J Am Acad Audiol 18:78–89 (2007)

Qualind™: A Method for Assessing the Accuracy of Automated Tests

Robert H. Margolis[*†]
George L. Saly[*‡]
Chap Le[§]
Jessica Laurence[*]

Abstract

As audiology strives for cost containment, standardization, accuracy of tests, and accountability, greater use of automated tests is likely. Highly skilled audiologists employ quality control factors that contribute to test accuracy, but they are not formally included in test protocols, resulting in a wide range of accuracy, owing to the various skill and experience levels of clinicians. A method that incorporates validated quality indicators may increase accuracy and enhance access to accurate hearing tests. This report describes a quality assessment method that can be applied to any test that (1) requires behavioral or physiologic responses, (2) is associated with factors that correlate with accuracy, and (3) has an available independent measure of the dimension being assessed, including tests of sensory sensitivity, cognitive function, aptitude, academic achievement, and personality. In this report the method is applied to AMTAS™, an automated method for diagnostic pure-tone audiometry.

Key Words: AMTAS™, audiometry, automated audiometry, cross-validation, Qualind™, quality assessment

Abbreviations: AMTAS™ = Automated Method for Testing Auditory Sensitivity; C_n = set of coefficients for QI_n; K = a constant; Q = psychometric dimension that is tested; Q_i = independent measure of Q; Q_m = measure of Q produced by a test; QA = absolute difference between Q_m and Q_i $|Q_m - Q_i|$; QA_{avg} = average of QA for all S_n; QA_{avg} = predicted value of QA_{avg}; $QA_{i\text{-}avg}$ = average absolute difference between two independent measures of Q; QI_n = set of quality indicators that are used to predict the accuracy of Q_m; Qualind™ = method for predicting the accuracy of an automated test result; S_n = set of n stimuli used to test Q; SD_i = standard deviation associated with $QA_{i\text{-}avg}$; STTR = Small Business Technology Transfer Program; Z_{QA} = Z score for QA_{avg} based on SD_i

Sumario

Conforme la audiología lucha por contener costos, es posible ver cada vez más estandarización, exactitud en las pruebas, rendición responsable de cuentas y use de pruebas automatizadas. Los audiólogos altamente calificados emplean factores de control de calidad que contribuyen en la exactitud de las pruebas, pero éstos no están formalmente incluidos en los protocolos de evaluación, resultando en una amplia gama de exactitudes, relacionada con

*University of Minnesota, Department of Otolaryngology; †Audiology Incorporated, Arden Hills, Minnesota; ‡Arris Systems, Edina, Minnesota; §University of Minnesota, Division of Biostatistics

Robert H. Margolis, University of Minnesota, Department of Otolaryngology, Minneapolis, MN 55455; E-mail: margo001@umn.edu; Phone: 612-626-5794

This work was supported by grant #R42 DC005110 from the National Institute on Deafness and Other Communication Disorders.

Disclosure: AMTAS™ and Qualind™ are the intellectual property of Robert H. Margolis and Audiology Incorporated and may become commercial products.

los diferentes de niveles de habilidad y experiencia de los clínicos. Un método que incorpore indicadores validados de calidad puede incrementar la exactitud y aumentar el acceso a pruebas precisas de audición. Este reporte describe un método de evaluación de calidad, que puede ser aplicado a cualquier prueba que (1) requiera de una respuesta conductual o psicológica, (2) esté asociada con factores que correlacionen con la exactitud, y (3) que posea una medida disponible independiente de la dimensión evaluada, incluyendo pruebas de sensibilidad sensorial, de función cognitiva, de aptitud, de logro académico y de personalidad. En este reporte el método es aplicado a AMTAS™, un método automatizado para audiometría diagnóstica de tonos puros.

Palabras Clave: AMTAS™, audiometría, audiometría automatizada, validación cruzada, Qualind™, evaluación de calidad

Abreviaturas: AMTAS™ = Método automatizado para Evaluar Sensibilidad Auditiva; C_n = set de coeficientes para QI_n; K = una constante; Q = dimensión psicométrica que es evaluada; Q_i = medida independiente de Q; Q_m = medida de Q producida por una prueba; QA = diferencia absoluta entre Q_m y Q_i $|Q_m - Q_i|$; $\mathrm{QA}_{\mathrm{avg}}$ = promedio de QA para todas las S_n; QA_{avg} = valor de predicción de $\mathrm{QA}_{\mathrm{avg}}$; $\mathrm{QA}_{\mathrm{i\text{-}avg}}$ = diferencia promedio absoluta entre dos medidas independientes de Q; QI_n = set de indicadores de calidad que se usa para predecir la exactitud de Q_m; Qualind™ = método para predecir la exactitud de un resultado de una prueba automatizada; S_n = set de n estímulo usados en la prueba Q; SD_i = desviación estándar asociada con $\mathrm{QA}_{\mathrm{i\text{-}avg}}$; STTR = Programa de Transferencia de Tecnología para Pequeños Negocios; Z_{QA} = puntaje Z para QA_{avg} con base en la DS_i

The availability of powerful, low-cost computers and their incorporation into diagnostic test equipment provides an increasing capability to automate diagnostic testing in audiology and other health fields. At the same time, current health-care economics requires greater attention to efficiency and cost containment. In addition to efficiency and lower cost, automation offers other advantages such as standardization, increased accuracy, quality assessment, increased accessibility, and integration into patient databases and electronic medical records.

When hearing tests are conducted by highly skilled audiologists, the clinician employs a variety of quality control factors that contribute to test accuracy. Because these factors are not formally included in hearing test protocols, there is a wide range of quality of results, owing to the various skill and experience levels of clinicians. Automation provides the capability to quantitatively track these factors and formally incorporate them into the test results. Quality assessment of test results is particularly important for automated procedures or other tests in which an expert observer is not present during the test to detect problems that may impact the accuracy of results.

In considering methods for assessing the accuracy of test results, it is important to distinguish between tests that produce binary results (pass-fail, disease–no disease, normal-abnormal) and those that produce continuous measurements, such as pure-tone audiometry, blood pressure, visual acuity, and intelligence tests. The latter are perhaps better referred to as "measurements," "the assignment of numerals to objects or events according to rules" (Stevens, 1951). Measurements, then, can become the basis for *tests*, rules for making decisions based on observations. From this point of view, pure-tone audiometry is a measurement. It becomes a test when rules are employed to make clinical decisions based on the results, such as the categorization of hearing loss severity or type. However, it is common in audiology, medicine, and psychology to refer to measurements that are made for the purpose of making a diagnosis or evaluating a treatment as "tests." Methods for evaluating the accuracy of binary tests are described in the biostatistics literature (Zhou et al, 2002; Pepe, 2003).

This report describes a quality assessment method (Qualind™) that can be applied to any measurement that meets the following requirements: (1) the procedure provides a continuous measurement of the dimension being tested; (2) the measurements result from behavioral or physiologic responses from the individual being tested; (3) quantifiable subject characteristics and behaviors exist that correlate with test accuracy; and (4) an independent measure of the dimension being assessed is available. Tests of this type include (1) tests of sensory sensitivity to physical stimuli including tests of hearing, vision, tactile sensation, and olfaction; (2) tests of cognitive function; (3) aptitude tests; (4) academic achievement tests; and (5) personality tests.

Qualind is based on a class of statistical techniques known as "response surface methodology." These techniques "seek to relate a *response, or output* variable to the levels of a number of *predictors,* or *input* variables, that affect it" (Box and Draper, 1987, p. 1). Although not specifically addressed in the audiology literature, it is widely recognized that certain observable variables are predictive of the accuracy of pure-tone test results. The ubiquitous "reliability" judgment found on audiogram forms assumes that a skilled audiologist can observe certain patient behaviors and characteristics that predict accuracy. If this is the case, then perhaps these variables can be quantitatively tracked and exploited for accuracy assessment, which is amenable to statistical validation. Once derived, the predictive equation can be validated by independent data sets, by the method of *cross-validation* (Schneider and Moore, 2000).

The development of Qualind for assessing the accuracy of an automated method for pure-tone audiometry, based on response surface methodology, is described. The automated audiometric method, AMTAS™, is described below. In Experiment 1, a predictive equation was derived that calculates the estimated average difference between AMTAS and manual thresholds. In Experiments 2 and 3, the accuracy of the predictive equation was cross-validated against two independent data sets.

AMTAS™: AUTOMATED METHOD FOR TESTING AUDITORY SENSITIVITY

AMTAS (U.S. Patent #6,496,585) is an automated method for obtaining a diagnostic pure-tone audiogram, including air- and bone-conduction thresholds with masking in the nontest ear and with quantitative quality indicators. The development and testing of the method was supported by the National Institutes of Health Small Business Technology Transfer (STTR) Program. The results of a comparison between AMTAS and manual testing will be reported separately.

AMTAS is a single interval, Yes-No, psychophysical procedure with feedback. Catch trials are presented randomly throughout the test to allow a quantitative measurement of false-alarm rate. A "quality check" is performed after each threshold determination by presenting a stimulus with a higher intensity than the threshold level. A "No" response to that stimulus is a "Quality Check Fail." Masking is always presented to the nontest ear by a proprietary method that estimates the appropriate masker level from the signal level and interaural attenuation of the transducers. At the conclusion of the test, AMTAS identifies "masking alerts," thresholds for which overmasking or undermasking may have occurred.

The method requires that the air- and bone-conduction transducers are placed at the beginning of the test without the requirement that transducers be moved during the evaluation. The bone vibrator is placed on the forehead. Prototype nonoccluding earphones are used so that bone-conduction thresholds could be determined without contamination by the occlusion effect. The development of a nonoccluding earphone for audiometry is the subject of a related STTR project.

The development of the quality assessment method was based on a comparison of audiograms obtained by AMTAS and by manual testing performed by expert audiologists.

QUALIND™: A METHOD OF ASSESSING THE ACCURACY OF AN AUTOMATED TEST

Qualind (patent pending) is a method for determining the accuracy of a test result from quantifiable factors (subject characteristics and behaviors) that are correlated with test accuracy. The accuracy of the prediction is dependent upon the strength of the association between the factors and the test results. The process of development and validation of Qualind for a specific test is comprised of the following steps.

1. Identify a psychometric quantity, "Q," that can be measured by a properly constructed test.
2. Develop a test of Q that requires behavioral or physiologic responses to a set of stimuli S_n such that the aggregate of the responses to S_n provides a quantitative measure Q_m of Q. S_n could be physical signals, images, or questions which require behavioral or physiologic responses in accordance with well-defined instructions and procedures.
3. Identify n measurable behaviors, QI_n, that may be related to the quality of the subjects' responses to S_n.
4. Identify an independent measure of Q, Q_i, against which Q_m can be compared.
5. Obtain a data set by testing a sample of a defined population providing values of Q_m, Q_i, and QI_n. For each S_n calculate the absolute difference $QA = |Q_i - Q_m|$, which is regarded as a measure of the accuracy of Q_m.
6. Calculate QA_{avg}, the average QA for all of Sn. QA_{avg} represents a global measure of test accuracy for the subject. QA_{avg} could be calculated on subsets of S_n as well to obtain accuracy estimations of various components of a test. In audiometry for example, separate accuracy predictions could be derived for air conduction versus bone conduction, right ear versus left ear, high frequencies versus low frequencies, and so on
7. Derive a predictive equation to estimate QA_{avg} from QI_n. The predictive equation may take the form

$$\mathit{QA}_{avg} = f(QI_n) \qquad (1)$$

The italics indicate a calculated estimate, to distinguish it from QA_{avg}, which is a measured average difference between Q_m and Q_i. One way to obtain $f(QI_n)$ is to perform a multiple regression of QI_n on QA_{avg}. The strength of the regression determines the accuracy of QA_{avg}.
8. If desired, QA_{avg} can be converted to categorical data such as a scale consisting of descriptive terms like "good," "fair," and "poor." These categories can be based on the variance associated with QA_{avg} for a certain population.

In the case of AMTAS, the psychometric quantity (Q) that is measured is hearing sensitivity as expressed by the pure-tone audiogram, including air- and bone-conduction thresholds for standard audiometric frequencies. The independent measure Q_i is the audiogram obtained manually by an expert audiologist. For each audiometric threshold, $QA = |Q_i - Q_m|$ is obtained by determining the absolute difference in auditory thresholds obtained by AMTAS and by the manual method. The average absolute difference for all test stimuli employed in the test is QA_{avg}.

QI candidates can be any quantifiable subject characteristic or behavior that may be related to test accuracy. Potential AMTAS quality indicators are shown in Table 1.

EXPERIMENT 1: DERIVATION AND VALIDATION OF PREDICTIVE EQUATION

Subjects and Methods

Data were collected at three sites chosen to sample a wide range of settings, patient demographics, and hearing loss characteristics. Subjects were recruited from the audiology clinics at each site. A patient was eligible for the study if the clinician judged that the patient was capable of understanding instructions that are typically provided for pure-tone audiometry. No constraints were placed on the degree or type of hearing loss. Immittance testing was not considered. The test sites, subject samples, and hearing loss characteristics are shown in Table 2.

Each site was equipped identically with commercial audiometers (Grason-Stadler GSI-61) and personal computers. Manual testing was performed with TDH-50

Table 1. QI Definitions

AMTAS QUALITY INDICATORS	DEFINITION
Patient Age§	Self-explanatory
Patient Gender§	Self-explanatory
Masker Alert Rate	The number of thresholds for which the masking noise presented to the nontest ear may have been either too low or too high divided by the number of measured thresholds.
Time per Trial	The elapsed time averaged across all observation intervals
Average Number of Trials for Threshold§	The total number of observation intervals divided by the number of measured thresholds
Elapsed Time§	The total elapsed time for the test
False Alarm Rate	The number of false alarms (trials in which the subject reported the presence of a stimulus when no stimulus was presented) divided by the total number of catch trials (trials in which there was no stimulus)
Average Test-Retest Difference	The average difference in threshold measures obtained for stimuli that were tested twice
Quality Check Fail Rate	The total number of occurrences of quality check fails (failure to respond to stimuli presented above threshold) divided by the number of measured thresholds
Air-Bone Gap >50 dB	Number of air-bone gaps (difference between thresholds obtained for air- and bone-conducted stimuli for each frequency/ear combination) that exceed 50 dB
Air Bone Gap <-10 dB	Number of air-bone gaps (difference between thresholds obtained for air- and bone-conducted stimuli for each frequency/ear combination) that are less than -10 dB
Average Air-Bone Gap	The average difference between air-conduction threshold and bone-conduction threshold

§Omitted from final analysis.

Table 2. Experiment 1 Subject Characteristics

Site	n	Gender Male Female	Age (Yrs.) Mean (SD) Range	Pure-Tone Average§ Better Ear (dB) Mean (SD) Range	Pure-Tone Average§ Poorer Ear (dB) Mean (SD) Range	Pure-Tone Average§ Interaural Difference (dB) Mean (SD) Range
University of Minnesota	54	32 22	52 (18) 16 to 87	29 (19) -3 to 70	43 (23) 1 to 104	11 (13) 0 to 69
University of Utah	21	11 10	47 (22) 12 to 76	29 (20) -5 to 55	35 (21) 0 to 64	6 (7) 0 to 33
VA Mountain Home	45	44 1	72 (9) 48 to 93	50 (15) 19 to 94	60 (18) 30 to 104	10 (12) 0 to 48
All	120	87 33	59 (18) 12 to 93	37 (20) -5 to 94	48 (23) 0 to 104	11 (13) 0 to 69

§ 0.5, 1.0, 2.0, and 4.0 kHz

earphones calibrated in accordance with ANSI S3.6-1996 (American National Standards Institute, 1996). For automated testing, a prototype, nonoccluding, circumaural earphone was used. Reference Equivalent Sound Pressure Levels for these earphones were derived by the method described in Annex D Par. D.4 of the standard.

For AMTAS testing, the computer controlled the audiometer, acquired a trial-by-trial history of the entire test, recorded thresholds, tracked the quality indicators, and transmitted the data files to a central server via a secure internet communication protocol. For manual testing, the computer logged the characteristics of each stimulus that was presented, including frequency, intensity, ear, mode (air or bone), and time of presentation, and stored threshold and masking values.

Data were collected by highly experienced audiologists at each test site. Testers were instructed to perform manual audiometry using the clinical methods that they normally use when testing patients. Each tester was validated against another audiologist (RHM) on six subjects at each site. The six subjects were chosen randomly and tested on site by the two testers in immediate succession. Inter-tester reliability for the three sites ranged from 0.95 to 0.97 (Table 3). The inter-tester validation study produced measurements of the differences in thresholds measured by the two audiologists. These data were used to establish accuracy categories for Qualind predictions.

Each subject was tested by AMTAS and by manual audiometry in immediate succession. The order of the tests was randomized. When manual testing was conducted second, the tester was not aware of the AMTAS results.

An equation that calculates the predicted average absolute difference between AMTAS and manual thresholds, QA_{avg}, was derived by performing a multiple regression between $\mathrm{QA}_{\mathrm{avg}}$ and the quality indicators that were selected from Table 1. QI measures that did not contribute to the strength of the regression were discarded. For the remaining factors, the multiple regression returned a set of coefficients (C_n) and an intercept (K). The resulting formula is

$$QA_{avg} = |\mathrm{Q_i} - \mathrm{Q_m}|_{avg} = \mathrm{f(QI_n)} = \Sigma(\mathrm{C{\cdot}QI}) + \mathrm{K} \qquad (2)$$

Note that QA_{avg} relies only on $\mathrm{QI_n}$ and not $\mathrm{Q_i}$ or $\mathrm{Q_m}$. It is an estimate of the accuracy of the test result (relative to results obtained by an expert professional). Its accuracy is determined by the strength of the multiple regression.

Results

A multiple regression of the quality indicators selected from Table 1 and $\mathrm{QA}_{\mathrm{avg}}$ revealed that the following four did not contribute to the strength of the regression–age, gender, average number of trials, and elapsed time. These were discarded and the multiple regression repeated with the remaining eight. The resulting regression coefficient was 0.84, indicating that the eight quality indicators account for 71% of the variance. This result suggests that QA_{avg} is a good predictor of $\mathrm{QA}_{\mathrm{avg}}$. However, the value of QA_{avg} is not a very useful measure for judging the accuracy of an individual audiogram. An additional step is necessary to provide the clinician with a useful quality assessment indicator.

One way to interpret QA_{avg} is to compare it to differences obtained between audiograms obtained by two experienced audiologists on the same patient. That is, two independent measures of $\mathrm{Q_i}$ are obtained. The inter-tester validation study provided a basis for this type of comparison. That study provided a measurement of the average absolute threshold difference for two audiologists:

$$\mathrm{QA_{i\text{-}avg}} = |\mathrm{QA_{i1}} - \mathrm{QA_{i2}}| \qquad (3)$$

The value of $\mathrm{QA_{i\text{-}avg}}$ from that study was 4.62 dB with a standard deviation SDi of 4.13. Using these results, each QA_{avg} was converted to a Z score that is the distance in standard deviation units from the mean average absolute difference for two audiologists. The Z score is calculated by

$$\mathrm{Z_{QA}} = (QA_{avg} - \mathrm{QA_{i\text{-}avg}})/\mathrm{SD_i} \qquad (4)$$

These Z scores were then categorized into three groups, Good, Fair, and Poor, using the rules given in Table 4. This process resulted in the occurrences of Good, Fair, and Poor results in the data set of 120 individuals with hearing loss shown in Table 5.

An examination of the audiograms revealed a high degree of face validity to the categorical data. That is, those in the Good category were generally free of obvious errors that would be evident to skilled audiologists, and those in the Poor category generally were associated with obvious errors such as unlikely audiometric configurations, high false alarm rates either in the aggregate or for specific stimuli, and theoretically impossible results such as air-bone gaps >50 dB or <-10 dB.

Table 3. Inter-tester Correlation Coefficients for Each Test Site

Site	University of Minnesota	University of Utah	VA Mountain Home
Inter-tester Correlation	0.95	0.97	0.97

Table 4. Category Definitions

Category	Z_{QA} Score
Good	Z_{QA} ±1
Fair	$1 < Z_{QA} < 2$
Poor	Z_{QA} ± 2

Table 5. Number of Occurrences of Each $\mathrm{QA}_{\mathrm{avg}}$ Category

Category	Number of Occurrences	%
Good	83	69
Fair	25	21
Poor	12	10

Agreement between categories based on predicted and actual accuracy (i.e., QA_{avg} vs. $\mathrm{QA}_{\mathrm{avg}}$) is shown in Table 6. Overall, for 78% of the cases, the predicted category was identical to the actual category (sum of values in bold divided by total). Sixty-nine percent of the cases were predicted to have "Good" accuracy, and of those, 92% had measured accuracy in the "Good" range. Agreement for the other two categories was not as good. Of 25 cases predicted to be in the "Fair" category, ten (40%) had actual accuracy in the "Fair" range. Of twelve cases predicted to have "Poor" accuracy, seven (58%) had actual accuracy in the "Poor" range.

Perhaps the most serious error is one in which accuracy is predicted to be "Good" and is actually "Poor." Only one error of this type occurred. The reverse case is one in which accuracy is predicted to be "Poor" but is actually "Good." Two errors of this type occurred. These may require unnecessary retesting but are not likely to result in a mischaracterization of the hearing loss if the retesting is done by an experienced audiologist.

QA_{avg} is plotted against $\mathrm{QA}_{\mathrm{av}}$g in Figure 1. The categories are shown by the shaded regions. The correlation between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg} is by definition identical to the multiple regression coefficient of 0.84. The figure illustrates that the seven most inaccurate audiograms (the boxed data points in Figure 1) were correctly categorized in the Poor category. Most of these cases were subjects who did not understand the instructions regarding masking and voted "Yes" when the masking was audible whether or not they heard the tone. Although these are clearly invalid audiograms, they were left in the analysis because it is important to test the power of Qualind to detect such cases.

EXPERIMENT 2: CROSS-VALIDATION AGAINST AN INDEPENDENT DATA SET

The method of cross-validation provides an evaluation of a predictive equation by determining the accuracy of predictions for data sets other than the one from which the predictive equation was derived (Schneider

Table 6. Predicted versus Actual Category Agreement between AMTAS and Manual Audiograms for the Subjects in Experiment 1

		Predicted Category			
		Good n % of column % of total	Fair n % of column % of total	Poor n % of column % of total	TOTAL n %
Actual Category	Good	76 92 63	11 44 9	2 17 2	89 74
	Fair	6 7 5	10 40 8	3 25 3	19 16
	Poor	1 1 1	4 16 3	7 58 6	12 10
	TOTAL n (%)	83 (69)	25 (21)	12 (10)	120

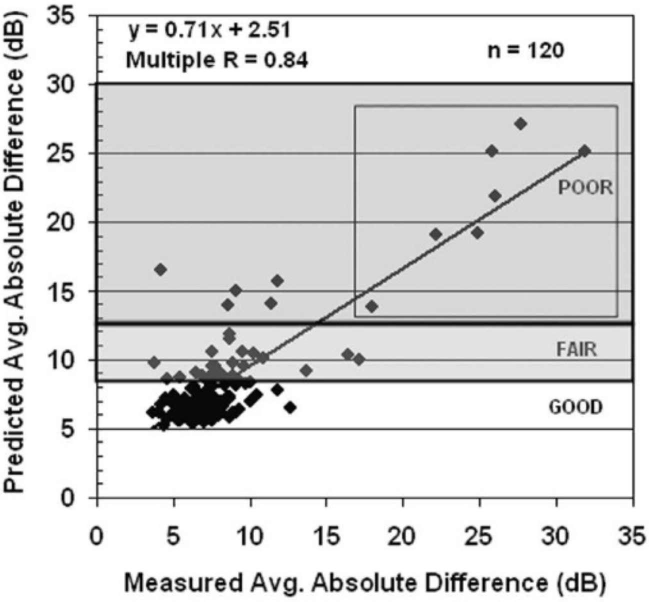

Figure 1. Predicted average absolute differences between AMTAS and manual thresholds plotted against the measured average absolute difference. Average absolute differences are the absolute values of the differences between AMTAS and manual thresholds averaged across all thresholds (air and bone, both ears) for each subject. The predicted average absolute differences are calculated from the regression formula produced by the process described in the text. The Good, Fair, and Poor regions are based on the average differences between manual thresholds obtained by two audiologists for the same subjects. (See text for full explanation.) The seven cases with the poorest accuracy (enclosed in the square) were correctly predicted to have the poorest accuracy.

and Moore, 2000). In Experiment 2 a small group of older subjects was tested by methods identical to those of Experiment 1 for this purpose.

Subjects and Methods

Eight adult subjects (six males, two females) ranging in age from 64 to 85 years were tested at one test site (University of Minnesota). Patient characteristics are summarized in Table 7. Hearing losses varied over a wide range indicated by the pure-tone averages shown in the table. The test procedure was identical to that of Experiment 1. It was not possible to derive a new predictive equation from this data set because of the small number of subjects. Instead, predicted absolute differences were calculated using the regression formula derived from Experiment 1.

Results

A comparison of predicted and actual categories of average absolute differences between AMTAS and manual audiograms for the subjects in Experiment 2 is shown in Table 8. For seven of eight cases (87%), the predicted and actual categories were the same.

A comparison of the predicted and measured average absolute differences between AMTAS and manual audiometry are shown in Figure 2. The figure demonstrates that the one case for which the predicted and measured differences fell into different categories was a near miss (see circled data point in Figure 2).

The correlation coefficient between the measured and predicted average absolute differences was 0.59, indicating that the quality indicators account for 35% of the variance in the differences between QA_{avg} and $\hat{QA}_{avg}$. The lower correlation coefficient (0.59) compared to the multiple regression coefficient obtained in Experiment 1 (0.84) may have resulted from the narrow range of

Table 7. Experiment 2 Subject Characteristics (n = 8) and Results

	Age (years)	Pure-Tone Average (dB) Better Ear (0.5, 1, 2, 4 kHz)	Pure-Tone Average (dB) Worse Ear (0.5, 1, 2, 4 kHz)	Interaural Difference in Pure-Tone Average (dB)	Difference between Predicted and Measured Absolute Difference (dB)
Mean	70.2	25	35	10	1.2
SD	7.2	12	21	11	0.9
Range	64–85	9–44	13–76	1–35	0.9–3.0

Table 8. Predicted versus Actual Category Agreement between AMTAS and Manual Audiograms for the Subjects in Experiment 2

		Predicted Category		
		Good n % of column % of total	Fair n % of column % of total	TOTAL n %
Actual Category	Good	5 100 63	1 33 13	6 75
	Fair	0 0 0	2 67 25	2 25
	TOTAL n (%)	5 (63)	3 (37)	8

the differences in Experiment 2. Unlike the heterogeneous distribution of accuracy in Experiment 1, these subjects were quite homogeneous, indicated by the substantial difference in the standard deviations for average absolute difference (4.7 in Experiment 1 vs. 2.4 in Experiment 2). All of the results fell into the "Good" and "Fair" categories with none in the "Poor" category. Because the correlation statistic is highly dependent on the range of the data (Games and Klare, 1967, p. 369), it is expected that the correlation in Experiment 2 would be lower than the multiple regression in Experiment 1.

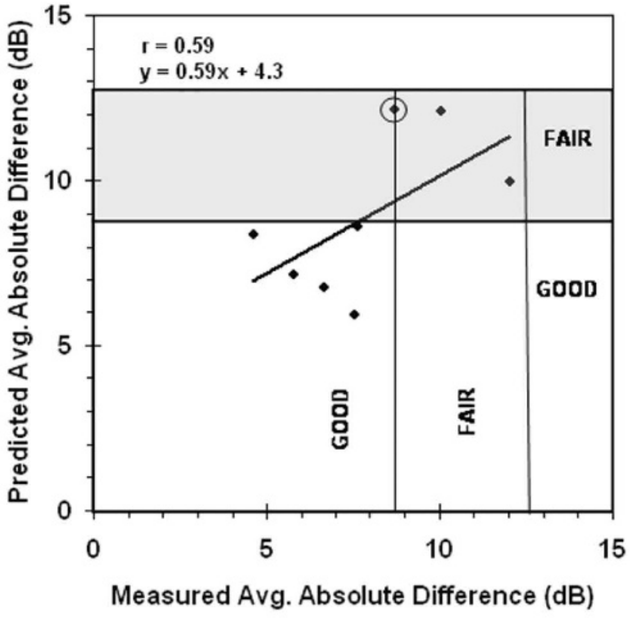

Figure 2. Predicted and measured average absolute differences between AMTAS and manual thresholds for the eight listeners tested in Experiment 2. Predictions were based on the regression equation derived in Experiment 1. The "GOOD" and "FAIR" labels indicate the ranges determined from average differences between manual testing performed by two audiologists for the same subjects. The horizontally aligned "GOOD" and "FAIR" labels indicate the predicted categories. The vertically aligned labels indicate the categories based on measured differences.

To examine the possibility that the lower correlation in Experiment 2 may have been influenced by the narrower range of results relative to Experiment 1, a subset of subjects from Experiment 1 was selected such that the range of values of QA_{avg} was identical to the range in Experiment 2. The correlation coefficient for this data set is 0.45, lower than the 0.59 obtained in Experiment 2. This result suggests that over that range, the relationship between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg} is stronger in Experiment 2 than in Experiment 1.

The average absolute difference between the measured and predicted average absolute differences was 1.9 dB (range = 0.4–2.4 dB), suggesting a high degree of predictability of QA from the quality indicators.

The categorical agreement (87%), the correlation between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg}, and the small mean difference between predicted and measured accuracy indicate that the regression equation produced in Experiment 1 provides a high degree of predictability for the results of Experiment 2. In addition, the slopes and intercepts of the best fit regression lines (see equations in Figures 1 and 2) are similar, suggesting very similar relationships between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg} in the two subject groups. The results of this cross-validation analysis suggest that a regression equation obtained from one data set obtained by AMTAS and manual testing are useful for predicting the accuracy of audiograms in another data set, provided the subjects are reasonably similar.

EXPERIMENT 3: CROSS-VALIDATION FOR INSERT EARPHONES

Subjects and Methods

To determine the extent to which the predictive equation derived in Experiment 1 might extend to variations in methodology, a cross-validation study was conducted at the Minnesota site for a group of subjects tested with a different earphone. In this study, AMTAS thresholds were measured with insert earphones (Etymotic Research ER3A), and manual thresholds were measured with supra-aural earphones (Telephonics TDH-50). Each was calibrated by standard calibration procedures (American National Standards Institute, 1996). The ER3A earphone was coupled to the ear with either the small or adult standard foam tips. The appropriate size was selected for each subject, the foam tip was compressed, and it was inserted into the ear canal such that the lateral surface of the tip was at the level of the entrance to the ear canal. With this insertion depth, an occlusion effect is expected that affects low-frequency bone-conduction thresholds (Dean and Martin, 2000). During AMTAS testing, both ears were occluded. During manual audiometry, only the nontest ear was occluded during bone-conduction testing. Therefore, a difference in low-frequency bone-conduction thresholds is expected that will affect the overall agreement between AMTAS and manual thresholds. Subject characteristics are summarized in Table 9.

Two predictions were derived for each subject. QA_{avg} was calculated using the coefficients produced by the multiple regression performed in Experiment 1. This is the cross-validation approach. In addition, a set of coefficients were obtained from a new multiple regression on the data from Experiment 3.

Results

The differences between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg} for the two predictive equations are shown in the last two columns of Table 9. In both cases the average differences were small with the new regression equation providing smaller differences and smaller ranges.

The performance of the two equations for assigning audiograms to categories is shown in Table 10. Again, the new predictive equation performed slightly better with 33 cases (92%) correctly classified as “Good” compared to 31 (86%) for the equation from Experiment 1. It is not surprising that the regression based on the data from Experiment 3 would be more predictive than a predictive equation derived from a different data set.

The superior performance of the new predictive equation is clearer in Figures 3 and 4, which show the relationship between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg} for each predictive equation. The new multiple regression is 0.68 (which is equivalent to the correlation between $\mathrm{QA}_{\mathrm{avg}}$ and QA_{avg}). The correlation based on the Experiment 1 predictions is 0.32, and the plot in Figure 4 shows substantially more

Table 9. Experiment 3 Subject Characteristics (n = 36) and Results

	Age (years)	Pure-Tone Average (dB) Better Ear (0.5, 1, 2, 4 kHz)	Pure-Tone Average (dB) Worse Ear (0.5, 1, 2, 4 kHz)	Absolute Interaural Difference in Pure-Tone Average (dB)	Cross-Validation: Absolute Difference between Predicted and Measured Average Absolute Difference (dB)[a]	New Regression: Absolute Difference between Predicted and Measured Average Absolute Difference (dB)[b]
Mean	61.3	39	49	10	1.5	1.0
SD	18.2	22	22	14	1.2	0.8
Range	13 to 86	-4 to 75	3 to 89	0 to 62	0.1 to 4.0	0 to 2.8

[a] From the coefficients derived in Experiment 1.

[b] Predictions from new multiple regression.

Table 10. Predicted versus Actual Category Agreement between AMTAS and Manual Audiograms for the Subjects in Experiment 3

Predicted Category							
		Cross-Validation			New Predictive Equation		
		Good n % of column % of total	Fair n % of column % of total	TOTAL n %	Good n % of column % of total	Fair n % of column % of total	TOTAL n %
Actual Category	Good	31 94 86	3 100 8	34 94	33 94 92	1 100 3	34 94
	Fair	2 6 6	0 0 0	2 6	2 6 6	0 0 0	2 6
	TOTAL (%)	5 (63)	3 (37)	36 (100)	35 (97)	1 (3)	36 (100)

scatter for the cross-validation than for the new predictive equation (Figure 3). In spite of the larger sample size, the cross-validation analysis indicated a weaker predictive relationship compared to the results of Experiment 2.

Contributing to the poorer performance of the cross-validation predictions may be the occlusion effect produced by the insert earphones. When subjects with conductive hearing losses were excluded, the bone conduction thresholds with insert earphones

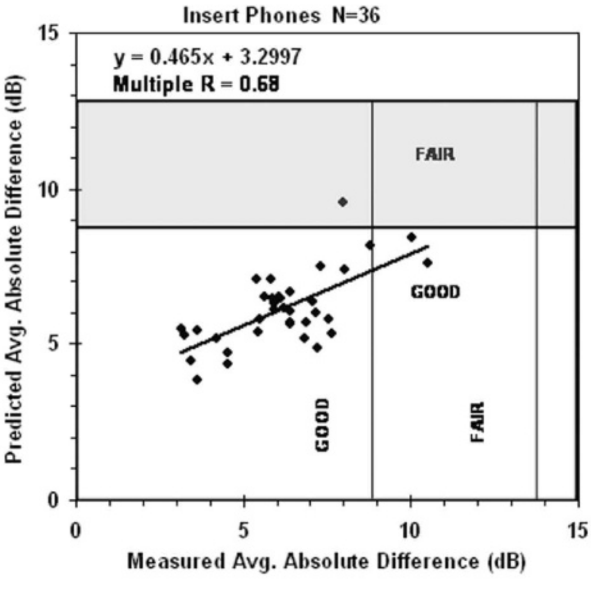

Figure 3. Predicted and measured average absolute differences between AMTAS and manual thresholds for the 36 listeners tested in Experiment 3. Predicted average absolute differences (QA_{avg}) were calculated from a multiple regression formula based on this data set. The "GOOD" and "FAIR" labels indicate the ranges determined from average differences between manual testing performed by two audiologists for the same subjects. The horizontally aligned "GOOD" and "FAIR" labels indicate the predicted categories. The vertically aligned labels indicate the categories based on measured differences.

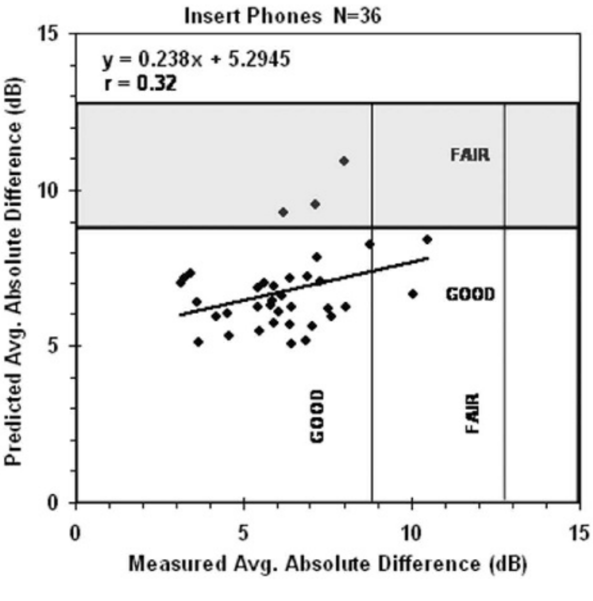

Figure 4. Predicted and measured average absolute differences between AMTAS and manual thresholds for the 36 listeners tested in Experiment 3. Predicted average absolute differences (QA_{avg}) were calculated from a multiple regression derived in Experiment 1. The "GOOD" and "FAIR" labels indicate the ranges determined from average differences between manual testing performed by two audiologists for the same subjects. The horizontally aligned "GOOD" and "FAIR" labels indicate the predicted categories. The vertically aligned labels indicate the categories based on measured differences.

averaged 9 dB and 6 dB lower (better) at 250 and 500 Hz, respectively, compared to the supra-aural earphones. These differences did not occur in Experiments 1 and 2, in which a nonoccluding earphone was used for AMTAS testing.

The results of this analysis suggest that the strength of the predictive equation derived in Experiment 1 is compromised when an earphone with different characteristics is employed. However, the regression derived from the Experiment 3 data set appeared to have similar predictive power. Thus, it may be necessary to derive predictive equations that are specific to differences in instrumentation that may differentially affect results, such as predicting unoccluded bone-conduction results from measurements obtained with an occluding earphone.

SUMMARY AND CONCLUSION

A method was developed based on response surface methodology for assessing the accuracy of a diagnostic audiogram obtained by a computer-controlled, automated procedure. 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.

Acknowledgments. Lisa Hunter and Richard Wilson have been invaluable colleagues in this project. M. Victor Berrett, Allison Kohtz, and Deborah Weakley collected the data, and their expertise and professionalism were critical to the study. We are grateful to Aaron R. Thornton and two anonymous reviewers for their critical reading of the manuscript.

REFERENCES

American National Standards Institute. (1996) *American National Standard Specification for Audiometers (ANSI3.6-1996).* New York: Acoustical Society of America.

Box GEP, Draper NR. (1987) *Empirical Model-Building and Response Surfaces.* New York: John Wiley.

Dean MS, Martin FN. (2000) Insert earphone depth and the occlusion effect. *J Am Acad Audiol* 9:131–134.

Games PA, Klare GR. (1967) *Elementary Statistics: Data Analysis for the Behavioral Sciences.* New York: McGraw-Hill.

Pepe MS. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

Schneider J, Moore A. (2000) *A Locally Weighted Learning Tutorial Using Vizier 1.0.* Tech. Report CMU-RI-TR-00-18, Robotics Institute, Carnegie Mellon University. http://www.cs.cmu.edu/~schneide/tut5/node42.html.

Stevens SS. (1951) Mathematics, measurements, and psychophysics. In: *Handbook of Experimental Psychology.* New York: John Wiley.

Zhou XH, Obuchowski NA, McClish DK. (2002) *Statistical Methods in Diagnostic Medicine.* New York: John Wiley.