2.9	-	3.1	4.2
N	136	196	196	196	196	-	196	-	196	360
Standard deviation	4.0	3.5	3.9	3.6	3.8	-	3.2	-	4.5	5.0
N	136	196	196	196	196	-	196	-	196	345

Table 4 provides a summary of weighted average differences between manual and automated audiometry in the adult population. Results indicate that the overall (n = 10) average differences between automated and manual audiometry ranged between −5.0 and 2.1 dB across the frequency spectrum with the standard deviations ranging from 5.3 to 8.7 dB. Furthermore, the average differences obtained between the automated method of limits and manual audiometry ranged between −1.7 and 3.8 dB with standard deviations between 4.4 and 7 dB. In addition, method of adjustment audiometry yielded lower results at 0.125, 0.25, 6, and 8 kHz (−0.1 to −2.3 dB) whereas manual audiometry yielded higher results at the remaining frequencies, with the standard deviations ranging from 5.3 to 9 dB. The combined absolute differences ranged from 2.9 to 4.2 dB with standard deviations ranging from 3.2 to 4.5 dB.

Last, it should be noted that data from the two studies on children (4 to 10 years of age) were excluded from the meta-analysis as only one study using the method of adjustment (Békésy fixed-frequency testing) reported results in the form of average differences. These ranged between 3.6 and 20.3 dB with standard deviations ranging from 2.6 to 7.2 dB for 0.25, 1 and 4 kHz (Hartly & Siengenthalar 1964). Another study reported results in terms of absolute differences across all frequencies (4.1 dB), with a standard deviation of 1.7 dB (Margolis et al. 2011), when using an automated method of limits technique.

ANOVA comparisons of the meta-analysis weighted averages were conducted between the test–retest differences for manual and automated audiometry and the average difference between manual and automated thresholds (accuracy comparison) for the real and absolute differences. This was done for the combined category (method of limits and method of adjustment) and between method of adjustment and method of limits average differences. No statistically significant differences ($p > 0.01$; summarized data ANOVA) were obtained between any of the comparisons of test–retest (manual and automated) threshold differences and automated compared with manual threshold differences.

DISCUSSION

Comparison of two audiometric threshold techniques, such as automated and manual audiometry, has been performed using a variety of statistical analyses (Supplemental Digital Content 3, Table). Measures of agreement determined by the two threshold-seeking methods most commonly included the average difference (with standard deviation), average thresholds (with standard deviation), and average absolute differences (with standard deviation). The average difference is valuable in showing a systematic effect but negative and positive differences may cancel each other out even when large differences in either direction exist. Bland and Altman (1986) recommend the use of absolute average differences and standard deviation as a more appropriate measure of correspondence because it provides an indicator of the expected spread in variability. With this in mind, the meta-analysis was conducted using average differences (real and absolute) and standard deviations to draw conclusions regarding the validity of automated audiometry when compared with manual audiometry.

Automated Audiometry Test–Retest Reliability

Test–retest reliability is defined as the repeatability of a technique and allows comparison of techniques to determine which is more precise (Bland & Altman 1986). Eleven reports in this systematic review included results on test–retest reliability, of which four used the method of limits and seven the method of adjustment for threshold audiometry. In each case, reported test–retest reliability for automated audiometry was indicated to be within typical variability when compared with the test– retest reliability of manual audiometry (Burns & Hinchcliffe 1957; Gosztonyi et al. 1971; Formby et al. 1996; Robinson & Whittle 1973; Erlandsson et al. 1979a, b; Lutman et al. 1989; Fautsi et al. 1990; Ho et al. 2009; Ishak et al. 2011; Swanepoel et al. 2011). Only Ishak et al. (2011) reported higher test–retest variability with Bèkèsy sweep-frequency audiometry, but reported that using a slower sweep rate of 20 seconds per octave would improve the acquired test–retest reliability.

Several reports indicated that the second test session produced slightly lower (i.e., better) thresholds than the first session when manual and automated audiometry were used (Burns & Hinchcliffe 1957; Gosztonyi et al. 1971; Robinson & Whittle 1973; Erlandsson et al. 1979a; Lutman et al. 1989; Fautsi et al. 1990; Fromby et al. 1996; Ho et al. 2009; Ishak et al. 2011, Swanepoel et al. 2011). Several of the reports attributed the lower thresholds during the second session to the learning effect (Erlandsson et al. 1979a, b; Lutman et al. 1989; Ishak et al. 2011). This suggests that subsequent studies should consider randomizing the order of testing techniques and control the previous experiences participants had with audiometric testing.

The meta-analysis showed overall test–retest variability to be similar for automated (5 reports) and manual AC audiometry (5 reports). Average differences obtained for manual and automated test–retest audiometry respectively were 1.3 dB (6.1 SD) and 0.3 dB (6.9 SD) and absolute differences of 3.2 dB (3.9 SD) and 2.9 dB (3.8 SD). The meta-analysis test–retest difference for automated compared with manual audiometry (Table 3) demonstrated no statistically significant difference (ANOVA; $p > 0.01$). Higher variability was noted at 6 kHz for both automated and manual AC audiometry, but this was because only one article reported data at 6 kHz (Burns & Hinchcliffe 1957). Burns and Hinchcliffe (1957) reported a high variability for 6 kHz, with standard deviations of 3 to 4 dB, higher than those obtained at the other tested frequencies in the study (Supplemental Digital Content 4, see Table, http://links.lww.com/EANDH/A103).

Meta-analysis test–retest results are consistent with previously reported standard deviations of average test–retest differences for manual audiometry, ranging between 4.4 and 6.2 dB for a group of adults and children (Stuart et al. 1991). A recent report (Swanepoel & Biagio 2011) on manual audiometry obtained absolute average test–retests differences (3.6 dB; 3.9 SD) that were in line with the meta-analysis results (2.9 dB; 3.8 SD). The AC test–retest threshold differences for automated audiometry fall well within present test–retest limits.

Ho et al. (2009) was the only study to report on automated BC test–retest reliability. Results were reported in terms of paired thresholds; the study concluded that test–retest reliability of automated BC audiometry was appropriate (Ho et al. 2009) and within typical manual BC test–retest reliability (Laukli & Fjermedal 1990; Margolis et al. 2010; Swanepoel & Biagio 2011).

Automated Audiometry Accuracy

Over the six decades since the first description of automated audiometry, only 29 reports (15 on method of adjustment, 13 on method of limits, and 1 using both method of limits and adjustment) have reported on the validation of automated audiometry by comparing results with the gold standard of manual audiometry.

The meta-analysis showed that overall average differences between manual and automated AC audiometry (0.4 dB; 6.1 SD) correspond to test–retest difference for manual (1.3 dB; 6.1 SD) and automated (0.3 dB; 6.9 SD) audiometry. No statistically significant difference (ANOVA; $p > 0.01$) was evident between overall absolute differences for manual and automated audiometry (4.2 dB; 5.0 SD) and the test–retest absolute differences for manual (3.2 dB; 3.9 SD) and automated (2.9 dB; 3.8 SD) audiometry (Table 3).

Average differences for manual and automated BC audiometry were only reported by nine studies. These studies used varied forms of analyses in terms of agreement (Supplemental Digital Content 3, Table) and as a result weighted averages for BC threshold audiometry could not be determined across studies.

Method of Adjustment

As demonstrated in Figure 1 the method of adjustment was the first type of automated threshold audiometry. Overall, 16 reports were identified including comparisons of manual and method of adjustment automated threshold audiometry. The manual audiometry threshold determination techniques in these reports included the modified Hughson-Westlake method and some variations thereof (Corso 1956; Burns et al. 1957; Hartley et al. 1964; Knight 1965; Jokinen 1969; Robinson & Whittle 1973; Erlandsson et al. 1979a; 1979b; Ishak et al. 2011) as indicated in Supplemental Digital Content 2 (see Table, http://links. lww.com/EANDH/A101).

Several reports included in the systematic review indicated that automated audiometry using the method of adjustment (Békésy sweep or Békésy fixed-frequency method) generally yields lower (i.e., better) thresholds compared with manual audiometry (Burns & Hinchcliffe 1957; Knight 1965; Jokinen 1969; Maiya & Kacker 1973; Robinson & Whittle 1973; Erlandsson et al. 1979a, 1979b; Harris 1979; Frampton & Courter 1989; Ishak et al. 2011). A single report showed manual audiometry having lower thresholds than the method of adjustment technique at certain frequencies (0.25, 6, and 8 kHz). The authors reported that the reason for this phenomenon was probably the threshold-seeking method used (Ishak et al. 2011).

The meta-analysis showed an average differences of 0.8 dB (6.2 SD) between automated (method of adjustment) and manual AC audiometry. There was no statistically significant difference (ANOVA; $p > 0.01$) when these results were compared with test–retest reliability of both manual (1.3 dB; 6.1 SD) and automated threshold audiometry (0.3 dB; 6.9 SD). The accuracy of automated (method of adjustment) threshold audiometry is therefore within the normal variability as defined by test–retest reliability. Margolis et al. (2010) compared automated and manual threshold differences between two audiologists using manual audiometry as opposed to test–retest reliability. The intertester differences (0.6 dB; 5.5 SD) for manual audiometry were similar to the average differences (0.8 dB; 6.2 SD) between manual and automated audiometry results obtained in the meta-analysis.

Four reports included screening audiometry, comparing manual and automated thresholds (method of adjustment). Three of these studies used children (Hartly & Siengenthalar 1964; Delany et al. 1966; McPherson et al. 2010) and one used an adult population (Gosztonyi et al. 1971). Delany et al. (1966) indicated that automated audiometry for participants provided results substantially in agreement with manual audiometry, however, as observed with adults, automated audiometry tends to produce thresholds that are slightly lower (−0.8 to −3.3 dB) than manual testing. In addition, the authors (Hartly & Siengenthalar 1964; Delany et al. 1966; McPherson et al. 2010) indicated that automated audiometry can produce useful threshold data with children down to the age of 6 years. As age decreases, however, a greater proportion of children are either unable to perform the test at all or frequently lose concentration so that portions of the test need to be repeated at a later stage to obtain a full audiogram.

Gosztonyi et al. (1971) reported on industrial screening conducted on salaried and hourly workers (N = 38 ears). This study indicated that manual audiometry thresholds may be significantly lower than automated thresholds but the authors later discovered that the reason for this phenomenon was the fact that all participants were involved in medicolegal cases. Thus the phenomenon of nonorganic hearing loss significantly increased the threshold differences obtained between manual and automated audiometry.

Although findings on the application of automated audiometry using the method of adjustment are promising, limited data are available for pediatric populations and BC testing. An important reason for no BC data in the method of adjustment technique is attributed to the difficulty in using masking with this method. It is challenging to use a masking noise on the contralateral ear as the narrowband noise level should theoretically change with the tested frequency (Meyer-Bisch 1996). In addition to the technical difficulties of such an operation, the test may become difficult to follow for the patient (Meyer- Bisch 1996).

Method of Limits

In the 1970s the focus of research on automated audiometry started to shift from method of adjustment techniques to the method of limits (Fig. 1). Overall, 13 reports used the method of limits for automated audiometry compared with manual audiometry. All the studies obtained in the systematic literature review reported no statistically significant difference for AC between manual and automated audiometry.

Meta-analysis weighted average difference (0.3 dB; 5.5 SD) obtained when comparing automated method of limits technique with manual audiometry was similar to the weighted average difference for the method of adjustment and manual audiometry (0.8 dB; 6.2 SD); no statistically significant difference was noted (ANOVA; $p > 0.01$). These findings correspond to test–retest reliability results of automated (1.3 dB; 6.1 SD) and manual (0.3 dB; 6.9 SD) audiometry, indicating no statistically significant difference (ANOVA; $p > 0.01$). The accuracy of method of limits automated audiometry is within normal variability as defined by test–retest reliability.

Seven of the 13 reports included findings on BC audiometry (Sparks, 1972; Wood et al. 1973; Picard et al. 1993; Margolis et al. 2007; Ho et al. 2009; Margolis et al. 2010; Margolis & Moore 2011). No statistically significant difference between manual and automated BC audiometry was noted across these studies. Margolis and Moore (2011) indicated a statistically significant difference between AC thresholds for manual and automated audiometry. The difference was partly attributed to the different transducers used (manual—TDH 50; automated— Sennheiser HDA 200) and the differential effect of low and high frequencies being tested.

CONCLUSIONS

Automated threshold audiometry has developed over six decades from method of adjustment (Békésy methods) automated procedures incorporating conventional manual audiometry (method of limits) threshold-seeking methods. Present evidence demonstrates similar test–retest reliability for automated compared with manual threshold audiometry, and automated audiometry thresholds being within typical test–retest and intertester variability of manual thresholds. Despite its long history, however, validation is still limited for (a) automated BC audiometry; (b) automated audiometry in children and difficultto- test populations, and (c) different types and degrees of hearing loss.

ACKNOWLEDGMENTS

The authors declare no conflict of interest.

Address for correspondence: De Wet Swanepoel, Department of Communication Pathology, University of Pretoria, Room 3–5, Level 3, Corner Lynnwood and University Road, Gauteng 0002, South Africa.

E-mail: dewet.swanepoel@up.ac.za

Received Novenber 20, 2012; accepted March 22, 2013.

REFERENCES

ANSI. (1996) Specification for audiometers. New York, NY: American National Standards Institute..

Békésy, V. G. (1947). A new audiometer. Acta Oto-Laryngol, 35, 411–422.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1, 307–310.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Stat Methods Med Res, 8, 135–160.

Burns, W., & Hinchcliffe, R. (1957). Comparison of the audiometry threshold as measured by individual pure tone and Bekesy audiometry.J Acoust Soc Am., 29, 1274–1277.

Corso, J. F. (1956). Effects of testing methods on hearing thresholds. AMA Arch Otolaryngol, 63, 78–91.

Delany, M. E., Whittle, L. S., Knox, E. C. (1966). A note on the use of selfrecording audiometry with children. J Laryngol Otol, 80, 1135–1143.

Dobie, R. A. (1983). Reliability and validity of industrial audiometry: Implications for hearing conservation program design. Laryngoscope, 93, 906–927.

Erlandsson, B., Håkanson, H., Ivarsson, A., et al. (1979a). Comparison of the hearing threshold measured by manual pure-tone and by self-recording (Békésy) audiometry. Audiology, 18, 414–429.

Erlandsson, B., Hakanson, H., Ivarsson, A., et al. (1979b). Comparison of the hearing threshold measured by pure-tone audiometry and by Bekesy sweep audiometry. Acta Oto-Laryngologica, 86, 54–55.

Fagan, J. J., & Jacobs, M. (2009). Survey of ENT services in Africa: Need for a comprehensive intervention. Glob Health Action, 2.

Fausti, S. A., Frey, R. H., Henry, J. A., et al. (1990). Reliability and validity of high-frequency (8-20 kHz) thresholds obtained on a computer-based audiometer as compared to a documented laboratory system. J Amer Acad Audiol, 1, 162-170.

Fausti, S. A., Frey, R. H., Henry, J. A., et al. (1993). Reliability and validity of high-frequency (8–20 kHz) thresholds obtained on a computer-based audiometer as compared to a documented laboratory system. J Am Acad Audiol, 1, 162–170.

Formby, C., Sherlock, L. P., Green, D. M. (1996). Evaluation of a maximum likelihood procedure for measuring pure-tone thresholds under computer control. J Am Acad Audiol, 7, 125–129.

Frampton, M. C., & Counter, R. T. (1989). A comparison of self recording audiometry in naval establishments and clinical audiometry in a hospital setting. J R Nav Med Serv, 75, 99–104.

Franks, J. R. (2001). Hearing measurement. In B. Goelzer, C. H. Hansen, G. A Sehrndt (Eds.). Occupational exposure to noise: evaluation, prevention and control. Dortmund, The Netherlands: World Health Organisation, Federal Institute for Occupational Safety and Health.

Gosztonyi Jr. R. E., Vassallo L. A., & Sataloff J. (1971). Audiometric reliability in industry. Archives of Environmental Health, 22, 113-118.

Frank, T., & Ragland, A. E. (1987). Repeatability of high-frequency bone conduction thresholds. Ear Hear, 8, 343–346.

Goulios, H., & Patuzzi, R. B. (2008). Audiology education and practice from an international perspective. Int J Audiol, 47, 647–664.

Harris, D. A. (1979). Microprocessor, self-recording and manual audiometry. J Aud Res, 19, 159–166.

Hartly, V. V., & Siengenthalar, B. M. (1964). Relationships between Bekesy fixed frequency and convential audiometry with children. J Audit Res, 4, 15–22.

Ho, A. T., Hildreth, A. J., Lindsey, L. (2009). Computer-assisted audiometry versus manual audiometry. Otol Neurotol, 30, 876–883.

Hughson, W. & Westlake, H. D. (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. Trans Am Acad Ophthalmol Otolaryngol, Suppl, pp, 3–15.

Ishak, W. S., Zhao, F., Stephens, D., et al. (2011). Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry. Audiol Med, 9, 40–46.

Jokinen, K. (1969). Presbyacusis. I. Comparison of manual and automatic thresholds. Acta Otolaryngol, 68, 327–335.

Laukli, E., & Fjermedal, O. (1990). Reproducibility of hearing threshold measurements. Supplementary data on bone-conduction and speech audiometry. Scand Audiol, 19, 187–190.

Lutman, M. E., Cane, M. A., Smith, P. A. (1989). Comparison of manual and computer-controlled self-recorded audiometric methods for serial monitoring of hearing. Br J Audiol, 23, 305–315.

Maiya, P. S., & Kacker, S. K. (1973). Comparison of threshold between pure tone and Bekesy audiometry. Silent Wld, 8, 16–20.

Margolis, R. H., Frisina, R., Walton, J. P. (2011). AMTAS: Automated method for testing auditory sensitivity: II. Air conduction audiograms in children and adults. Int J Audiol, 50, 434–439.

Margolis, R. H., Glasberg, B. R., Creeke, S., et al. (2010). AMTAS: Automated method for testing auditory sensitivity: Validation studies. Int J Audiol, 49, 185–194.

Margolis, R. H., & Moore, B. C. (2011). AMTAS(®): Automated method for testing auditory sensitivity: III. Sensorineural hearing loss and airbone gaps. Int J Audiol, 50, 440–447.

Margolis, R. H., & Morgan, D. E. (2008). Automated pure-tone audiometry: An analysis of capacity, need and benefit. Am J Audiol, 17, 109–113.

Margolis, R. H., Saly, G. L., Le, C., et al. (2007). Qualind: A method for assessing the accuracy of automated tests. J Am Acad Audiol, 18, 78–89.

McPherson, B., Law, M. M., Wong, M. S. (2010). Hearing screening for school children: Comparison of low-cost, computer-based and conventional audiometry. Child Care Health Dev, 36, 323–331.

Meyer-Bisch, C. (1996). Audioscan: A high-definition audiometry technique based on constant-level frequency sweeps—A new method with new hearing indicators. Audiology, 35, 63–72.

Picard, M., Ilecki, H. J., Baxter, J. D. (1993). Clinical use of BOBCAT: Testing reliability and validity of computerized pure-tone audiometry with noise-exposed workers, children and the aged. Audiology, 32, 55–67.

Raza, S. N. (2008). Computer audiogram. J Coll Physicians Surg Pak, 18, 463–464.

Robinson, D. W., & Whittle, L. S. (1973). A comparison of self-recording and manual audiometry: Some systematic effects shown by unpractised subjects. J Sound and Vib, 26, 41–62.

Sparks, D. W. (1972). The feasibility of computerized audiometry. J Aud Res, 12, 62–66.

Stuart, A., Stenstrom, R., Tompkins, C., et al. (1991). Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults. Audiology, 30, 82–90.

Swanepoel, D., & Biagio, L. (2011). Validity of diagnostic computer-based air and forehead bone conduction audiometry. J Occup Environ Hygien, 8, 210-214.

Swanepoel, D., & Hall, J. W., III. (2010). A systematic review of telehealth applications in audiology. Telemed J E Health, 16, 181–200.

Swanepoel, D., Mngemane, S., Molemong, S., et al. (2010). Hearing assessment—Reliability, accuracy, and efficiency of automated audiometry. Telemed J E Health, 16, 557–563.

Swanepoel, D., Olusanya, B. O., Mars, M. (2010). Tele-audiology in sub- Saharan Africa. J Telemed Telecare, 16, 53–56.

White, A., & Schmidt, K. (2005). Systematic literature reviews. Complement Ther Med, 13, 54–60.

Wood, T. J., Wittich, W. W., Mahaffey, R. B. (1973). Computerized pure tone audiometric procedures. J Speech Hear Res, 16, 676–684.

Supplemental Digital Content 1. Table (Databases and search strategy details)

Database	Search strategy	Identifiers	Results	Limiters
Medline	Reports indicating findings of automated audiological testing. Terms occurring in the title, abstract, or keywords of articles.	“Automatic” OR “computerized” OR “computer-based”OR "pc-based" OR "automation" OR "automated" OR “audioscan” AND "audiometry" OR "hearing measurement" OR "hearing thresholds" OR "auditory thresholds" OR "hearing assessment" OR "hearing evaluation"	463	Reports published prior to 1946 not included
PubMed	MeSH terms related to automated audiological testing, occurring in the title and abstract.	“automatic” OR “computerized” OR “computer-based” OR “pc-based” OR “automation” OR “automated” OR “audioscan” AND “audiometry”	195	MeSH terms utilized only
Scopus	Reports indicating findings of automated audiological testing. Terms occurring in all fields.	“automatic” OR “computerized” OR “computer-based” OR “pc-based” OR “automation” OR “audioscan” OR “automation” “automated”, “self-recording”, “self-recorded” OR “Békésy” AND “audiometry”, “hearing measurement”, “hearing thresholds”, “auditory thresholds”, “hearing assessment” and “hearing evaluation”.	1274	None

Supplemental Digital Content 2. Table (Summary of reports included in review)

Author	Year	Subject description	Test parameters	Automated audiometry threshold seeking method	Research findings		Conclusion
					Accuracy	Test-retest	
Corso	1956	105 subjects (210 ears), 17-25 years old. Normal hearing adults	Diagnostic AC audiometry. Frequencies: .25, .5, 1, 1.5, 2, 3, 4, & 8 kHz). Transducers: Auto- oscillator type 1011 manual-oscillator type 1304-A Audiometer: Manual-Bekesy type audiometer, Reager Model, Automated-ADC audiometer, Model 50-E2	Method of Adjustment-Békésy fixed frequency. Frequency range of 2-8 kHz, starting at 40 dB. Testing time: 10min per ear was used with 0.5 dB rate per second. Thresholds obtained by the intersection of the midpoint curves and specific frequency lines.	- Average absolute thresholds and standard deviations -Test of significance (t-ration). -Difference in variability (F-ratio). -Pearson product-moment correlation coefficient.	-	Manual testing obtained thresholds that were lower than for automated testing (midpoint Békésy testing). Less variability in thresholds was noted between .25 and 2 kHz when manual testing was utilized. A low statistically significant positive correlation was noted at given frequencies between manual and automated audiometry.
Burns & Hichcliffe	1957	20 subjects (40 ears), 20-58 years of age. Hearing status not indicated	Diagnostic AC testing. Frequencies: .5, 1, 2, 3, 4, 6 kHz. Transducer: Standard Telephones Model 4026	Method of Adjustment - Békésy sweep frequency. Frequency range of .5-6 kHz was swept with a continuous tone, in 7 min 55 sec, paper speed of 1cm/min.	- Average difference and standard deviation - t-Test values	- Average difference and standard deviations - Product moment	Overall, manual and automated (Békésy) threshold audiometry gives essentially similar results. A significant difference was noted at 1000Hz, where Bekesy testing yielded a lower threshold of approximately 3 dB.

				Rate of change of intensity, increasing and decreasing, approximately 2 dB/sec. Thresholds obtained by the intersection of the midpoint curves and specific frequency lines.		correlation coefficients. -t-Test	Reliability was satisfactory at all frequencies utilizing both audiometric testing methods, besides at 500 Hz where the second automated test yielded a lowering of thresholds of 1-2 dB.
Hartely & Siengenthalar.	1964	30 subjects (60 ears) 13 children: 4 - 5 years old; 17 children: 8-10 years old. Normal hearing children.	Diagnostic AC Testing. Frequencies: .25, 1, 4 kHz. Audiometer: Manual- Audiovox Model 7-B, automated- Granson-Stadler Model E-800,	Method of Adjustment- Békésy fixed frequency. 1 min fixed frequency tracing (timed to begin after 3 reversals on the tracing) were obtained. Thresholds read using the mean mid-point between peaks and valleys.	- Average thresholds - Average difference and Standard deviations -t-Test -Within subject variability – t-Test	**-**	Better standard of acuity for manual compared to automated threshold audiometry were obtained. The difference was greater for younger children than older children. Within subject variability for automated threshold testing was higher than manual testing. Significant difference of variability at .25 kHz for the older group and at 4 kHz for the younger group.
Delany *et al.*	1966	66 ears, 17-29 years old. Hearing status not indicated.	Diagnostic AC testing. Frequencies: .5, 1, 2, 3, 4, 6 kHz. Transducer: 4026A earphones Audiometer:	Method of Adjustment - Békésy fixed frequency. Frequencies tested at kHz/sec. Tone burst	- Average difference	-	Automated threshold audiometry gives results substantially in accord with manual audiometry. The differences over most frequencies are small, but automated threshold

			Automated- mobile audiometric laboratory, manual- not indicated	presentation rate: 2 tones/sec.			audiometry gives lower threshold levels.
Knight	1965	66 ears. Normal hearing subjects.	Diagnostic AC testing. Frequencies: .5, 1, 2, 3, 4, 6 kHz. Audiometer: Manual and automated- Grason-Stadler model E 800	Method of Adjustment. Attenuator speed: 5 dB/sec, tone pulsed 2/sec.	-Average difference and standard deviation	-	Manual and automated audiometry is equivalent, as they yield threshold levels on average that are within 1 dB.
Jokinen	1969	**4 groups:** 1) 19 subjects (30 ears), 19-24 years old, inexperienced, normal hearing subjects. 2)15 subjects (30 ears), 19-24 years old, experienced outpatients, normal hearing. 3) 9 subjects	Diagnostic AC testing. Frequencies: .125 .25, .5, 1, 2, 3, 4, 6, 8 kHz. Audiometer: Manual- Madsen Model OB 60, Automated- Granson Stdler model E800	Method of Adjustment - Békésy fixed frequency. Tones presented for 30 sec at a frequency, first with 200 msec pulsed tones, secondly with a continuous tone. Tone pulse, rise and fall time of 25 msec, with on and off ratio of 1: 1. Intensity changes: 0.25dB steps, rate: 2.1 dB/sec.	-Average differences and standard deviations	-	Various differences were seen in the 4 groups. The normal hearing, inexperienced and experienced groups, obtained better results with automated testing (both continues and pulsed tones) than with manual testing. The presbycusis group, with and without the acoustic trauma, indicated that manual and continues Békésy testing obtained the same results, however, pulsed Békésy testing obtained better thresholds

		(17 ears), 52-73 years old, presbycusis with drop at 4000Hz indicating an acoustic trauma. 4) 22 patients (39 ears), 53-81 years old, subjects had presbycusis					than manual testing.
Gosztonyi et al.	1971	***Accuracy*** 19 subjects. ***Test-retest reliability*** 46 salaried employees and 25 hourly employees. All noise exposed adults.	Industrial screening AC testing. Frequencies: .5, 1, 2, 4, 8 kHz. Audiometer: Automated-self-recording audiometer, manual-standard clinical audiometer.	Method of Adjustment.	- Average thresholds	- Average difference	Manual testing produced better thresholds than automated testing , there was a difference of 10 dB between the two. Test- retest reliability for salaried employees indicated a difference no more than 10 dB. In this study it was investigated that the reason for the great difference between thresholds was as a result of subjects either being influenced to claim for HL or had compensation cases or had compensation legislations in progress.
Sparks	1972	15 subjects.	Diagnostic AC	Method of limits.		-	It was apparent that if

		Bi-modal population of mild or severe hearing loss participants used.	and BC testing, with masking. Frequencies: .25, .5, 1, 2, 4, 8 kHz. Transducers: AC- TDH-39 housed in a MX-41 AR cushion. BC- Radioear B-70A oscillator Audiometer: Manual and automated- Beltone 15-C	A computer program using Hughson-Westlake procedure for threshold seeking, masking programmed according to Hood (1960). Computer program provided instructions, which were followed by an assistant who was familiar with the use of Teletype system. If a response was elicited the assistant would type 1, no response the assistant would type 2. The computer would indicate next step.	-Average thresholds and standard deviations. -t-Test conducted on mean values. -Product moment correlation coefficient.		subjects were consistent in their response, automated testing could obtain thresholds similar to that of manual testing. The t-test: no significant difference between AC and BC values between two methods of testing. Correlation coefficients: high correlation between the two methods of testing.
Maiya, & Kacker.	1973	20 subjects, 15-30 years. Normal hearing subjects.	Diagnostic AC testing. Frequencies: .125, .25, .5, 1, 2, 4, 6, 8 kHz. Audiometer: Manual- Maico-MA-8, Automated- Grason-Stadler Company model E- 800.	Method of Adjustment - Békésy sweep frequency. Rate: 1 octave/min, chart travel period of 6 2/3 min. Rate of change of intensity: 2.5dB/sec. Thresholds read using the mid-point mean value between	- Average thresholds	-	Automated and manual testing yielded similar thresholds, however automated testing seemed to be more sensitive than manual testing.

				ascending or descending tracing at the frequency level.			
Robinson & Whittle	1973	***Accuracy:*** 64 subjects (128 ears), 26-73 years old. ***Test-retest reliability:*** 48 subjects (96 ears), 29-73 years old. Hearing status not indicated.	Diagnostic AC testing. Frequencies: .25, .5, 1, 2, 4, 6, 8 kHz. Transducers: TDH-39 earphones and MX-41-AR cushions. Audiometer: Manual and automated- Rudmose type ARJ-5	Method of Adjustment - Békésy fixed frequency. Pulsed tones with a repetition rate: 2 Hz, cycle consisting of a silent period of 185 ms and a tone pulse with 65 ms rise, fall times and a dwell of 185 ms at maximum amplitude, attenuator: 5dB/s. Thresholds read as the mid-point of the excursions, extraneous deviations being ignored.	- Average differences and standard deviations -Linear regression and correlation coefficients. - Estimation of asymptomatic data.	- Average differences and standard deviations of initial test - Average differences and standard deviations of second test	Automated threshold yield better results than manual testing, except at .25 kHz where no diff was noted. Test-retest reliability: manual and automated testing yield lower thresholds when tested for the second time.
Wood *et al.*	1973	20 subjects, 7-72 years old. Hearing status of subjects included: 1 normal hearing	Diagnostic AC, BC testing with masking. Frequencies: .25, .5, 1, 2, 4, 8 kHz. Audiometer: Automated- Grason Stadler model 829E, manual- not	Method of limits. Functional generator controlled frequency of tonal signal. Rise and fall time: 30 sec, duration of the tone: 1500msec. ***Unmasked air and bone:*** Tones presented using	- Average deviations	-	A high positive relationship between manual and automated testing for air and bone testing was noted. Automated testing reduces examiner bias and causes direct standardization of testing. Additionally, the use of computerized program will

		subject, 14 sensorineural, 4 conductive and 1 mixed hearing loss subject/s.	indicated.	an initial bracketing of 10 dB, then a bracketing of 5dB. ***Masking:*** AC Masking- 40dB gap between AC of test ear and BC of non-test ear. BC Masking- if AC of the test ear exceeded the midline BC by more than 10dB. Minimal effective masking (Martin 1976) was used / if patient did not respond to minimal masking than platue masking was administered.			give the audiologist time for direct patient contact, counselling and aural rehabilitation.
Almqvist & Aursnen	1978	82 subjects (41 ears), 7-82 years. Hearing status not indicated.	Screening AC, Frequencies: .5, 1, 2, 3, 4, 6 Hz. Audiometer: Manual- not indicated, Automated- minicomputer, type PDP-8.	Method of limits. Computer program utilized principles based on manual audiometry.	-Standard deviation	-	Automated audiometry appeared to be a fast and a reliable method for screening audiometry. A total standard deviation of 4.8 dB was noted between manual and automated audiometry, standard deviation varied across frequencies and was the smallest in the speech frequencies.

Sakabe *et al.*	1978	***2 groups used:*** 1) 31 subjects (62 ears), 19- 22 years old. Normal hearing subjects. 2) 124 subjects (248 ears). Hearing status not indicated.	Diagnostic AC testing. Frequencies: .125, .25, .5, 1, 2, 4, 6, 8 kHz.	Method of limits. Automatically interrupted tone, on-off time: 2sec, rise- fall time: 25ms. Tone presented at 30dB, if not heard, raised to 60dB, if heard lowered again to 30dB and increased by 5dB till heard again. The tone is lowered to 30dB again and raised in 5dB steps till a response is elicited. Once a response is obtained a comparison between the 2 'thresholds' are made. The smaller value is the threshold obtained at that frequency.	- Error analysis	-	Automated audiometry has sufficient accuracy for practical use. Automated audiometry coincides with manual audiometry within 10 dB. Additionally it would take 5-15min to conduct.
Erlandsson *et al.*	1979	***Accuracy :*** 115 subjects (230 ears), 25 to 63 years. ***Test-retest reliability:*** 10 subjects (20 ears).	Diagnostic AC. Frequencies: .25, .5, 1, 1.5, 2, 3, 4, 6, 8 kHz. Transducers: Manual- TDH-39M with MX-41/AR cushions. Automated- TDH-49P with MX-	Method of adjustment- Békésy sweep frequency. Attenuation rate: 2.5 dB/s, pulsed tone-presentation; sweep time from .25 -10 kHz was 400s.	- Regression equations and α and β coefficients. -	-standard deviations	Automated audiometry yields a lower and more reliable hearing threshold than manual audiometry. Manual audiometry SD are about twice as much for automated testing. Test-retest reliability of automated audiometry indicated that the standard

		All subjects were noise exposed shipyard workers.	41/AR cushions. Audiometer: Manual- Madsen OB60, automated- Type Delmar 120.		Estimated standard deviations		deviations between the 5 successive tests had their lowest values for 1 kHz, increasing slowly towards lower and higher frequencies.
Erlandsson *et al.*	1979	***Accuracy :*** 115 subjects (230 ears), 25 to 63 years. ***Test-retest reliability:*** 10 subjects (20 ears). All subjects were noise exposed shipyard workers.	Diagnostic AC. Frequencies: .5, 1, 1.5, 2, 3, 4, 6, 8 kHz. Audiometer: Manual- Madsen OB60, automated- Type Delmar 120.	Method of adjustment- Békésy sweep frequency. Attenuation rate: 2.5 dB/s with a pulsed tone-presentation, sweep time from .25-1 kHz was 400s.	- Regression equation - Estimated standard deviations	- Average thresholds and standard deviations	Automated audiometry yields a lower and more reliable hearing threshold than manual audiometry. Test-retest reliability of automated audiometry indicated that the standard deviations between the 5 successive tests had their lowest values for 1 kHz, increasing slowly towards lower and higher frequencies.
Harris	1979	12 subjects (24 ears), 20 - 26 years old. Hearing status not indicated.	Diagnostic AC. Frequencies: .5, 1, 2, 3, 4, 6, 8 kHz. Audiometer: Manual- Tracor Model RA-115, automated- Self-recording- Tracor Model	Method of adjustment- Békésy fixed frequency. Tone pulse rate: 2.5pulses/sec was used; tones were presented for 30sec at each frequency. Attenuation rate of 5dB/sec in 0.25dB	- Average threshold and standard deviation - Average differences	-	Automated audiometry, utilizing the method of limits, indicated results that agree more with manual than automated audiometry utilizing the method of adjustment. At all frequencies, automated audiometry utilizing the method of

			ARJ-4C, Microprocessor-Tracor Moder RA-40 ** Two automated methods compared to manual testing.	steps. Thresholds read as the mid-point of the excursions at each frequency. Method of limits. An 800msec tone presented at random intervals of 1,2, sec. The Hughston-westlake method was utilized by the computer program.			adjustment showed lower thresholds than the other 2 tests. Automated audiometry utilizing the method of limits showed higher thresholds for all frequencies except 4 KHz, over manual audiometry. The two automated audiometry tests differed significantly at the 0.01 level in all frequencies. Time differences between each test were less than a minute.
Frampton & Counter	1989	42 subjects (84ears). All subjects were noise exposed adults.	Diagnostic AC testing. Frequencies: .5, 1, 2, 3, 4, 6, 8 kHz. Audiometer: Manual-Grason Stadler GSI 10, automated-Grason sStadler 1703 B	Method of Adjustment - Békésy sweep frequency. 7 frequency sweep with a pulsed tone mode.	- Average differences	-	Automated audiometry produced lower thresholds than manual testing. Automated audiometry is reliable and sensitive in the 'real world' setting. It allows large numbers of audiograms to be collected quickly by medical assistants with no training.
Lutman *et al.*	1989	120 subjects (240 ears), 40 – 65 years old.	Diagnostic AC thresholds. Frequencies: .5, 1, 2, 3, 4 kHz.	Method of adjustment-Békésy fixed frequency. Stimulus tone pulsed	- Average thresholds and	- Average differences and	Automated audiometry produced better results than manual audiometry. Overall automated

		Hearing status not indicated. Longitudinal study, subjects retest 2-3 years later.	Transducers: Manual- TDH-39P with MX 41/AR cushions Automatic- TDH-49P with MX -41/AR cushions	at a rate: 2.5pulses/sec, with duration of 200ms (3dB down points). The tracking procedure : 2dB step occurring every 2 pulses. Tracking at each frequency lasted 40sec, 50 levels were visited for each frequency.	standard deviations - Ranges of thresholds - Average difference	standard deviations - Standard of variance	audiometry was 4.4 dB better than manual audiometry; the difference was lower at .5 kHz and increased as the frequency increased. Test-retest reliability- manual audiometry indicated a worsening of hearing at .5,1, 2 kHz and an improvement at 4 kHz. Automated audiometry produced correlation coefficients which were statistically significant, however it suggests the shift is due to random measurement error rather than actual shifts in the threshold.
Fausti *et al.*	1990	20 subjects (40 ears), 18-25 years old. Normal hearing adults.	Diagnostic AC testing. Frequencies: .25, 0.5, 1, 2, 4, 8 kHz. Audiometer: Manual- GS1701, Automated- V320	Method of limits. V 320 Audiometer used, tones presented: 50% duty cycle, duration: 250 ms , rise-fall time: 25-50ms. Modified Hughson Westlake Ascending-descending audiometric test technique .	- Two-way analysis of variance with repeated measures on frequency and systems - Sheffé's	- Average absolute differences	No significant difference was noted between automated and manual testing over all test frequencies. Test-retest reliability: indicated no significant difference between the two tests conducted.

					to determine statistical significance.		
Picard *et al.*	1993	***3 groups used:*** 1) 420 subjects (840 ears), 18-64 years old. Noise exposed workers. 2) 36 elderly subjects (72 ears), 65-80 years old. Hearing status not indicated. 3) 12 subjects (24 ears), 7.5- 12 years old. Normal hearing children.	Diagnostic AC and BC testing with masking. Frequencies: AC- .5, 1, 2, 3, 4, 6 kHz. BC- .5 ,1, 2, 4 kHz. Audiometer: Automated- MADSEN, Model OB 822, manual not indicated.	Method of limits- BOBCAT. Tone duration of 700ms, 2s time interval. The computer program made use of the ascending- descending method (ISO 6189). ***Masking:*** Hood technique of masking used. AC Masking- 40dB gap between AC of test ear and BC of non-test ear. BC Masking- AC of the test ear exceeded the midline BC by more than 10dB.	- Reliability coefficients using Hoyt's solution. - Average thresholds and standard deviation - Dispersion relationships	-	Manual and automated procedures produce similar results, regardless of subject age, degree of hearing loss or nature of hearing loss. Mean thresholds across the populations comparable between automated and manual testing. Automated testing with the child population did not reveal consistent results when compared to manual audiometry, especially at 2 and 6 kHz. Automated testing takes longer to determine thresholds than manual testing (automated- 42 sec, manual- 34 sec). It was noted as population changed to 'difficult to test' patients (children) manual testing started to take more time. It was also noted that examiner takes shortcuts to

							obtain results but automated testing maintains rigid adherence to full procedure.
Fromby *et al.*	1996	***Accuracy:*** 101 subjects (202 ears), mean age of 43 years. Noise exposed workers. ***Test-retest reliability:*** 20 subjects (39 ears), Mean age of 43 years. Noise exposed workers.	Diagnostic AC testing. Frequencies: .25. .5, 1, 2, 3, 4, 6, 8 kHz Transducer: Telephonics TDH-39. Audiometer: Manual- Madsen, model OB822, automated- digital-to-analog converter (DAC) (TDT, model Quikki QDA1).	Method of limits- Maximum likelihood method was used (ML). Threshold for each frequency was measured in 15-trial block to yield 60% correct detection. On a trial, a 200msec pure-tone signal presented in a visually cued 200msec observation interval. Signals: 10-msec rise-fall times as part of the nominal durations. Subjects had 1000 msec to make a "yes-only" response which attenuated the signal level. If the subject did not respond during the 1000-msec response period, the computer assumed a "no" response for the trial, and the signal level was increased	- Average threshold - Standard error bars	- Average threshold - Standard error bars	Automated testing and manual testing yielded similar results. Threshold differences between the two methods were not statistically significant at any test frequency except .25 kHz, automated threshold was higher, but was within 3 dB of the threshold obtained manually. Test- retest reliability for automated testing: no significant test-retest differences at any test frequency. Additionally, manual testing took less time than automated testing (manual-3 min 46 sec, auto-6 min 43 sec).

				according to the ML algorithm.			
Margolis *et al.*	2007	***3 groups:*** 1) 120 subjects, 16-93 years old. Hearing status varied. 2) 8 subjects, 64- 85 years old. Varying degrees of hearing loss. 3) 6 subjects, 13- 86 years old. Varying degrees of hearing loss.	Diagnostic AC, BC and masking. Frequencies: not indicated. Transducers varied for different groups tested. ***Group 1 and 2:*** Manual- TDH-50, automated- prototype, non-occluding circumaural earphones ***Group 3:*** Manual- TDH-50 (not test ear occluded during BC testing), automated- insert earphones ER3A (both ears occluded during BC testing)	Method of limits- AMTAS. Tonal stimuli presented in a temporal observation interval that is visually marked for the listener, following the observation interval, the listener responds YES or NO by touching 'buttons' on a touchscreen monitor. The signal level is changed in an adaptive fashion to find the threshold of audibility. A threshold is obtained using a bracketing procedure. Masking noise presented to the non-test ear at levels that are selected to maximize the likelihood that neither under-masking nor over-masking will occur.	-Average absolute differences (QAave) -Regression coefficients -QUALIND -Correlation coefficients	-	The aim of this study was to develop a quality assessment method (QUALIND) based on a comparison of audiograms obtained utilizing automated (AMTAS) and manual testing. A predictive equation was derived from a multiple regression of a set of quantitative quality indicators on a measure of test accuracy, defined as the average absolute difference between automated and manually tested thresholds. For a large subject sample (n=120), a strong relationship was found between predicted and measured accuracy. The predictive equation was cross validated against two independent data sets. The results suggest that the predictions retain their accuracy for independent data sets if similar subjects

							and methods are employed, and that new predictive equations may be required for significant variations in test methodology. The method may be useful for automated test procedures when skilled professionals are not available to provide quality assurance.
Ho *et al*.	2009	***3 groups used:*** 1) 16 subjects (32 ears), 20- 80 years old. 2) 16 subjects (32 ears), 23-80 years old. 3)16 subjects (32 ears), 23-81 years old. Hearing status of all 3 groups unknown.	Diagnostic AC and BC testing with masking. Frequencies: AC- .25, .5, 1, 2, 3, 4, 6, 8 kHz. BC- .5,1, 2, 4 kHz Transducer: EAR 5A. Audiometer: Manual- not indicated, Automated- Otogram.	Method of limits- Otogram. Assesses AC and BC thresholds, administers masking when appropriate. Uses touch-screen technology programmed according to the Hughson-Westlake algorithm.	- Average Differences and standard deviations. - Levels of agreement were analysed and expressed by weighted κ coefficients, using SPSS version 15 and StatXact version 8.0.	- Average Differences and standard deviations. - Levels of agreement were analysed and expressed by weighted κ coefficients, using SPSS version	AC and BC results when tested with automated and manual testing produced similar results. AC thresholds when tested using automated and manual testing indicated 94% of automated thresholds that fell within 10 dB of those obtained manually and indicated 10 paired thresholds that fell within 15 dB of manual testing. BC unmasked thresholds showed that 93% of automated thresholds fell within 10 dB of each other and 96% fell within 15 dB of each other. BC masked thresholds between the 2 tests showed

						15 and StatXact version 8.0.	a lower level of agreement but still a good level of agreement. Test-retest reliability indicated good intrarater agreement between the automated and manual testing conducted.
McPherson *et al.*	2010	80 subjects (160 ears), 7-8 years old.	Screening AC tested. Frequencies: .5, 1, 2, 3, 4 kHz. Transducers: Manual- Circumaural ME-70 enclosures over TDH-39 supra-aural earphones. Automated- Circumaural headphone Ovann OV880V. Audiometer: Manual- Madsen Micromate, automated- IBM ThinkPad laptop PC, model T22.	Methods of adjustment. Békésy fixed frequency. Continues tones of 1 sec were presented in left ear at .5 kHz at 40 dB, and were raised or lowered in 3dB steps depending on response. Thereafter 1-4 kHz tested.	-X^2-test -Sensitivity or specificity analysis - Individual test results for each ear was compared using kappa values of agreement.	-	Automated screening procedure produced higher referral rate than manual screening (56% versus 13%). However, when .5 kHz was excluded from the data the referral rate between the two methods indicated no significant difference. The reason for .5 kHz producing errors could be as a result of ambient environmental noise and that automated audiometry started at .5 kHz and subjects were unfamiliar to test procedures.
Margolis *et al.*	2010	***Accuracy:*** 30 subjects (60 ears).	Diagnostic AC, BC and masking. Frequencies: AC-	Method of limits- AMTAS (see Margolis *et al*, 2007).	- Average differences	-	The differences between automated and manual testing were compared to

		Hearing status: 5 normal hearing subjects, 25 hearing loss subjects. ***Test-retest reliability:*** 18 subjects (36 ears). Hearing status: 3 normal hearing subjects, 15 sensorineural hearing loss subjects.	.25, .5, 1, 2, 3, 4, 6, 8 kHz. BC- .5, 1, 2, 4 kHz Transducer: AC- Sennheiser HDA200 BC manual- Radioear B71(mastoid placement) BC automated- B71 vibrator (forehead placement). Audiometer: Manual and automated- Madsen Conera.		-Average Absolute differences - Confidence intervals		differences obtained when the same subjects are tested manually by two audiologists. AC thresholds obtained by manual and automated testing indicated similar differences that were obtained when the same patients were tested manually by two audiologists. BC thresholds obtained with automated testing were lower than thresholds obtained with manual testing. The difference could be due to the placement of the bone conductor.
Swanepoel *et al.*	2010	***2 groups used:*** 1) 30 subjects (60 ears), 18- 31 years old. Normal hearing adults.	Diagnostic AC and masking. Frequencies: .125, .25, .5, 1, 2, 4, 8 kHz. Audiometer: Manual and automated- KUDUwave 5000.	Method of limits. Modified Hughson-Westlake method. Software presented a tone for 1.25s, subjects had to respond within 1.5 s before the next tone was presented.	- Absolute average differences and standard deviations - Two sided	- Absolute average differences and standard deviations	Thresholds determined by manual and automated testing were within 5 dB of each other, indicating no significant difference between the two test procedures, in both the hearing and hearing loss group.

		2) 8 subjects (16 ears), average age of 55 years old. Subjects had a sensorineural hearing loss ranging from mild to severe hearing loss.		Threshold was accepted if there was a minimum of 3 responses. Software automatically determined if contralateral masking was necessary and applied when required in an adaptive manner.	paired *t*-test - Pearson correlation coefficients	- Two sided paired *t*-test - Pearson correlation coefficients	Test-retest reliability of automated testing indicated reliability equivalent to that of manual testing. Additionally, both manual and automated testing took more or less the same time to administer (manual- 7.2-7.7 min, automated- 7.2-7.4 min).
Ishak *et al.*	2011	***Accuracy:*** 13 subjects (13 ears), a8-60 years old. Normal hearing adults. ***Test-retest reliability:*** 21 subjects (21 ears), 18-60 years old. Normal hearing adults.	Diagnostic AC testing. Frequencies: .25, .5, .75, 1, 1.5, 2, 3, 4, 6, 8 kHz. Audiometer: Manual and automated-Essilor Audioscan system. ** Test-retest reliability was determined by testing subjects 4 times with each test producer.	Method of adjustment-Békésy sweep frequency and Audioscan. ***Békésy:*** Sweep rate: 15 s per octave, pulse rate: 2.5 pulses/s, attention rate: 2.5dB/s was used. Hearing thresholds determined by calculating averaged values of three consecutive audiometric data obtained around each octave or half-octave frequencies.	- Repeated measures ANOVA - Contrasts analysis to compare mean thresholds.	- Thresholds from each test session were subtracted - Variance of hearing threshold ($\sigma 2$)	The results showed that the thresholds obtained with Békésy testing were significantly better than those obtained from the manual testing at most frequencies. Audioscan produces better thresholds than Békésy, showing no significant differences in hearing thresholds at frequencies from .5 kHz- 4 kHz. Hearing thresholds obtained from Audioscan were significantly poorer than manual testing at frequencies of .25, 6 and 8 kHz. This was probably due

				These values were rounded to the nearest 5dB for the analysis. ***Audioscan:*** Sweep rate: 15sec/octave, tones swept 1- 8 kHz, back to 1 kHz and swept again from 1 kHz to .25 Hz. A straight line was produced when the subjects pressed the response button. The level was then increased by 5dB at frequencies to which the subjects did not respond.			to the threshold seeking procedure, which does not allow the intensity level to go either higher or lower than the current screening intensity level. High test-retest reliability for manual and audioscan testing, however, Békésy testing indicated poor test-retest reliability.
Margolis *et al.*	2011	***2 groups:*** 1) 68 subjects (136 ears), 4- 8 years old (1 group of 4-5 year olds and another group of 6-8 year olds). Normal hearing	Diagnostic AC testing. Frequencies: .5, 1, 2, 4, 8 kHz. Transducers: Automated- HDA 200 Manual- TDH-50. Audiometer: Manual and automated (children)- Benson CCA-100	Method of limits- AMTAS was used for the adult group (see Margolis *et al*, 2007). KIDTAS was used for the child population. It differed from AMTAS, used a smiley and sad face and a visual reinforcement picture for a correct response. Additionally, QUALIND was used. QUALIND is	- Average absolute average difference and standard deviation	-	The differences obtained between automated testing (AMTAS/KIDTAS) and manual testing produces thresholds with variability that is comparable to thresholds obtained using manual testing by two audiologists, only if QUALIND identifies and excludes 'poor' audiograms. No significant differences between manual and

		children. 2) 15 subjects , Adults. Hearing status: 11 normal hearing, 1 unilateral hearing loss, 3 mild-to-moderate bilateral hearing loss subjects.	Mini. Manua (adults)l- Grason Stadler, automated- Benson CCA. **Different transducers were only used in the adult population.	a method for estimating accuracy by tracking variables that are known to predict agreement between automated and manual thresholds, and calculating the predicted average absolute difference with a formula derived from a regression analysis of the relationship between the quality indicators and the measured average absolute differences. The strength of the regression coefficient indicates the degree to which accuracy can be predicted by QUALIND.			automated thresholds were noted when using different earphones in the adult subjects.
Margolis & Moore	2011	13 subjects (19 ears), 21-65 years old. All subjects had a sensorineural	Diagnostic AC, BC and masking. Frequencies: .25, .5, 1, 2, 4, 8 kHz. Audiometer: Manual- Grason Stadler	Method of limits- AMTAS (see Margolis *et al*, 2007).	- Average thresholds -Average differences -Average absolute	-	Automated testing produced thresholds similar to those obtained by manual testing results. Automated thresholds were higher than those obtained manual by 7 dB at .25, .5, 1, 2 kHz, with

		hearing loss.	GSI 61, Automated- Madsen Aurical.		differences -Analysis of variance (ANOVA)		smaller differences at higher frequencies. According to Margolis et al (2010) results between manual and automated testing should be similar, thus it was concluded by this study that the difference noted between the two test results was due to the use of different earphones.

Supplemental Digital Content 3. Table, Statistical measures of accuracy and test-retest reliability employed in systematic review reports (n=29)

Type of analysis	**Number of studies**
Accuracy (threshold comparison with manual audiometry)	
Average differences and standard deviation	11
Average thresholds and standard deviation	11
Absolute average differences and standard deviation	6
t-Test	4
Linear regression and correlation coefficients	4
Pearsons product	3
Standard deviations only	3
ANOVA analysis	2
Average deviation	1
Error analysis	1
Contrast analysis	1
X^2 Test	1
Sensitivity and specificity analysis	1
Comparison of Kappa values of agreement	1
Standard error bars	1
Test of significance	1
Within subject variability test	1
F-ratio	1
Two way analysis of variance	1
Reliability coefficients- Hoyts solution	1
Sheffe’s test of statistical significance	1
Dispersion relationships	1
K-coefficients	1
Confidence intervals	1
Estimation of asymptomatic data	1
Test-retest reliability	
Average differences and standard deviation	4
Average thresholds and standard deviation	3
Absolute average differences and standard deviation	2
t-test	2
Pearson Product moment correlation coefficients	2
Standard deviation	1
Standard of variance	1
Standard error bars	1
k-coefficients	1
Repeated ANOVA	1
Variance of hearing threshold (σ^2)	1

Supplementary Digital Content 4. Table (Summary of data included in meta-analysis, test-retest reliability)

	Author	Year	Number of ears	Statistical analysis	Frequencies (Hz)									
					125	**250**	**500**	**1000**	**2000**	**3000**	**4000**	**6000**	**8000**	**All**
MANUAL TESTING	**AVEARAGE DIFFERENCES**													
	Burns & Hichcliffe.	1957	40	Average differences	-	-	1.0	2.2	1.5	2.0	1.4	-1.7	-	1.0
				Standard deviation	-	-	4.9	4.2	4.7	4.7	4.7	7.6	-	5.1
	Lutman et al.	1989	460	Average differences	-	-	2.4	2.1	1.4	-	-0.5	-	-	1.3
				Standard deviation	-	-	6.9	4.8	5	-	7.1	-	-	6.1
	Ho et al.	2009	32	Average differences	-	-	-	-	-	-	-	-	-	1.8
				Standard deviation	-	-	-	-	-	-	-	-	-	6.6
	ABSOLUET AVERAGE DIFFERENCES													
	Fausti et al.	1990	20	Absolute Average difference	-	2.3	2	1.8	1.8	-	2.3	-	2.3	2.1
	Swanepoel et al.	2010	60	Absolute Average difference	4.8	3.8	3.3	3.7	3.0	-	3.0	-	3.3	3.6
				Standard deviations	5	3.7	3.7	3.4	3.6	-	3.5	-	4.3	3.9
AUTOMATED TESTING	**AVERAGE DIFFERENCES**													
	Burns & Hichcliffe.	1957	40	Average differences	-	-	1.0	2.0	1.0	2.1	1.2	1.7	-	1.5
				Standard deviation	-	-	6.4	5.2	3.8	6.2	6.4	10.4	-	6.4
	Lutman et al.	1989	460	Average differences	-	-	0.2	-1.3	-0.1	-	0.6	-	-	0.1
				Standard deviation	-	-	7.2	6.9	6.6	-	7.2	-	-	7.0
	Ho et al.	2009	32	Average differences	-	-	-	-	-	-	-	-	-	0.3
				Standard deviation	-	-	-	-	-	-	-	-	-	5.9
	ABSOLUTE AVERAGE DIFFERENCES													
	Fausti et al.	1990	20	Absolute Average difference	-	2.3	1.8	1.8	1.8	-	2.0	-	1.5	1.9
	Swanepoel et al.	2010	60	Absolute Average difference	4.9	3.8	3.2	2.8	2.8	-	2.4	-	2.2	3.2
				Standard deviations	4.8	3.5	3.6	3.2	4.1	-	3.0	-	3.2	3.8

Supplementary Digital Content 5. Table (Summary of reports included in the meta-analysis, accuracy)

	Author	Year	Number of ears	Statistical analysis	Frequencies									
					125	250	500	1000	2000	3000	4000	6000	8000	All
METHOD OF ADJUSTMENTS	**AVERAGE DIFFERENCES**													
	Burns & Hichcliffe	1957	40	Average differences	-	-	-1.1	3.2	1.3	1.6	1.2	-0.5	-	1.0
				Standard deviation	-	-	5.5	5.1	4.7	6.0	7.1	9.2	-	6.3
	Knight J.J.	1965	66	Average differences	-	-	-0.3	1.0	1.5	1.1	1.3	-0.1	-	0.8
				Standard deviation	-	-	4.2	4.9	4.9	4.9	3.8	5.3	-	4.7
	Delany et al.	1966	66	Average differences	-	-	1.2	-0.8	-0.9	-1.4	-1.5	-3.3	-	-1.1
	Jokinen K	1969	30	Average differences	5.1	2.1	-0.6	-0.6	2.3	4.9	-2.5	-0.6	-2.7	0.8
				Standard deviation	7.0	6.0	5.0	5.6	7.0	4.4	5.4	6.6	6.4	5.9
			30	Average differences	-1.7	-3.1	-3.3	-2.8	-0.6	4.1	-4.4	-4.5	-5.5	-2.4
				Standard deviation	8.1	6.1	4.2	4.9	5.2	5.6	5.3	6.4	7.9	6.0
			17	Average differences	-8.4	-7.7	-5.4	-7.8	-4.4	2.6	-10	-5.6	-3.9	-5.6
				Standard deviation	6.3	6.4	6.4	3.8	6.2	7.4	5.6	6.7	6.4	6.1
			39	Average differences	-5.2	-7.0	-5.6	-6.4	-4.1	-1.0	-12.6	-8.1	-12.3	-6.9
				Standard deviation	9.9	9.5	6.7	7.0	7.6	7.4	7.0	9.0	11.5	8.4
			30	Average differences	4.3	0.3	-2.0	-0.4	1.2	5.6	-2.0	-0.7	-3.6	0.3
				Standard deviation	7.7	7.2	5.5	5.7	6.8	5.2	7.0	8.1	8.5	6.9
			30	Average differences	-4.0	-5.7	-4.1	-2.8	-1.5	3.2	-5.1	-6.1	-4.8	-3.4
				Standard deviation	8.9	6.1	4.0	5.2	5.1	6.8	5.9	7.4	8.2	6.4
			17	Average differences	-6.4	-6.0	-2.1	-3.0	0.9	8.5	-3.6	1.9	2.0	-0.9
				Standard deviation	6.7	6.7	8.4	5.0	7.3	6.8	4.8	7.2	6.1	6.6
			39	Average differences	-3.1	-4.7	-0.2	-1.7	-0.5	-4.2	-4.9	-2.8	-10.0	-3.6
				Standard deviation	11.4	10.0	7.9	7.7	6.6	8.1	8.2	10.1	14.6	9.4
	Robinson & Whittle	1973	128	Average differences	-	0.4	2.9	1.5	2.8	-	2.7	4.2	2.1	2.4
				Standard deviation	-	5.9	4.4	4.1	4.3	-	5.3	8.2	8.5	5.8
	Harris	1979	24	Average differences	-	-	-2.1	-4.0	-5.6	-4.0	-9.0	-1.0	-2.9	-4.1

	Lutman et al.	1989	240	Average differences	-	-	3.0	2.8	6.4	-	5.3	-	-	4.4
				Standard deviation	-	-	5.8	5.6	5.2	-	6.1	-	-	5.8
METHOD OF LIMITS	**AVERAGE DIFFERENCES**													
	Harris	1979	24	Average differences			-3.5	-2.3	-1.3	-2.9	3.8	-4.4	-0.2	-1.5
				Standard deviation	-	-	6.4	5.2	3.8	6.2	6.4	10.4	-	6.4
	Ho et al.	2009	32	Average differences	-	-	-	-	-	-	-	-	-	0.76
				Standard deviation	-	-	-	-	-	-	-	-	-	5.7
	Margolis et al.	2010	60	Average differences	-	-0.4	0.4	1.5	1.4	-	0.1	-	-2.3	0.1
				Standard deviation	-	5.1	4.4	5.3	5.8	-	4.9	-	7.0	5.4
	ABSOLUTE AVERAGE DIFFERENCES													
	Sparks	1972	15	Absolute Average differences	-	-	-	-	-	-	-	-	-	4.5
	Swanepoel et al.	2010	60	Absolute Average differences	4.8	3.8	3.8	3.7	3.2	-	2.9	-	2.8	3.6
				Standard deviation	4.1	3.4	4.5	3.7	3.3	-	3.5	-	4.5	3.9
			60	Absolute Average differences	4.2	3.8	3.6	3.8	3.3	-	2.2	-	2.3	3.3
				Standard deviation	4.2	3.5	4.5	3.1	4.0	-	3.0	-	3.6	3.8
			16	Absolute Average differences	2.3	3.3	2.2	2.2	2.2	-	2.8	-	1.4	2.4
				Standard deviation	3.2	2.4	2.6	2.6	2.6	-	3.1	-	3.1	2.8
	Margolis et al.	2010	60	Absolute Average differences	-	3.2	3.0	3.3	4.0	-	3.7	-	4.5	3.6
				Standard deviation	-	4.0	3.2	4.4	4.4	-	3.2	-	5.8	4.2
	Margolis et al.	2011	15	Absolute Average differences	-	-	-	-	-	-	-	-	-	3.9
				Standard deviation	-	-	-	-	-	-	-	-	-	1.7