Clinical validation of automated audiometry with continuous noise-monitoring in a clinically heterogeneous population outside a sound-treated environment

Christopher G. Brennan-Jones [a b]

Robert H. Eikelboom [a b c]

De Wet Swanepoel [a b c]

Peter L. Friedland [a b d e]

Marcus D. Atlas [a b e]

a) Ear Science Institute Australia, Subiaco, Perth, Australia
b) Ear Sciences Centre, School of Surgery, The University of Western Australia, Crawley, Australia
c) Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa
d) School of Medicine, University of Notre Dame, Fremantle, Australia
e) Department of Otolaryngology, Head, Neck and Skull Base Surgery, Sir Charles Gairdner Hospital, Nedlands, Australia

ABBREVIATIONS:

4FA: Four-frequency average (500, 1000, 2000 and 4000 Hz)

ISO: International Standards Organisation

SD: standard deviation

CONTACT INFORMATION:

Christopher Brennan-Jones
Ear Science Institute Australia
Suite 1, Level 2, 1 Salvado Road, Subiaco, Western Australia 6008
Tel: +61 (8) 6380 4900 Fax: +61 (8) 6380 4901
chris.brennan-jones@earscience.org.au

ABSTRACT:

Objective: Examine the accuracy of automated audiometry in a clinically heterogeneous population of adults using the KUDUwave automated audiometer.
Design: Prospective accuracy study. Manual audiometry was performed in a sound-treated room and automated audiometry was not conducted in a sound-treated environment.
Study Sample: 42 consecutively recruited participants from a tertiary otolaryngology department in Western Australia.
Results: Absolute mean differences ranged between 5.12 – 9.68 dB (air-conduction) and 8.26 – 15.00 dB (bone-conduction). 86.5% of manual and automated 4FAs were within 10 dB (i.e. ±5 dB); 94.8% were within 15 dB. However, there were significant (p<0.05) differences between automated and manual audiometry at 0.25, 0.5, 1 and 2 kHz (air-conduction) and 0.5 and 1 kHz (bone-conduction). The effect of age (≥55 years) on accuracy (p = 0.014) was not significant on linear regression (p>0.05; R^2 = 0.11). The presence of a hearing loss (better ear ≥26 dB) did not significantly affect accuracy (p = 0.604; air-conduction), (p = 0.218; bone-conduction).
Conclusions: This study provides clinical validation of automated audiometry using the KUDUwave in a clinically heterogeneous population, without the use of a sound-treated environment. Whilst threshold variations were statistically significant, future research is needed to ascertain the clinical significance of such variation.

KEYWORDS: automated audiometry, audiometry, hearing loss, teleaudiology, KUDUwave

INTRODUCTION

Assessment of hearing sensitivity thresholds is one of the key tests conducted by audiologists. The methods of assessment are well defined by the modified Hughson-Westlake protocols ISO 8253-1:2010 (ISO, 2010). In a standard manual audiometry procedure, frequency-specific sound stimuli are presented to a patient and the hearing level of the stimuli is adjusted, either decreasing or increasing, according to the patient's response or lack of response, respectively. This method is also termed a 'method of limits' approach and is performed according to ISO 8253-1:2010 standards on equipment calibrated to ISO389-1:1998 (ISO, 1998) standards. In the past decade there has been an increasing interest in

systems that automate these procedures (Eikelboom et al., 2013; Ho et al., 2009; Margolis et al., 2010; Swanepoel et al., 2010).

Automated audiometers are not new. Georg von Bekesy (von Bekesy, 1947) was the first to describe a self-recording threshold audiometer which automatically increased or decreased sound level whilst sweeping a specified frequency test range. Whist this technique is still in use by some, the Hughson-Westlake method is now the most common technique for performing audiometry. A number of automated audiometry systems have implemented computerised versions of the Hughson-Westlake procedure, with the first reports of this method of automation appearing more than four decades ago (Sparks, 1972).

Following the successful clinical validation of a number of automated audiometers (Eikelboom et al., 2013; Margolis et al., 2011; Swanepoel et al., 2010), and a systematic review of their accuracy (Mahomed et al., 2013), the potential scope of these devices has expanded to include full diagnostic hearing assessments for adults, encompassing masked and not-masked air and bone-conduction thresholds.

In the meta-analysis conducted by Mahomed et al. (2013), automated audiometry showed comparable accuracy to manual audiometry, with overall average differences of 0.4 dB (6.1 SD). However, the authors noted that there was limited data on automated bone conduction audiometry and patients with different types and degrees of hearing loss. A number of studies included in the systematic review reported the accuracy of automated audiometry on participants with normal hearing only.

The inclusion of participants without hearing loss introduces significant bias into accuracy studies (Rutjes et al., 2006). The potential for bias is clear; normal hearing patients are known to have hearing within a certain range, thereby limiting the potential range of variation between two methods of assessment. To limit bias it is therefore essential that the accuracy of automated audiometry be examined in a population that is likely to include participants with a range of hearing threshold levels, but who are not pre-selected according to hearing status or level of impairment. The exclusion of patients with known conductive hearing impairments (e.g. Storey et al., 2014) is also a source of potential bias. These patients represent a significant part of the clinical population and it is just as important to have accuracy estimates in such cases as for patients with sensorineural hearing losses.

The inclusion of participants with normal hearing threshold levels has been a necessary and valuable step in establishing the accuracy of automated audiometry. However, the development of studies that reduce bias by examining participants from a true clinical population will provide the most valid estimates of the accuracy of automated audiometry in practice.

One of the appeals of automated audiometry over conventional manual audiometry is its potential application in teleaudiology and its use in situations where sound treated rooms are unavailable or inaccessible. Recent reports have emphasized the global shortage of audiological services, and highlighted that these shortages are not exclusive to low and middle-income countries (Windmill & Freeman, 2013). It has also been reported that patients living in rural and remote areas of developed countries are more likely to present to primary care with a self-reported hearing loss (Brennan-Jones et al., 2015). The ability to provide automated audiometric testing in the absence of a sound-treated environment has a great potential to increase service provision to low and middle-income countries, and rural and remote areas of high-income countries that do not have these facilities. At least two of the contemporary clinically available automation-capable audiometers use audiocups to provide attenuation from environmental sounds (Margolis et al., 2010; Swanepoel & Biagio, 2011), and studies have demonstrated their potential feasibility in environments that are not sound-treated (Eikelboom et al., 2013; Maclennan-Smith et al., 2013).

The device used in this study (KUDUwave 5000) has previously been validated in an environment that is not sound-treated using its manual-mode (Maclennan-Smith et al., 2013), and in a controlled noise environment in automated-mode (Storey et al., 2014). The present study therefore aims to address a gap in the evidence-base by combining automated testing in an uncontrolled environment that is not sound-treated, using an unselected clinical population of patients attending otolaryngology and audiology appointments at a tertiary public hospital. The potential influence of age and presence of hearing loss will also be examined to investigate the influence of patient-related variables on accuracy of automated audiometry.

METHODS

Participants

42 participants (20 male, 22 female) were recruited from a publicly funded combined otolaryngology and audiology clinic at Sir Charles Gairdner Hospital, Perth, Western Australia. Attendance at the clinic was free at the point of service for patients. Inclusion criteria were: 18 years or over, no known cognitive disorder, English spoken as a first language, both ears suitable for hearing assessment. Ethics approval was granted by the University of Western Australia Human Research Ethics Committee (Reference: RA/4/1/4877).

Participant sampling and recruitment

Patient recruitment was by consecutive series, with all patients attending the clinic offered enrolment in the study, subject to inclusion criteria. Recruitment was not based on presenting symptoms (except where they contra-indicated audiological assessment) or results from previous audiometry. No incentives were given to participants involved in the study.

Data collection

Data collection was prospectively designed. The order of test administration was not randomised. Five patients had the index test administered prior to the reference test, and all other participants (n=37) received the reference test first.

Test methods

Reference test: Manual audiometry

Manual audiometry is considered the gold standard assessment of hearing thresholds in adults and children over five years of age and therefore served as the reference test for this study (ASHA, 2004). The Hughson-Westlake method (i.e. ascending method according to ISO8253-1:2010), or adaptations of this method according to local protocols, is typically used when determining hearing thresholds with manual audiometry. Manual audiometry was conducted within a sound-treated room (mean ambient noise level 37 dBA) using Acoustic Analyser AA30 audiometer (Starkey Hearing Technologies; Minnesota), calibrated to ISO389-1:1998 and TDH-39P (Telephonics; North Carolina) supra-aural headphones and Radioear B-71 bone-conductor (Radioear Corp.; Pennsylvania), calibrated to ISO389-3:1994.

The bone-conductor was placed on the patient's mastoid for manual testing. Patient history, otoscopy and tympanometry using a GSI 38 Auto Tymp (Grason-Stradler; Minnesota) preceded audiometry testing.

Index test: Automated audiometry

Automated audiometry was conducted using the KUDUwave (eMoyoDotNet; Pretoria, South Africa) a mobile Type 2B screening, diagnostic and clinical audiometer (IEC 60645-1/2) using the ascending method according to ISO8253-1:2010. A key advantage of the KUDUwave audiometer is its double attenuation via use of insert earphones and circumaural earcups and its use of continual noise monitoring, which pauses audiometric testing if ambient noise levels exceed prescribed limits, enabling accurate testing down to 0 dB with an ambient noise level of up to 59 dB SPL. The mean ambient noise level when there was no outpatient clinic in progress was measured at 46 dBA. Placing insert earphones down to the bony part of the ear canal also reduces the occlusion effect allowing for bone-conduction evaluation with occluded ears using insert earphones (Slevin et al., 2000; Swanepoel & Biagio, 2011). However, not removing the insert earphone is a limitation to the technique as insertion down to the bony portion of the ear canal cannot be confirmed or guaranteed. If the contralateral insert earphone is removed, this can adjust for the occlusion effect, however it also means losing some attenuation that is added by the insert. The insert earphone frequency response approximated that of the ER3A within 1 dB across test frequencies. This allowed for the use of the international insert earphone standard (ISO 389-2: 1994) for calibration. These features make the KUDUwave especially suited for use without a sound-treated environment, making it appropriate for use in rural, remote or community settings, where the availability of a sound-treated environment for testing is unlikely. The audiometry procedures were automated and recorded on a laptop using the eMOYO (v3.6.7) interface developed by eMoyoDotNet. Whenever the difference between the air conduction thresholds in the test and non-test ear was 75 dB or more at frequencies ≤1000Hz and 50 dB or more at frequencies >1000 Hz, air conduction thresholds were masked according to current guidance Munro & Agnew, 1999; Edwards, 2010). A masking level of 30 dB above the air conduction threshold of the non-test ear was used. Bone conduction thresholds (using a B-71 bone oscillator (Kimmetrics, Smithsburg, USA)) were determined with continuous masking in the contralateral ear. A continuous masking level of 20 dB above the air conduction threshold of the non-test ear was used. Testing took place in a quiet room that was not sound treated (mean ambient noise level when there was no outpatient clinic in progress was 46 dBA). The

researcher gave standard instructions, placed the insert earphones, bone-conductor and headset on the participant and monitored the progress of the test in case of malfunction or patient discomfort.

Definitions

Hearing thresholds were presented in dB hearing level (dBHL). Participants were tested at air conduction frequencies of 250, 500, 1000, 2000, 4000 and 8000 Hz and bone-conduction frequencies of 500, 1000, 2000, 4000 Hz for both the reference test and index test. The audiologists administering the reference test obtained hearing thresholds at additional frequencies for participants as clinically indicated; however, these additional thresholds were not examined in this analysis as in most cases no corresponding threshold from the index test was available.

The index test had lower maximum sound level limits compared to the reference test (KUDUwave limits for air conduction were 95 dB for 0.25 kHz, 100 dB for 0.5, 1, 2 and 4 kHz and 90 dB at 8 kHz; for bone-conduction 55 dB at 0.5 kHz and 70 dB at 1-4 kHz). In cases where no response was recorded because the index test reached its maximum testable limits at a lower level than the reference test, the hearing threshold level of the reference test was corrected to the maximum output level of the index test.

Test procedure

The reference test (manual audiometry) was administered by tertiary-qualified clinical audiologists (five clinical audiologists were involved in administering the reference test throughout the study). The audiologists were all registered with the Audiological Society of Australia. Interpretation of the reference test was conducted by the clinical audiologist responsible for the patient's care. Automated audiometry was administered by researchers involved in the project. The time interval between the reference test and the index test being conducted was less than 60 minutes for all participants, as patients proceeded directly to the next test, or after a short break if requested or required.

Blinding

The audiologist administering the reference test was blinded to the results of the index test. The researcher administering the index test was not blinded to the results of the reference test as the index test was automated and therefore could not influence the results. Other

information available to the audiologist and researcher were a clinical history, and a combination of tympanometry, acoustic reflexes and speech recognition threshold testing scores, as conducted by the clinical audiologist.

Statistical methods

The validation analysis used air-conduction thresholds for 250, 500, 1000, 2000, 4000 and 8000 Hz and bone-conduction thresholds of 500, 1000, 2000 and 4000 Hz for both manual and automated audiometry. Mean and standard deviations were calculated for each frequency as well as real and absolute mean differences between the reference and index test hearing thresholds. Absolute mean differences are a preferable measure compared to real mean differences as absolute differences can account for positive and negative variation, whereas positive and negative variance can cancel each other out when using real mean differences (Eikelboom et al., 2013). Reference test (manual audiometry) thresholds were subtracted from index test (automated audiometry) thresholds to calculate the difference, in keeping with methodologies from similar studies (Eikelboom et al., 2013; Swanepoel et al., 2010). A paired-samples t-test and ANOVA with Bonferroni's correction applied were used to calculate significant differences in hearing thresholds, an independent samples t-test was used for age (using 55 years of age as an arbitrary cut-point) and presence of hearing loss analysis (using better ear hearing of 4FA ≥26 dB as a cut-point). Simple linear regression was also used for the analysis of age on accuracy of automated audiometry. Excel 2010 (Microsoft®, Washington) and SPSS v21 (New York: IBM Corp) were used for the analysis.

RESULTS

Participants

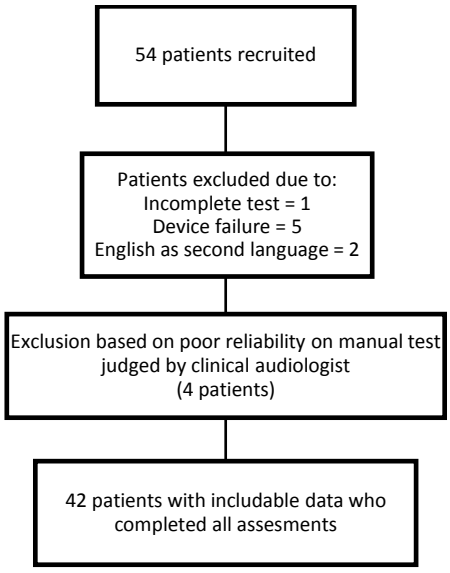
The mean age of participants was 49.93 years (SD = 17.35, range of 19.33 to 92.55 years). Patients presented with a diverse range of clinical conditions, symptoms and co-morbidities, including but not limited to: sensorineural hearing loss, tinnitus, conductive hearing loss, otosclerosis, otitis media, acoustic neuroma, Meniere's disease, benign paroxysmal positional vertigo, perforated tympanic membrane, Eustachian tube dysfunction, ototoxic hearing loss, skull base fracture and unilateral hearing loss, as well as pre-surgical assessment and post-surgical assessments. Hearing loss was not always the primary complaint for participants and many had more than one ear or hearing related symptom at the time of testing. The patients had a wide range of hearing losses (Table 1). Patients who had incomplete assessment data on

Table 1: Ear specific level of hearing loss and cumulative percentage differences.

Hearing level (dB)	Total	%	c.%[#]
0 to 10	8	9.5	9.5
10-20	22	26.2	35.7
20-30	17	20.2	55.9
30-40	15	17.9	73.8
40-50	7	8.3	82.1
50-60	4	4.8	86.9
60-70	8	9.5	96.4
70-80	1	1.2	97.6
>80	2	2.4	100.0
Totals	84		

either manual or automated audiometry (n = 8) or had reliability questioned by the clinical audiologist (n = 4) were not included in the analysis (Figure 1).

Figure 1: Flow diagram of patient recruitment and drop-outs



Accuracy of KUDUwave automated audiometry

Summary tables of mean absolute and real differences are presented in Tables 2 and 3 respectively. The range of absolute mean differences for all air-conduction thresholds was 5.12 to 9.68 dB (SDs 5.17 to 9.59 dB) and for bone-conduction was 8.26 to 15.00 dB (SDs 7.44 to 10.58) (Table 2). The range of absolute mean differences for 4FA (500, 1000, 2000

Table 2: Absolute mean differences, standard deviation for 42 participants for automated audiometry compared to manual audiometry.

Hearing Thresholds			Frequency (HZ)					
			250	500	1000	2000	4000	8000
Air	Right	AMD	7.86	6.43	6.19	6.19	4.76	8.41
		SD	6.26	5.98	5.93	5.39	5.17	9.84
	Left	AMD	9.17*	7.02*	6.79*	6.67*	7.02*	6.79*
		SD	8.69	6.54	5.50	6.31	6.54	7.23
Bone	Right	AMD	-	10.83*	13.45*	9.76*	8.93*	-
		SD	-	9.23	8.59	7.65	8.94	-
	Left	AMD	-	10.60*	14.40*	10.36*	8.21*	-
		SD	-	7.42	9.77	10.20	9.36	-

AMD: Absolute mean difference (in dB); SD: Standard deviation; - not measured
*indicates a significant ($p<0.05$) difference in threshold accuracy according to a one-way ANOVA with Bonferroni's correction applied.

Table 3: Real mean differences and standard deviation and *p* value for pair-wise t-test for 42 participants for index test (automated audiometry) compared to reference test (manual audiometry)

Hearing Thresholds			Frequency (Hz)											
			250	*p*	500	*p*	1000	*p*	2000	*p*	4000	*p*	8000	*p*
Air	Right	RMD	3.33	.029	4.76	<.001	5.00	<.001	4.29	<.001	-1.90	.077	-2.74	.165
		SD	9.54	-	7.40	-	6.98	-	7.03	-	6.80	-	12.55	-
	Left	RMD	5.60	.003	6.07	<.001	5.36	<.001	3.57	<.001	-1.31	.380	-1.55	.314
		SD	11.38	-	7.45	-	6.93	-	8.50	-	9.57	-	9.85	-
Bone	Right	RMD	-	-	6.79	.001	12.74	<.001	3.33	.080	2.02	.302	-	-
		SD	-	-	12.58	-	9.64	-	12.03	-	12.55	-	-	-
	Left	RMD	-	-	8.93	<.001	13.69	<.001	3.21	.289	2.26	.479	-	-
		SD	-	-	9.41	-	10.77	-	14.26	-	12.31	-	-	-

RMD: Real mean difference (in dB); SD: Standard deviation; - not measured

and 4000 Hz) air-conduction thresholds was 5.12 to 6.98 dB (SDs 5.17 to 6.46 dB) and for all bone-conduction frequencies was 8.26 to 15.00 dB (SDs 7.44 to 10.58) (Table 2). The percentage of 4FA automated air-conduction thresholds falling within an absolute mean difference of 5dB of the reference test was 67.8%, within 10 dB was 86.5% and within 15 dB were 94.8% of hearing thresholds (Table 4).

Analysis of variance (ANOVA) with Bonferroni's correction applied was used to compare the mean difference between manual and automated audiometry and these results are provided in Table 2. For air-conduction audiometry, the mean differences in hearing thresholds determined by manual and automated audiometry were not significantly different ($p>0.05$) in the right ear, but were significantly different across all frequencies in the left ear.

Table 4: Difference distribution of air-conduction hearing thresholds for mid-frequencies (500, 1000, 2000 and 4000 Hz).

dB Diff	500 Hz		1000 Hz		2000 Hz		4000 Hz	
	right	left	right	left	right	left	right	left
0	11	11	10	7	11	14	16	11
5	18	13	21	21	18	10	16	20
10	6	13	6	10	6	11	7	5
15	5	3	3	1	6	5	1	3
20	1	0	0	2	1	1	2	2
25	1	1	2	1	0	1	0	0
30	0	1	0	0	0	0	0	1
35	0	0	0	0	0	0	0	0
Totals	42	42	42	42	42	42	42	42
%within 5 dB	69.0	57.1	73.8	66.7	69.0	57.1	76.2	73.8
Total % ≤5 dB	67.8							
%within 10 dB	83.3	88.1	88.1	87.5	83.3	83.3	92.9	85.7
Total % ≤10 dB	86.5							
%within 15 dB	98.4	95.2	95.2	89.5	97.6	95.2	94.9	92.8
Total % ≤15 dB	94.8							

For bone-conduction audiometry, the mean differences in hearing thresholds determined by manual and automated audiometry were significantly different at all frequencies.

Pair-wise comparisons for the real mean difference between manual and automated audiometry are provided in Table 3. For air-conduction audiometry, the mean differences in hearing thresholds determined by manual and automated audiometry for the frequencies 4000 and 8000 Hz were not significantly different ($p>0.05$). For bone-conduction audiometry, the mean differences in hearing thresholds determined by manual and automated audiometry for the frequencies 2000 and 4000 Hz were also significantly associated bilaterally ($p>0.05$). All other pair-wise comparisons in Table 3 were significantly different ($p<0.05$).

Age differences and presence of hearing loss

To calculate the effect of age differences and the presence of a hearing loss on the accuracy of automated audiometry each participant had an individual average absolute mean difference (individual AMD) calculated using frequencies tested for air-conduction (250–8000 Hz) and bone-conduction (500–4000 Hz) on both manual and automated audiometry. This created a summary score of audiometric variance between manual and automated audiometry for each participant to enable analysis. 16 participants had a hearing loss of ≥26 dB in both ears. Bilateral hearing loss was not significantly associated with increased variation in individual AMDs in thresholds between automated and manual audiometry for air-conduction $t(40) = -$

0.523; $p = 0.604$ or bone-conduction $t(40) = 1.251$; $p = 0.218$. The mean age of participants with a bilateral hearing loss was 56.18 (SD 18.93) compared to 46.08 (SD 14.43) for those without a hearing loss, this difference was marginally not significant (β = 10.10 [95%CI -0.706, 20.913]; $p = 0.066$).

A statistically significant difference was found between age (≥55 years, n = 14) and hearing threshold accuracy for air-conduction $t(40) = 1.599$; $p = 0.014$, but not for bone-conduction $t(40) = 1.334$; $p = 0.190$. However, whilst linear regression showed a slight upward trend of increased individual AMDs (i.e. decreased accuracy) with age ($R^2 = 0.11$), the relationship was not statistically significant when analysed independently (β = 0.019 [95%CI -0.037, 0.074]; $p = 0.504$) (See Figure 2) or once adjusted for presence of hearing loss (β = 0.025 [95%CI -0.034, 0.083]; $p = 0.398$).

Excluded participants

Details of excluded participants are provided in Table 5. The mean age is lower than included population and there were equal number of normal hearing and hearing impaired participants excluded.

Table 5: Data from excluded participants (XP)

Participant	Gender	Age (yrs)	4FA (dB)	Reason for exclusion
XP1	M	73.59	70.00	Spoke English as second language
XP2	F	49.62	13.75	Spoke English as second language
XP3	F	25.37	7.50	Audiologist reported poor reliability
XP4	F	69.80	41.25	Audiologist reported poor reliability
XP5	F	37.86	18.75	Audiologist reported poor reliability
XP6	F	n/a	n/a	Device failure
XP7	F	61.23	50.00	Audiologist reported poor reliability
XP8	M	61.83	n/a	Incomplete test
XP9	F	50.85	n/a	Device failure
XP10	F	21.05	n/a	Device failure
XP11	F	56.38	n/a	Device failure
XP12	M	75.72	n/a	Device failure
Mean age		53.03		
SD age		18.46		

DISCUSSION

This study presents data relating to the accuracy of automated audiometry in a clinical population, using consecutive series recruitment. The present study also examined a range of patient-related factors that may affect accuracy between manual and automated audiometry. The study cohort represents a wide range of type and severity of hearing losses, which has been highlighted as a key limitation of previous studies (Mahomed et al., 2013).

Accuracy

According to the current ISO standard 8253-1:2010, the standard variability for determining a hearing threshold level at frequencies below 4 kHz is 4.9 dB, in a sound-treated environment, without masking and assuming no other uncertainties. To account for uncertainties, the standard acceptable variability in audiometry is an absolute difference of 10 dB, representing the typical ±5 dB test-retest criteria that is practiced widely by audiologists and present in most audiological standards (ASHA, 2004).

Participants in this study had their automated audiometry thresholds tested in a room that was not sound-treated, whilst the manual audiometry testing was performed in a sound-treated environment, potentially introducing further accuracy variation from previous studies (Maclennan-Smith et al., 2013; Storey et al., 2014). The placement of the bone-conductor also differed, with mastoid placement used for the reference test and forehead placement used for the index test and ambient noise levels for the reference test also exceeded the recommended ISO 8253-1:2010 standard. These factors may therefore have contributed additional variation between the reference and index tests.

Therefore, a number of potential influences may have affected variability for this study, with standard test variability, inter-tester differences for manual audiometry, calibration differences and the test environment all likely to influence variation in addition to that caused by automation. Considering the compounded variation of these variables, the variability due to automation appears acceptable, with 86.5% of four-frequency thresholds (500, 1000, 2000 and 4000 Hz) within the accepted ISO absolute variation of 10 dB and 94.8% of participant's thresholds within a further 5 dB of this (i.e. absolute mean threshold difference within 15 dB, real mean difference of ±7.5 dB). This shows that for approximately 95% of participants in this study the additional variation introduced was within an additional 5 dB absolute difference (±2.5 dB relative difference) of ISO standards. It is also comparable to previous

clinical validation studies of the KUDUwave which showed 91% accuracy (Swanepoel et al., 2010) and 92% accuracy (Storey et al., 2014) of obtaining hearing thresholds within a 10 dB absolute difference in sound-treated and 40 dBA multi-talker background noise environments.

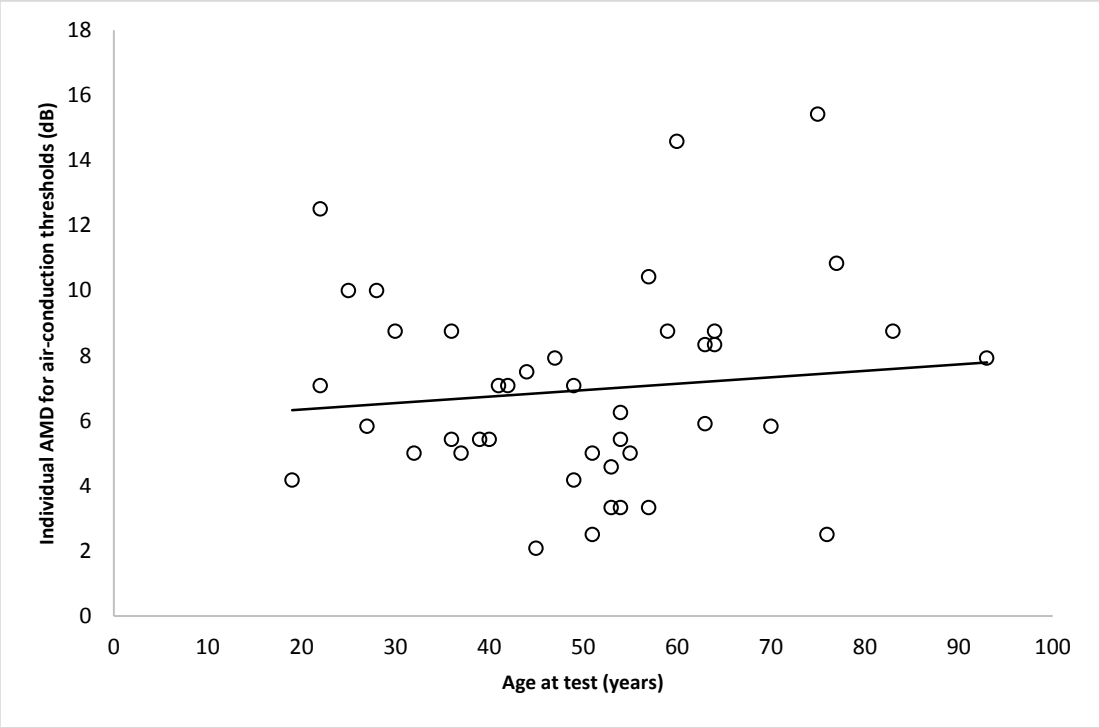
Whilst the additional 5 dB variation above the ISO standard may be considered low once the confounding factors are accounted for, the accuracy of automated audiometry in this study is lower than previous studies which have fallen within the ISO standard threshold variation limits (Eikelboom et al., 2013; Margolis et al., 2011; Storey et al., 2014; Swanepoel et al., 2010). Use of the same KUDUwave audiometer in manual-and automated mode, as in (Swanepoel et al., 2010), eliminates the calibration differences that were introduced in the present study, and could easily account for some of the increased differences in hearing thresholds. The use of sound-treated environments in previous studies (Eikelboom et al., 2013; Margolis et al., 2011; Swanepoel et al., 2010), or artificial background noise environments (Storey et al. 2010) can also help account for the slightly better accuracy estimates in these studies. This study therefore provides clinical validation and accuracy levels for automated audiometry in environments with variable noise levels, typical of an outpatient clinic or health professional's office.

There were a number of outlying hearing thresholds in the study cohort, with six individual threshold differences of 30-35 dB occurring across five different participants, which may have skewed mean differences given the size of our study sample. We have found no clear reason for these outliers but future studies with increased sample sizes may be able to better account for such variation. Whilst previous studies have excluded participants as outliers (Storey et al., 2014), the outliers in the present study were specific to thresholds, not individuals themselves. Inter and intra-tester variability scores were not possible to assess in this study, but previous studies have indicated that the average level of inter-tester variation can be between 2.3–6.0 dB for air-conduction thresholds and 2.9–7.9 dB for bone-conduction thresholds (Margolis et al., 2010). This does not include clinician-specific variation around certain frequencies such as 4000 Hz (Margolis et al., 2013), which may also have increased variation in the results. Data from participants with an outlying threshold have not been removed as it could be argued that their results give a clearer picture of the potential issues facing automated audiometry, namely identifying participants who would benefit from manual audiometry due to patient-related factors that may potentially affect reliability, such

as memory, attention, reaction times and physiological aspects of ageing (e.g. ear canal structure) (Landry, 1999).

To date, the analysis of results comparing automated and manual devices has been predominately descriptive; here we have included analysis of the pair-wise relationship between automated and manual audiometry results in conjunction with an analysis of variance approach. We have also included independent tests of the effect of age, presence of hearing loss on accuracy using a multivariate regression model and used simple linear regression to examine the influence of age on accuracy (Figure 2). The results from these

Figure 2: Scatterplot showing individual absolute mean differences (AMD) for air-conduction between automated and manual audiometry against participant age at testing ($R^2 = 0.11$).



statistical tests are mixed. For the pair-wise tests of automated versus manual audiometry accuracy, the low and mid frequencies for both air and bone conduction showed statistically significant differences with one another. However, at high frequencies for air conduction (4000 and 8000 Hz) and bone conduction (2000 and 4000 Hz) there was no significant statistical differences between the two testing methods. These results from pair-wise tests at high-frequencies were unanticipated. There is a recognized positive skew in manual thresholds referred to by Margolis et al., (2015) as the "Good enough" bias which is believed

to be the result of manual testers not acquiring accurate thresholds below 0 dB because this is deemed to be a sufficient level of hearing. Automated audiometry does not have this basis and may therefore introduce variation. There are also recognised calibration discrepancies at 4000 Hz for bone conduction that may account for additional variation, but in this case, the influence of these differences were not significant on pair-wise analysis (Margolis et al., 2013). For the ANOVA analysis, a significant difference was seen for left ear air conduction thresholds but not right ear air conduction thresholds. These results highlight the random error (as opposed to systematic error) associated with behavioural testing and order effects. We therefore suggest that automated audiometry include a number of pre-assessment trials to familiarise the patient with the automated procedure before the hearing threshold assessment begins.

For air-conduction, 8000 Hz thresholds presented the most variability according to absolute mean difference, although these differences were not statistically significant according to pair-wise tests. It has been established that high-frequency audiometric testing at or above 8000 Hz is more susceptible to variation from differences in the coupling of headphones or earphones and individual physiological differences, with additional variation differences of up to 10 dB (Gössing, 2003). It is therefore possible that the use of insert earphones (KUDUwave) compared to manual audiometry (supra-aural headphones) may have introduced additional variation at higher frequencies.

Age was examined as a potential source of variability in this study. We found statistically significant differences in threshold accuracy between manual and automated audiometry for participants aged ≥55 years using air-conduction audiometry when using a t-test approach. However, this effect was not significant using linear regression or once adjusted for the presence of hearing loss. Therefore, any effect of age on threshold accuracy using automated audiometers (see Figure 2) was either very weak or non-existent (i.e. a Type I error or "false-positive"). However, future studies examining this variable would be beneficial. Despite the variation that may have been introduced by a heterogeneous clinical population, we detected no significant association between the presence of hearing loss and accuracy of automated audiometry.

It could be argued that whilst statistically significant differences have been identified at certain frequencies, the variation in hearing thresholds may be of minimal clinical significance. Whilst the absolute mean variation in thresholds exceeded the ISO standard by 5

dB in this study, this is still the minimal measurable difference in most conventional audiometers and the clinical implications of the difference may be similar to equivalent inter-tester, environmental or calibration differences. Future research should focus on isolating key variables that increase threshold variation as optimal test conditions and patient factors deteriorate and investigate the clinical, rather than statistical, significance of audiometric variation and the effect on audiometry interpretation due to automation.

CONCLUSION

This study has described the clinical validation of automated audiometry in an unselected, clinically heterogeneous population, without the use of a sound-treated environment and with numerous manual audiometry testers, potentially introducing a high degree of inter- and intra-tester variability. Considering this, the difference in hearing thresholds is low, with 86.5% of 4FAs within 10 dB and 94.8% within 15 dB. This study did however reveal that in these least optimal conditions for automated audiometry, the majority of automated hearing thresholds were statistically different to manual thresholds, with the exception of high frequency air-conduction (4000 and 8000 Hz) and bone-conduction (2000 and 4000 Hz) frequencies, and this should be considered when interpreting audiograms produced via this method. However, whilst this variation was statistically significant, future research is needed to ascertain the clinical significance of such variation.

ACKNOWLEDGMENTS

We would like to acknowledge all the participants in this study for the generous use of their time. We would also like to thank all the medical, nursing and support staff from the Department of Otolaryngology, Sir Charles Gairdner Hospital and audiologists Amie Grandidge, Jonathan Constantine, Susan Tegg-Quin, Varsha Matthews and Joanne Sahdeo.

REFERENCES

ASHA. (2004). *Guidelines for the Audiologic Assessment of Children From Birth to 5 Years of Age*. Rockville: American Speech-Language-Hearing Association.

Brennan-Jones, C. G., Taljaard, D.S., Brennan-Jones, S.E.F., Bennett, R.J., Swanepoel, D., Eikelboom, R.E. (2015). Self-reported hearing loss and manual audiometry: a rural versus urban comparison *Australian Journal of Rural Health*, In Press.

Edwards, A. L. (2010). *Evaluation of the Kuduwave 5000 audiometer for compliance with standards for hearing conservation purposes* Johannesburg, South Africa: Council for Scientific and Industrial Research.

Eikelboom, R. H., Swanepoel, D., Motakef, S., & Upson, G. S. (2013). Clinical validation of the AMTAS automated audiometer. *Int J Audiol, 52*(5), 342-349. doi: 10.3109/14992027.2013.769065

Gössing, P., Richter, U. (2003). Characteristic data of the circumaural earphone Sennheiser HDA 200 in the conventional and the extended high frequency range. In U. Richter (Ed.), *Characteristic data of different kinds of earphones used in the extended high frequency range for pure-tone audiometry*. Braunschweig: PTB report PTB-MA-72.

Ho, A. T., Hildreth, A. J., & Lindsey, L. (2009). Computer-assisted audiometry versus manual audiometry. *Otol Neurotol, 30*(7), 876-883. doi: 10.1097/MAO.0b013e3181b120d0

ISO. (1994). Acoustics -- Reference zero for the calibration of audiometric equipment- Part 2: Reference equivalent threshold sound pressure levels for pure tones and insert earphones. ISO 389-2:1994. Geneva, Switazerland: International Organization for Standardization.

ISO. (1998). Acoustics -- Reference zero for the calibration of audiometric equipment -- Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones: ISO 389-1:1998. Geneva, Switzerland: International Organization for Standardization.

ISO. (2010). Acoustics-audiometric Test Methods-Part 1: Basic Pure Tone Air and Bone Conduction Audiometry. ISO 8253-1 2010. Geneva, Switzerland: International Organization for Standardization.

Landry, J. A., Green, W. B. . (1999). Pure-Tone Audiometric Threshold Test-Retest Variability in Young and Elderly Adults *Journal of Speech-Language Pathology and Audiology, 23*(2), 74-80.

Macfarlan, D. (1928). Circuit plans for inexpensive audiometers. *Archives of Otolaryngology, 7*(5), 527-532. doi: 10.1001/archotol.1928.00620010551004

Macfarlan, D. (1939). History of audiometry. *Archives of Otolaryngology, 29*(3), 514-519. doi: 10.1001/archotol.1939.00650050548011

Maclennan-Smith, F., Swanepoel, D., & Hall, J. W., 3rd. (2013). Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. *Int J Audiol, 52*(2), 66-73. doi: 10.3109/14992027.2012.736692

Mahomed, F., Swanepoel, D., Eikelboom, R. H., & Soer, M. (2013). Validity of Automated Threshold Audiometry: A Systematic Review and Meta-Analysis. *Ear and Hearing, Publish Ahead of Print*. doi: 10.1097/AUD.0b013e3182944bdf

Margolis, R. H., Eikelboom, R. H., Johnson, C., Ginter, S. M., Swanepoel, D., & Moore, B. C. (2013). False air-bone gaps at 4 kHz in listeners with normal hearing and sensorineural hearing loss. *Int J Audiol, 52*(8), 526-532. doi: 10.3109/14992027.2013.792437

Margolis, R. H., Frisina, R., & Walton, J. P. (2011). AMTAS®: Automated method for testing auditory sensitivity: II. Air conduction audiograms in children and adults. *International Journal of Audiology, 50*(7), 434-439. doi: doi:10.3109/14992027.2011.553206

Margolis, R. H., Glasberg, B. R., Creeke, S., & Moore, B. C. (2010). AMTAS: automated method for testing auditory sensitivity: validation studies. *Int J Audiol, 49*(3), 185-194. doi: 10.3109/14992020903092608

Margolis, R. H., Wilson, R. H., Popelka, G. R., Eikelboom, R. H., Swanepoel de, W., & Saly, G. L. (2015). Distribution characteristics of normal pure-tone thresholds. *Int J Audiol, 54*(11), 796-805. doi: 10.3109/14992027.2015.1033656

Munro, K. J., & Agnew, N. (1999). A comparison of inter-aural attenuation with the Etymotic ER-3A insert earphone and the Telephonics TDH-39 supra-aural earphone. *Br J Audiol, 33*(4), 259-262.

Rutjes, A. W., Reitsma, J. B., Di Nisio, M., Smidt, N., van Rijn, J. C., & Bossuyt, P. M. (2006). Evidence of bias and variation in diagnostic accuracy studies. *CMAJ, 174*(4), 469-476. doi: 10.1503/cmaj.050090

Slevin, M., Farrington, N., Duffy, G., Daly, L., & Murphy, J. F. A. (2000). Altering the NICU and measuring infants' responses. *Acta Pædiatrica, 89*(5), 577-581. doi: 10.1111/j.1651-2227.2000.tb00342.x

Sparks, D. W. (1972). The feasibility of computerized audiometry. *Journal of Auditory Research, 12*, 62-66.

Storey, K. K., Munoz, K., Nelson, L., Larsen, J., & White, K. (2014). Ambient noise impact on accuracy of automated hearing assessment. *Int J Audiol, 53*(10), 730-736. doi: 10.3109/14992027.2014.920110

Swanepoel, D., & Biagio, L. (2011). Validity of diagnostic computer-based air and forehead bone conduction audiometry. *J Occup Environ Hyg, 8*(4), 210-214. doi: 10.1080/15459624.2011.559417

Swanepoel, D., Mngemane, S., Molemong, S., Mkwanazi, H., & Tutshini, S. (2010). Hearing assessment-reliability, accuracy, and efficiency of automated audiometry. *Telemed J E Health, 16*(5), 557-563. doi: 10.1089/tmj.2009.0143

von Bekesy, G. (1947). A New Audiometer. *Acta Oto-Laryngologica, 35*(5-6), 411-422. doi: doi:10.3109/00016484709123756

Windmill, I. M., & Freeman, B. A. (2013). Demand for Audiology Services: 30-Yr Projections and Impact on Academic Programs. *Journal of the American Academy of Audiology, 24*(5), 407-416.