AJA

Research Article

Reliability of Broadband Middle-Ear Power Reflectance in Younger and Older Adults: Application of Generalizability Theory

Marty J. Mahoney,[a] Dennis J. McFarland,[b] MiChelle S. Carpenter,[a] Sabahet Rizvi,[a] and Anthony T. Cacace[a]

Purpose: To assess the reliability of broadband middle-ear power reflectance (BMEPR) and transmittance profiles for chirp and tonal stimuli using generalizability theory (GT).
Method: In adults without a history of middle-ear disease, the authors assessed the reliability of BMEPR to chirp and tonal stimuli using a multivariate approach based on an analysis of variance model (GT). For comparisons with other published studies, Pearson's product–moment correlation coefficients (Pearson's *r*) also were used.
Results: Based on GT with chirp stimuli, overall BMEPR measures had good reliability; however, the reliability of individual profiles across frequencies and ears was less than optimal. Lower generalizability coefficients were found when transmittance was evaluated. Test–retest reliability using Pearson's *r* was better for right versus left ears, and mid-frequencies were generally more reliable than those at either extreme of the frequency range. In contrast, tonal stimuli had higher generalizability coefficients and Pearson's *r* values than chirps for all frequencies tested; Pearson's *r* values were also higher for right versus left ears.
Conclusion: The authors extended the use of GT as a preferred way to evaluate reliability of BMEPR and transmittance profiles for chirps and tones because it allows for a more comprehensive evaluation compared with unidimensional pairwise correlations.

Key Words: adults, audiology, hearing, middle ear, power reflectance, generalizability theory

Measurement of broadband middle-ear power reflectance (BMEPR) represents an emerging technology for evaluating electroacoustic characteristics of human middle-ear function in vivo (Allen, Jeng, & Levitt, 2005; Jeng, Allen, Lapsey-Miller, & Levitt, 2008).[1] With this method, high-resolution frequency reflectance, absorbance, and/or transmittance profiles offer bio-inspired assessment opportunities for evaluating the middle ear under normal and pathological conditions (see, e.g., Feeney, Grant, & Marryott, 2003; Feeney, Grant, & Mills, 2009; Hunter, Tubaugh, Jackson, & Propes, 2008; Keefe & Simmons, 2003; Shahnaz, Bork, et al., 2009; Shahnaz, Longridge, & Bell, 2009). Nevertheless, as BMEPR measures transition from the laboratory to the clinic, the need for establishing the reliability of these measures is an important factor for test evaluation and clinical decision making.

Given the broadband characteristics of this metric, methodological and design considerations should take into account whether to base reliability on individual data points (frequencies; Hunter et al., 2008), select bands of frequencies (see, e.g., Beers, Shahnaz, Westerberg, & Kozak, 2010;

[a]Wayne State University, Detroit, MI
[b]The Wadsworth Center, New York State Department of Health, Albany
Correspondence to Anthony T. Cacace: cacacea@wayne.edu

Editor: Larry Humes
Received November 6, 2012
Revision received March 18, 2013
Accepted March 21, 2013
DOI: 10.1044/1059-0889(2013/12-0063)

[1]Broadband electroacoustic measures of middle-ear function can be represented in a number of different formats: power reflectance, absorbance, transmittance, and so on. *Middle-ear power reflectance* is defined as the ratio of reflected power to the incident power, which, when normalized, can range from 0 to 1 (where 0 = no reflectance and 1 = maximum reflectance), or it can be expressed as a percentage, from 0% to 100%. Absorbance is just a linear transformation of reflectance (1 – reflectance) representing the amount of energy that is absorbed versus reflected from the tympanic membrane/middle-ear system. Transmittance transforms the absorbance metric into a decibel scale (see Equation 1). Furthermore, there is no standard at present for representing how these measures should be expressed. To be clear, expressing these data as either a power reflectance or absorbance metric is a preference and not a requirement; it will not fundamentally change the result. It has been suggested that, by converting absorbance to transmittance, this metric would be less variable and may be more amenable to comparisons with hearing loss, because hearing loss is also expressed on a logarithmic (dB) scale (e.g., Allen et al., 2005; Jeng et al., 2008; Keefe & Simmons, 2003).

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

Shahanaz, Bork, et al., 2009; Vander Werff, Prieve, & Georgantas, 2007), or the shape of the entire frequency reflectance profile (the present study). Moreover, special consideration should be given as to whether to compute reliability metrics based on absolute difference measures of central tendency and dispersion (*M*s and *SD*s), Pearson product–moment correlation coefficients (Pearson's *r*), or generalizability theory (GT).

Although researchers have used different approaches to assess test–retest reliability, some methods are on more secure scientific ground than others. For example, absolute differences of BMEPR have been used as indices of test–retest reliability when values are measured over two or more points in time (see, e.g., Beers et al., 2010; Rosowski et al., 2012; Shahnaz, Bork, et al., 2009; Vander Werff et al., 2007; Werner, Levi, & Keefe, 2010). On the surface, this approach makes intuitive sense because a reliable test is one that produces similar results on two different occasions. However, with the absolute-difference method, the interpretation of good or poor reliability is unduly subjective. This is due to the fact that this method does not take variability into account, and its meaning depends on the scale of measurement applied. In contrast, the more conventional Pearson's *r* provides a standardized metric for reliability calculations because it is based on the proportion of variance that is repeatable. Because Pearson's *r* represents a normalized difference metric that is signed, it allows for direct comparisons to be made with other studies because it is independent of the unit of measurement (see, e.g., Hunter et al., 2008; Werner et al., 2010; the present investigation). Although the advantages of using Pearson's *r* over the absolute-difference method are apparent, this approach is limited to pairwise comparisons; therefore, the univariate nature of this metric is not well suited for complex data sets. In comparison to these latter two measures, GT is a multivariate approach to reliability based on an analysis of variance (ANOVA) model, in which more than two points in time and multiple independent variables can be jointly considered in the computations. Furthermore, GT is unique when compared with the typical ANOVA model. Whereas the typical ANOVA model considers subjects as the source of error and considers trends over time as the effect of interest, in GT, the variance associated with subjects is the effect of interest, and the variance over time is the source of error. With this strategy, the resultant generalizability coefficient becomes a measure of the effect size (i.e., the size of the main effect for subjects) that represents the proportion of variance that is due to consistent individual differences. Thus, GT provides a framework for assessing multiple time points, including main effects and interactions between multiple independent variables. Of particular relevance to the current area of interest is the interaction between subjects and stimulus frequency because this relationship allows for the reliability of individual frequency reflectance profiles to be assessed.

It is our contention that the choice and rationale of whether to base reliability estimates on individual data points, on select bands of frequencies, or on profiles should depend on how clinicians actually use these measurements to diagnose middle-ear disorders. For example, if a diagnosis is based on a single point or on a single band, independent of the overall shape of the profile, then the reliability of individual data points or bands would be the appropriate index. However, as Keefe and Simmons (2003) noted,

> There is no evidence to suggest that the use of a single frequency, as in clinical tympanometry, is optimal for assessing middle-ear function at all frequencies important in auditory communication systems, no more than would a single frequency suffice for assessing cochlear, behavioral, or neural function. Wideband measurements of middle-ear functioning appear to have promise as a clinical diagnostic test. (p. 3217)

In this article, we extend the logic of Keefe and Simmons to include the fact that if clinicians base their diagnosis on the shape of the entire profile, then the reliability of the profile would be the most appropriate feature to evaluate. We focus herein on the use of GT for establishing test–retest reliability of BMEPR data, whereby the effects of multiple variables are to be considered (see, e.g., Crocker & Algina, 1986; Cronbach, Nageswari, & Gleser, 1963; Laenen, Vangeneugden, Geys, & Molenberghs, 2006). Last, because GT has not been used extensively in the audiological/hearing science literature, we provide a concise overview to acquaint readers with this topic (see the Appendix).

Method and Materials

Fifty-six adults, categorized into two age groups (Group 1: 18–25 years, $n = 28$; Group 2: ≥ 50 and ≤ 66 years, $n = 28$), were studied. Each age group was stratified by gender (14 men, 14 women) and ear (56 left, 56 right) and, therefore, provided a balanced design among age group, gender, ear, and frequency. Because subjects were recruited by word of mouth from friends, relatives, and students, the data obtained were considered a convenience sample. Inclusion criteria were a negative history of middle-ear disease; no air–bone gaps exceeding 10 dB for any frequency; and ear canals free of obstruction or debris, based on a screening otoscopic exam. The Human Investigation Committee at Wayne State University approved this study, and we obtained signed informed consent from each individual prior to data collection.

Audiometric testing was conducted in a commercial sound booth (Acoustic Systems, Model RE-144) through use of a clinical audiometer (Grason-Stradler, Model 61) with standard earphones (Telephonics, Model TDH-50P) enclosed in supra-aural ear cushions (MX-41/AR). Pure-tone air-conduction audiometry was performed at octave frequencies from 250 Hz through 8000 Hz and at one-half–octave frequency (3000 Hz) bilaterally. Bone-conduction testing used a standard oscillator (Radioear B-71) and a standard headband. Bone-conduction thresholds were assessed at octave frequencies ranging from 250 Hz through 4000 Hz.

BMEPR was measured using commercially available hardware and software (Mimosa Acoustics, MEPA3 Clinical

Reflectance System) and a high-quality probe assembly (Etymotics, Model ER10C) to transduce acoustic stimuli and record acoustic responses from the ear canal. Before each recording session, the MEPA3 system was successfully calibrated in a four-chamber coupler (Model CC4-V) in accordance with guidelines provided by the manufacturer. Particular care was taken to ensure that the foam ear tip of the probe was properly seated and stable in the ear canal. There were no crimps in the foam, and this coupling device was fully expanded in the ear canal before testing was initiated. After measurements from chirp and tonal stimuli were obtained from each ear, the probe was removed and reinserted into the same ear canal, and a second set of chirp and tonal measures was acquired. Then, the second ear was tested using the same approach. For the present investigation, both sets of within-session data were used in this analysis. The SPL of the chirp stimulus was set to 60 dB (re: 20 uPa), and data were collected over a 1-s time epoch at ambient ear canal air pressure. This allowed for 24 individual chirps (~5 ms in total duration) to be collected and averaged. The SPL of the tonal stimuli was also set to 60 dB (re: 20 uPa), and nine individual pure tones (ranging from 257 Hz to 6000 Hz) were analyzed. Individual tonal stimuli were 300 ms in total duration, presented sequentially from low to high frequency and separated by a 150-ms silent interstimulus interval. Selection of the initial ear of measurement was randomized by a physical coin toss (heads = left ear, tails = right ear) to avoid potential order effects that might confound data interpretation (see, e.g., Thornton, Marotta, & Kennedy, 2003). The same medium-sized foam ear tips (14A) were used during instrument calibration and data collection. With respect to chirp stimuli, out of a possible 248 frequencies measured, we selectively sampled a subset of 16 frequencies (258, 307, 398, 492, 633, 750, 796, 1008, 1270, 1500, 1590, 1992, 2530, 3000, 4060, and 5040 Hz) for this investigation that clearly outlined the frequency reflectance profile. With respect to tonal stimuli, we used default values and sampled nine separate frequencies (258, 492, 750, 1007, 1500, 1992, 3000, 4007, and 6000 Hz). Power-reflectance values associated with both chirp and tonal stimuli were extracted from separate stored output files that were available from each subject.

To allow for reliability to be evaluated from a multivariate perspective, we conducted an ANOVA to compute the generalizability coefficients. To allow for comparisons with other published studies in the literature, we also used Pearson's *r* to evaluate test–retest reliability for individual frequencies.

Results

Figure 1 shows grand averaged frequency reflectance profiles separately for chirp and tonal stimuli collapsed across all variables and for all combinations of age, gender, and ear variables. Except for the highest frequency studied, average power-reflectance values corresponding to each stimulus type (chirp and tone) were very similar. In Figure 2, individual scatter plots are shown for 16 frequencies and all test–retest conditions for chirp stimuli. The general trends observed in these plots show that within-session variability increased from low to high frequencies. Figure 3 shows individual scatter plots for nine separate tonal frequencies and for all test–retest conditions. Although a similar trend for increased within-session variability with increases of stimulus frequency was also observed, tonal stimuli showed less within-session variability than did chirps.

For comparison with previous studies, in Table 1 (left side) we provide Pearson's *r* values for the test–retest reliability of 16 individual frequencies for chirp stimuli, separately for each ear. Trends in these data reveal higher reliabilities for right versus left ears, with mid-frequency reliabilities generally higher than those at either extreme. Test–retest reliabilities of Pearson's *r* values for the nine individual frequencies are presented in the right side of Table 1, separately for each ear obtained for tonal stimuli. These data show higher test–retest correlation values in comparison to chirps at corresponding frequencies; right ears also showed higher test–retest correlations than left ears.

Next, we analyzed both chirp and tonal data sets with an ANOVA, in two ways. First, the effect of subjects was used in the error terms to evaluate the consistency of age, gender, ear, and test effects across subjects (i.e., traditional null hypothesis significance testing). The second set of analyses used tests in the error terms to evaluate the proportion of variance due to subject effects that was consistent across tests (i.e., generalizability or test–retest reliability). Results were calculated for both power reflectance and transmittance. We analyzed the transmittance metric because we thought that this transformation might reduce variability and thus potentially improve the generalizability coefficients (Allen et al., 2005), keeping in mind that although this assertion was suggested by Allen and colleagues (2005), it has never been proven empirically.

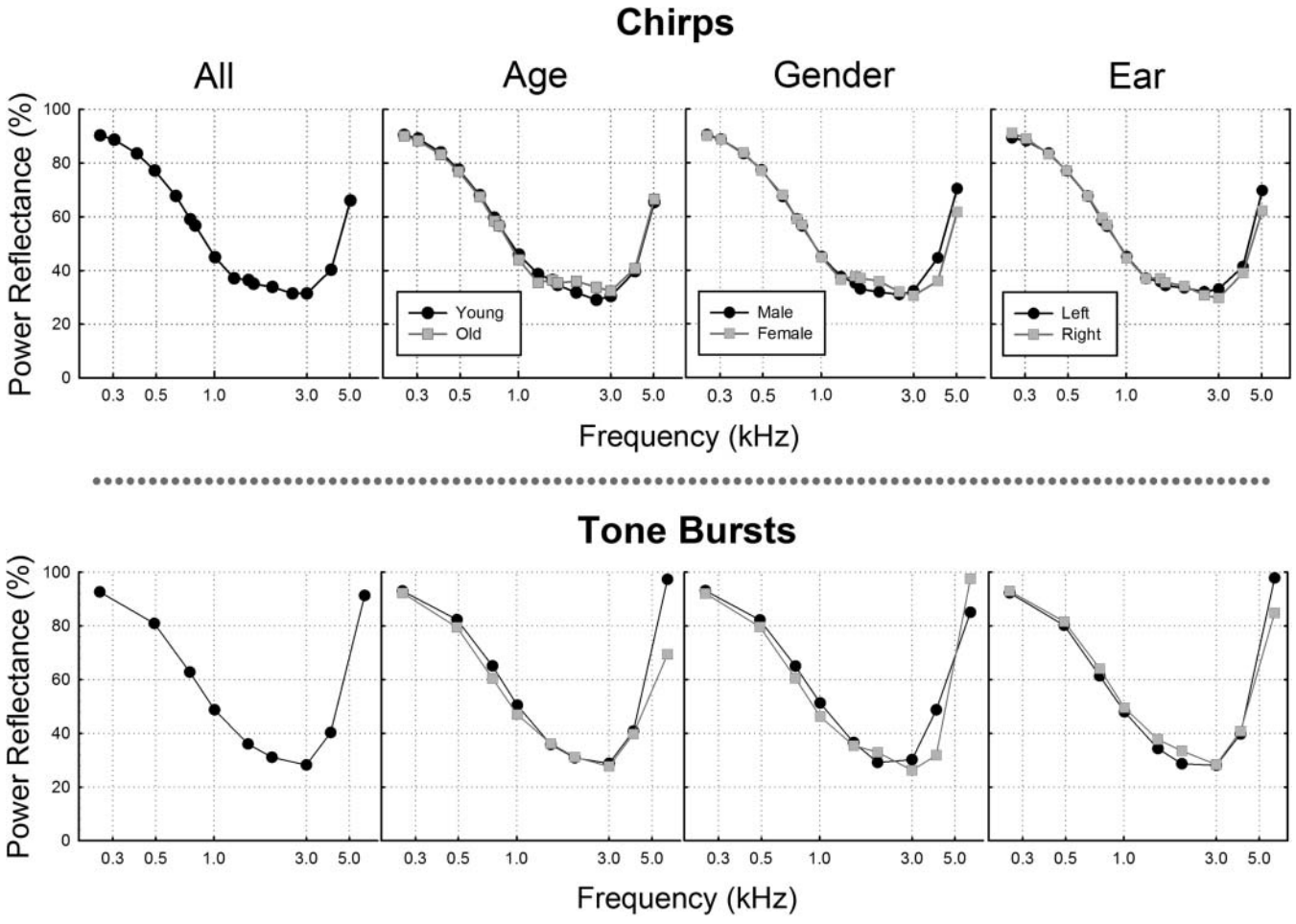
Chirps

We conducted a six-way ANOVA in which the effects of age and gender were used as between-subjects variables and ear, frequency, and time were used as within-subject variables. With subject effects used as the error term, the ANOVA showed a significant main effect of frequency ($F = 350.31$, $p < .0001$), resulting from the lower power reflectance in mid-frequencies, as seen in all plots of Figure 1. There were also Gender × Ear ($F = 4.22$, $p < .045$), Gender × Frequency ($F = 2.38$, $p < .002$), and Age × Gender × Ear ($F = 4.60$, $p < .037$) interactions. The Gender × Frequency interaction was due in part to greater reflectance in men at the highest frequencies. The three-way interaction was associated with greater reflectance in the right ear of older women and the left ear of older men, with less difference in younger women and men.

The results of the ANOVA in which test effects were used as the error term resulted in the generalizability coefficients shown in Table 2 (left side). Also shown in Table 2 (right side) are generalizability coefficients for the transmittance values, computed as follows:

$$T = 10 \times \log_{10}[1 - (|R|^2)](\mathrm{dB}), \quad (1)$$

Figure 1. Grand averaged frequency–reflectance profiles for chirps and tonal stimuli. The demarcations noted at the top of each figure ("All," "Age," "Gender," "Ear") represent individual categories of data. "All" indicates that data were collapsed across age, gender, and ear; "Age" (young and old) indicates that data were collapsed across gender and ear; "Gender" (male and female) indicates that data were collapsed across age and ear; and "Ear" (left and right) indicates that data were collapsed across age and gender.



where T is the transmittance and $|R|^2$ is the power reflectance expressed as a proportion in decibels.

These data show that the generalizability coefficient associated with the main effect of subjects was 0.82. This score represents the average for each subject collapsed over frequency and ear. The reliability associated with the Subjects × Frequency interaction was 0.56. This effect corresponds to the profile for individual subjects across frequencies averaged across both ears. The reliability of the Subjects × Ear × Frequency interaction was 0.35. This corresponds to the shape of the profile for individual subjects across frequencies for individual ears and probably represents the feature of greatest interest to the clinician. Generalizability coefficients for transmittance were somewhat lower, particularly for the three-way interaction.

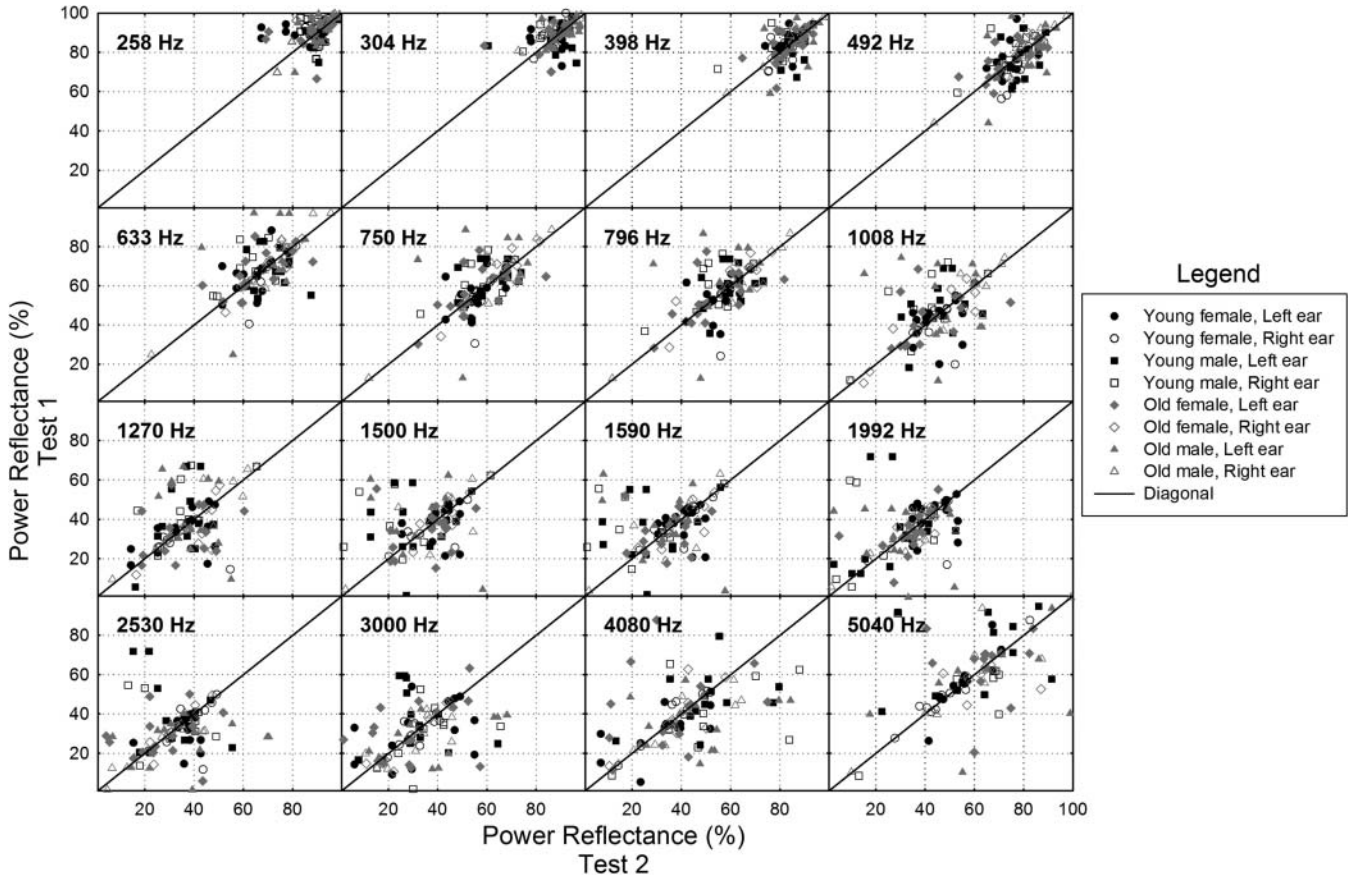
Tonal Stimuli

Generalizability coefficients for middle-ear power reflectance using tonal stimuli are shown in Table 3 (left side). The generalizability coefficient associated with the main effects of subjects was 0.86. The reliability associated with the Subjects × Frequency interaction was 0.72. The reliability associated with the Subjects × Ear interaction was 0.75, and that for the Subjects × Ear × Frequency interaction was 0.63. Because reliability coefficients are correlations, differences can be evaluated with Fisher's z transformation, and significance levels depend on the number of cases studied. In the present comparison (0.63 vs. 0.75), in which there were 56 subjects, the difference was not significant (see Ramseyer, 1979). Nevertheless, these values were considerably higher than those reported for chirps. Values for the transmittance data for tonal stimuli (see Table 3, right side) were similar to the power-reflectance data.

Discussion

Cronbach and colleagues (1963) and Cronbach, Gleser, Nanda, and Rajaratnam (1972) initially introduced GT

Figure 2. Composite scatter plots for 16 frequencies for different variables studied. Data collected from Test 1 are plotted on the *y*-axis, and data collected from Test 2 are plotted on the *x*-axis. If the within-session data from Test 1 and Test 2 were identical for each of the different frequencies, then data points would fall directly on the solid diagonal line in each of the plots. On the basis of the scatter of data points observed, the degree of within-session variability appears rank ordered from low, to middle, to higher frequencies.



into the educational psychology literature to evaluate the reliability of profiles of standardized test scores in school-age children. Because of its unique capabilities of assessing reliability from a multivariate perspective, GT has garnered increased interest in other fields of inquiry, including speech, hearing, vestibular, and physical sciences, where a wide range of topics—physiology, electroacoustics, perception, responses to questionnaires, and so forth—have already been investigated. Relevant examples include the reliability of distortion product otoacoustic emissions over a 24-hr time period (Cacace, McClelland, Weiner, & McFarland, 1996), postural control in the evaluation of concussion (Broglio, Zhu, Sopiarz, & Oark, 2009), speechreading abilities (Demorest & Bernstein, 1992), perceptual scaling (S. O'Brian, Packman, Onslow, & O'Brian, 2003), analysis of observational data (N. O'Brian, O'Brian, Packman, & Onslow, 2003; Scarsellone, 1998), videographic representation of tooth and lip position in smiling and speech following orthodontic and dentofacial surgery (van de Geld, Oosterveld, van Waas, & Kuijpers-Jaqtman, 2007), and in force measurements used in physical therapy and rehabilitation (Roebroeck, Hariaar, & Lankhors, 1993). Use of GT in assessing the reliability of BMEPR profiles expands this list of testing domains to include another form of auditory-based electroacoustic analysis, in which profiles involving multiple frequencies evaluated at two or more points in time and numerous independent variables (age, gender, and ear) are under consideration.

We chose to analyze power-reflectance profiles of individual frequencies as our primary metric because this is the relevant feature derived from commercially available instrumentation and the one that clinicians would actually use to make inferences about normal or pathological states of the middle ear. Moreover, advanced textbooks on research design and statistics consider GT the most comprehensive technique available for estimating test measurement reliability (Schiavetti & Metz, 2006, p. 123), and they do not even recognize the absolute-difference method, as described in this article, as a metric of reliability (Maxwell & Satake, 2006).

Established diagnostic exemplars of this methodology include categories of tympanometric types (i.e., profiles of immittance shape as a function of positive and negative air pressures; see, e.g., Jerger, Jerger, & Mauldin, 1972) or on

Figure 3. Composite scatter plots for nine separate tonal frequencies for different variables studied. Data collected from Test 1 are plotted on the *y*-axis, and data collected from Test 2 are plotted on the *x*-axis.

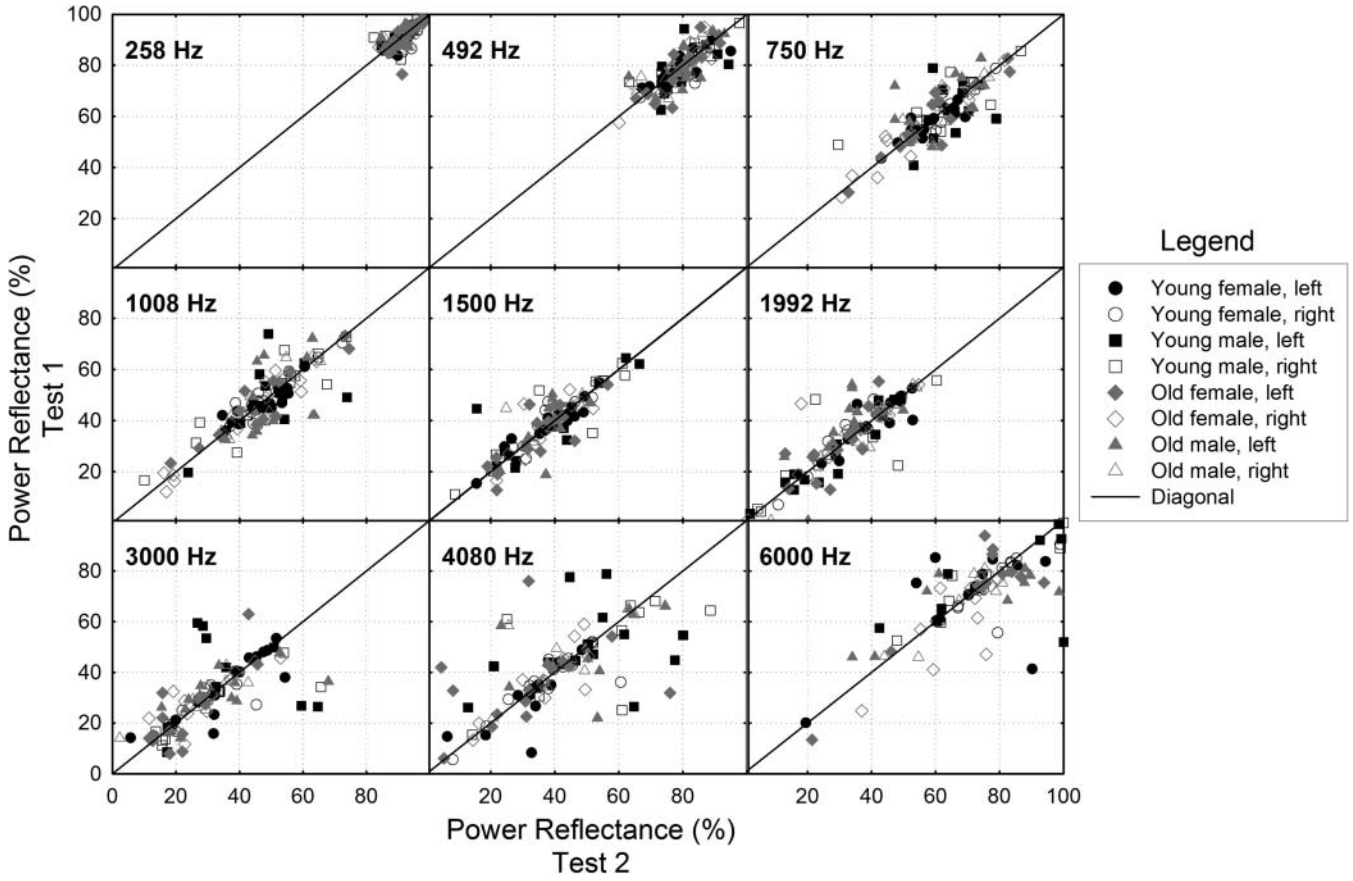


Table 1. Pearson's *r* correlations for chirps and tones.

	Chirps		Tones	
Frequency (Hz)	Left ear	Right ear	Left ear	Right ear
257	0.24	0.50**	0.75**	0.66**
304	0.17	0.70**		
398	0.44**	0.83**		
492	0.52**	0.83**	0.78**	0.83**
632	0.50**	0.85**		
750	0.60**	0.85**	0.82**	0.87**
796	0.57**	0.82**		
1007	0.53**	0.81**	0.82**	0.92**
1265	0.44**	0.69**		
1500	0.30*	0.60**	0.52**	0.85**
1593	0.34**	0.64**		
1992	0.55**	0.61**	0.58**	0.87**
2531	0.31*	0.63**		
3000	0.38**	0.73**	0.67**	0.84**
4007	0.51**	0.66**	0.64**	0.78**
5039	0.26	0.72**	0.79	0.89**

*$p < .05$. **$p < .01$.

more quantitative immittance typologies based on the theoretical model of Vanhuyse (see Margolis, Van Camp, Wilson, & Creten, 1985; Van Camp, Margolis, Wilson, Creten, & Shanks, 1986; Vanhuyse, Creten, & Van Camp, 1975). The potential for BMEPR measures to identify and delineate different pathological conditions of the middle ear provides the rationale and support for instituting a profile analysis because diagnostic interpretations are already being made using this approach (see, e.g., Allen et al., 2005; Feeney et al., 2003, 2009; Keefe & Simmons, 2003; Shahnaz, Longridge, & Bell, 2009).

Based on GT using chirps, the reliability for the overall effect (i.e., the power reflectance averaged across all frequencies) was 0.82; this is acceptable by standard convention (Cicchetti, 1994). The reliability of the frequency reflectance profile (the shape of the profile independent of height) was 0.56, which would be considered fair. Repeating the test and averaging the results yielded a profile reliability of 0.72, which is considered good. However, profiles involving ears and Ear × Frequency interactions were at levels conventionally considered to be poor (0.35). For tonal stimuli, the reliability of the overall effect was 0.86. The reliability of the frequency

Table 2. Chirp generalizability for reflectance and transmittance.

Effect/error	Reflectance			Transmittance		
	MS	ρ^2	*SE*	MS	ρ^2	*SE*
Subjects	2,360.880	0.8196	0.0241	55.95144	0.7788	0.0296
Subjects × Time	260.713			6.95624		
Subjects × Frequency	291.408	0.5621	0.0585	18.49421	0.4771	0.0699
Subjects × Frequency × Time	81.681			6.54339		
Subjects × Ear	609.427	0.3877	0.0818	11.19549	0.2581	0.0991
Subjects × Ear × Time	268.925			6.60240		
Subjects × Ear × Frequency	108.633	0.3484	0.0871	6.98064	0.0133	0.1319
Subjects × Ear × Frequency × Time	52.494			6.79778		

Note. MS = mean square; ρ^2 = generalizability coefficient; *SE* = standard error.

reflectance profile was 0.72. Repeating the test and averaging the results yielded a profile reliability of 0.75, and profiles involving ears and Ear × Frequency interactions were at 0.63. Thus, in comparison to chirps, the reliability for tones was better.

Using chirps, we made within-session test–retest reliability measures using Pearson's *r* for 16 individual frequencies ranging from 258 Hz to 5040 Hz. These data, which were collapsed across gender and age group, ranged from 0.30 to 0.85 for the right ear and from 0.18 to 0.57 for the left ear. These values agreed favorably with the adult data of Werner et al. (2010), who assessed 15 frequencies across a similar bandwidth (ranging from 281 Hz to 7336 Hz). Their correlation values from adults ranged from 0.28 to 0.95, where data were collapsed across gender and were presented for right ears only. In addition, Hunter and colleagues (2008) provided test–retest correlation coefficients for nine frequencies, ranging from 258 Hz to 6000 Hz, with data collapsed across age, gender, and ear and focused on children ranging in age from 3 days to 47 months. Their correlation values ranged from 0.68 to 0.97 and were higher than those reported in adults. Test–retest correlation values for tonal stimuli from the present study ranged from 0.52 to 0.92.

Werner and colleagues (2010) used a hybrid approach to assess reliability of power reflectance in infants and adults using three different metrics: (a) absolute-difference measures, (b) test–retest correlations of 15 individual frequency bands using Pearson's *r,* and (c) the cross-correlation method to examine reliability of the entire profile for the same ear and for left and right ears on two occasions. In their study, test–retest correlations for individual frequency bands were predominantly positive and were statistically significant. The highest correlation values were generally observed in the lower frequency range; the lowest correlations were observed in the higher frequency range. When test–retest correlations were averaged across frequency for the individual age groups (our computations are based on Table 1 of Werner et al., 2010), they were rank ordered, being lowest for 5- to 9-month-olds (M = 0.274, SD = 0.158, range: –0.05 to 0.44), intermediate for 2- to 3-month-olds (M = 0.401, SD = 0.139, range: 0.16 to 0.57), and highest for adults (M = 0.551, SD = 0.196, range: 0.30 to 0.95). In regard to adults, the cross-correlation method that was used to assess reliability of the shape of the profile and collapsed across age group produced a value of 0.85; the average between-ear cross-correlation was 0.84. However, it is noteworthy that the cross-correlation statistic is typically performed by comparing two time series, using a lag term to shift one function against the other as a way to determine the maximum correlation. Werner and colleagues indicated that "frequency was the lag variable" (p. 7), but they did not elaborate on the nonconventional use of this statistic. Applying the cross-correlation in this way results in profiles being aligned at different frequencies, a practice that contrasts with typical usage and one that has a questionable theoretical rationale because direct comparisons across frequency are not possible. Nevertheless, it is our contention

Table 3. Tone generalizability for reflectance and transmittance.

Effect/error	Reflectance			Transmittance		
	MS	ρ^2	*SE*	MS	ρ^2	*SE*
Subjects	988.5041	0.8546	0.0194	10.39252	0.8505	0.0200
Subjects × Time	77.5189			0.83962		
Subjects × Frequency	258.4342	0.7218	0.0372	3.06893	0.7888	0.0282
Subjects × Frequency × Time	41.7633			0.36238		
Subjects × Ear	321.4741	0.7541	0.0315	4.30111	0.7114	0.0386
Subjects × Ear × Time	42.9896			0.72534		
Subjects × Ear × Frequency	100.8526	0.6296	0.0495	0.83801	0.5346	0.0622
Subjects × Ear × Frequency × Time	22.9267			0.25412		

that the overall shape of the profile (e.g., power-reflectance values across frequencies) within individual ears is the primary feature of interest to clinicians using it for diagnostic purposes.

As noted above, we found that the reliability estimates of BMEPR and transmittance values were better for tones than for chirps. Although the precise reason for this disparity remains to be determined, several factors can be investigated in future studies. Consider that a broadband chirp is a continuously changing dynamic waveform in the time domain that is spectrally complex in the frequency domain. Even though the input waveform has a short duration (~5 ms), it is evaluated over a longer time epoch (1 s), such that acoustic reflections from the eardrum can be captured by the probe microphone, allowing for computations to be made on the metric of interest. With this in mind, single cycles of individual frequencies may be susceptible to interference and/or perturbations from physiologic events that may be present in a closed ear canal, such as respirations, blood flow pulsations from surface vessels, spontaneous or evoked otoacoustic emissions from the cochlea, subject movement, cord noise, and so forth. Thus, we speculate that, either alone or in combination, these factors can lead to increased variability in measurement. On the other hand, individual tonal stimuli are repetitive steady-state oscillations and perhaps are less affected by physiologic (state) and subject (trait) variables.

Ear-specific profiles of measurements across frequencies are commonplace in the field of diagnostic audiology; obvious examples include audiograms, iso-level/frequency profiles of distortion-product otoacoustic emissions (aka DPGrams), tympanograms, and BMEPR profiles. Thus, the problem of assessing reliability is ubiquitous in this field, and GT provides a robust solution to this problem. As noted above, we have already applied GT to DPGrams and evaluated the influences of time of day, stimulus frequency, stimulus SPL, and gender in adults with normal hearing (Cacace et al., 1996). We found that DPGrams were reliable measures within subjects over a contiguous 24-hr time period. Significant and reliable differences and interactions across frequency, SPL, and gender were also observed.

Another issue that requires further study concerns which characteristics of a profile might be meaningful clinically and how clinicians and researchers could potentially use this information to improve reliability estimates. As an example, Feeney et al. (2003) and Shahnaz, Bork, et al. (2009) have described trends associated with middle-ear pathology as alterations in broadband features (e.g., increased low-frequency power reflectance associated with otosclerosis). Because broadband changes such as those seen in otosclerosis involve lower-order trends, we speculate that it might be useful to fit these profiles with simple functions that capture these trends and smooth over measurement noise. Therefore, trend analysis might be a useful way to model these effects and, thereby, improve reliability. However, whether all clinically relevant information would be captured in lower-order trends remains to be seen; we suspect that this would not be the case. One alternative is to consider the possibility that the variability of a profile—made multiple times in the same individual—might represent a pathologic feature. In this context, it would be interesting to determine whether high variance in the residual after the removal of lower-order trends would be associated with clinically useful information. Pathological states of the middle ear that might show such effects include tympanic membrane perforations, monomeric tympanic membranes secondary to healed perforations, and ossicular discontinuities. Thus, further work in this area will be necessary to assess this hypothesis.

On the basis of the arguments presented in this article, and with respect to current clinical usage, profile analysis of BMEPR values appears to be the most relevant metric for evaluating and diagnosing middle-ear disorders. In contrast to tympanometry, which typically uses only one (226 Hz) or perhaps only a few specific probe-tone frequencies (e.g., 226 Hz, 660 Hz, 1200 Hz) to estimate characteristics of middle-ear function, power-reflectance techniques can rapidly measure hundreds of points over a much broader bandwidth and in a considerably shorter period of time (seconds). This is noteworthy because it allows for a more comprehensive account of energy transfer characteristics of the middle ear than is currently available from other methods. Combined with computer-controlled hardware and based on rapidity of measurement, BMEPR has many desirable attributes consistent with a viable clinical tool. Therefore, improving reliability of measurement is an essential requisite in the evolution of this method if it is to transition effectively from the laboratory to the clinic.

In conclusion, the reliability of BMEPR measurements is an important consideration in establishing this method for clinical decision making. The application of GT allows for a more comprehensive evaluation of these types of data compared with other approaches, and we advocate for the strategic use of this metric in future investigations.

Acknowledgments

Data were collected by the first and third authors as part of a capstone research project for satisfying the Doctor of Audiology degree; Sabahet Rizvi analyzed the tonal data, as part of a separate research project. Portions of this work were presented as a poster at the 2010 annual meeting of the American Auditory Society in Scottsdale, Arizona, and at the 2010 annual Michigan–Toledo P30 Post-ARO Podium and Poster Meeting in Ann Arbor, Michigan, sponsored by the Kresge Hearing Research Laboratory, University of Michigan. We thank Robert H. Margolis for constructive comments and suggestions.

References

Allen, J. B., Jeng, P. S., & Levitt, H. (2005). Evaluation of human middle ear function via an acoustic power assessment. *Journal of Rehabilitation Research and Development, 42,* 63–78.

Beers, A. N., Shahnaz, N., Westerberg, B. D., & Kozak, F. K. (2010). Wideband reflectance in normal Caucasian and Chinese school-aged children and in children with otitis media with effusion. *Ear and Hearing, 31,* 221–233.

Broglio, A. T. C., Zhu, W., Sopiarz, K., & Oark, Y. (2009). Generalizability theory analysis of balance error scoring system reliability in healthy young adults. *Journal of Athletic Training, 44,* 497–502.

Cacace, A. T., McClelland, W. A., Weiner, J., & McFarland, D. J. (1996). Individual differences and the reliability of 2F1–F2 distortion-product otoacoustic emissions: Effects of time-of-day, stimulus variables, and gender. *Journal of Speech and Hearing Research, 39,* 1138–1148.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Harcourt Brace.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speech reading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876–891.

Di Nocera, F., Ferlazzo, F., & Borghi, V. (2001). G theory and the reliability of psychophysiological measures: A tutorial. *Psychophysiology, 38,* 796–806.

Feeney, M. P., Grant, I. L., & Marryott, L. P. (2003). Wideband energy reflectance measurements in adults with middle-ear disorders. *Journal of Speech, Language, and Hearing Research, 46,* 901–911.

Feeney, M. P., Grant, I. L., & Mills, D. M. (2009). Wideband energy reflectance measurements of ossicular chain discontinuity and repair in human temporal bone. *Ear and Hearing, 30,* 391–400.

Hunter, L. L., Tubaugh, L., Jackson, A., & Propes, S. (2008). Wideband middle ear power measurement in infants and children. *Journal of the American Academy of Audiology, 19,* 309–324.

Jeng, P. S., Allen, J. B., Lapsey-Miller, J. A., & Levitt, H. (2008). Wideband power reflectance and power transmittance as tools of assessing middle-ear function. *Perspectives on Hearing and Hearing Disorders in Children, 18,* 44–57.

Jerger, J., Jerger, S., & Mauldin, L. (1972). Studies in impedance audiometry: Normal and sensorineural ears. *Archives of Otolaryngology, 96,* 513–523.

Keefe, D. H., & Simmons, J. L. (2003). Energy transmittance predicts conductive hearing loss in older children and adults. *The Journal of the Acoustical Society of America, 114,* 3217–3238.

Laenen, A., Vangeneugden, T., Geys, H., & Molenberghs, G. (2006). Generalized reliability estimation using repeated measures. *British Journal of Mathematical and Statistical Psychology, 59,* 113–131.

Margolis, R. H., Van Camp, K. J., Wilson, R. H., & Creten, W. L. (1985). Multifrequency tympanometry in normal ears. *Audiology, 24,* 44–53.

Maxwell, D. L., & Satake, E. (2006). *Research and statistical methods in communication sciences and disorders*. Boston, MA: Thomson Delmar Learning.

Mushquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analysis. *Behavioral Research Methods, 38,* 542–547.

O'Brian, N., O'Brian, S., Packman, A., & Onslow, M. (2003). Generalizability theory I: Assessing reliability of observational data in the communication sciences. *Journal of Speech, Language, and Hearing Research, 46,* 711–717.

O'Brian, S., Packman, A., Onslow, M., & O'Brian, N. (2003). Generalizability theory II: Application to perceptual scaling of speech naturalness in adults who stutter. *Journal of Speech, Language, and Hearing Research, 46,* 718–723.

Ramseyer, G. C. (1979). Testing the difference between dependent correlations using Fisher *z*. *Journal of Experimental Education, 47,* 307–310.

Roebroeck, M. E., Hariaar, J., & Lankhors, G. J. (1993). The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical Therapy, 73,* 386–395.

Rosowski, J. J., Nakajima, H. H., Hamade, M. A., Mahfoud, L., Merchant, G. R., Halpin, C. F., & Merchant, S. N. (2012). Ear-canal reflectance, umbo velocity, and tympanometry in normal-hearing adults. *Ear and Hearing, 33,* 19–34.

Scarsellone, J. M. (1998). Analysis of observational data in speech and language research using generalizability theory. *Journal of Speech, Language, and Hearing Research, 41,* 1341–1347.

Schiavetti, N., & Metz, D. E. (2006). *Evaluating research in communicative disorders* (5th ed.). Boston, MA: Pearson Education.

Shahnaz, N., Bork, K., Polka, L., Longridge, N., Bell, D., & Westerberg, B. D. (2009). Energy reflectance and tympanometry in normal and otosclerotic ears. *Ear and Hearing, 30,* 219–233.

Shahnaz, N., Longridge, N., & Bell, D. (2009). Wideband energy reflectance patterns in preoperative and post-operative otosclerotic ears. *International Journal of Audiology, 48,* 240–247.

Thornton, A. R. D., Marotta, N., & Kennedy, C. (2003). The order of testing effect in otoacoustic emissions and its consequences for sex and ear differences in neonates. *Hearing Research, 184,* 123–130.

Van Camp, K. J., Margolis, R. H., Wilson, R. H., Creten, W. L., & Shanks, J. E. (1986). Principles of tympanometry. *American Speech and Hearing Association Monographs, 24,* 1–88.

van de Geld, P. A., Oosterveld, P., van Waas, M. A., & Kuijpers-Jaqtman, A. M. (2007). Digital videographic measurement of tooth display and lip position in smiling and speech: Reliability and clinical application. *American Journal of Orthodontics and Dentofacial Orthopedics, 131,* e1–e8.

Vander Werff, K., Prieve, B., & Georgantas, L. (2007). Test–retest reliability of wideband reflectance measures in infants under screening and diagnostic test conditions. *Ear and Hearing, 28,* 669–681.

Vanhuyse, V. J., Creten, W. L., & Van Camp, K. J. (1975). On the W-notching of tympanograms. *Scandinavian Audiology, 4,* 45–50.

Werner, L. A., Levi, E. C., & Keefe, D. H. (2010). Ear-canal wideband transfer functions of adults and two- to nine-month-old infants. *Ear and Hearing, 31,* 1–12.

Appendix (p. 1 of 2)

A Concise Overview of Generalizability Theory (GT)

The conceptual basis of GT requires a detailed understanding of repeated measures analysis of variance (ANOVA); tutorial and computational methods are available to assist clinicians or researchers in this regard (see Di Nocera, Ferlazzo, & Borghi, 2001; Mushquash & O'Connor, 2006). However, we caution that there is no simple "cookbook" approach to data analysis with GT. This is due in part to the many design features that are possible within a given experimental framework. Thus, as would be the case for any experimental design using ANOVA, the approach taken will depend on the complexity of the experiment and the statistical model being used.

In Figure A1, two hypothetical cases involving four test sessions in three subjects are shown in diagrammatic form. The figure illustrates a case in which the relative ranking of each subject's test scores is consistent across test sessions (Panel A). In this case, the main effect of subjects is large relative to the interaction between subjects and test sessions. In a second case, the ranking of the subjects' test scores varies considerably across test sessions (Panel B). This results in a large interaction between subjects and test sessions relative to the main effect of subjects. The generalizability coefficient is a ratio of the main effect of subjects to the sum of that main effect and the interaction between subjects and test sessions. Thus, this ratio would be much larger in the first case than in the second.

The metric used in GT is expressed as the proportion of the total variance due to subjects that is common to the testing occasions of interest. For a single measure determined on two testing occasions, this can be computed as Pearson's *r*; for *k* testing occasions, it is computed as

$$\rho^2 = \sigma^2_{subj}/(\sigma^2_{subj} + \sigma^2_{err}), \quad (A1)$$

where σ^2_{subj} is the variance due to the main effects of subjects and σ^2_{err} is the variance due to error (i.e., the nonadditive or inconsistent effect of subjects across testing occasions).

As a relevant example, we present a "thought experiment" whereby the reliability of middle-ear power reflectance is assessed at 1500 Hz for the right ear on two occasions. Here, our sample consists of 56 adults without a history of middle-ear disease.

Results from the ANOVA are depicted in Table A1. Degrees of freedom, sum of squares, mean square, and expected mean square can be obtained from many statistical programs that are commercially available in the marketplace (e.g., SAS, SPSS, Statistica, etc.). Statistical programs such as these report *F* values and probability estimates (*p* values) associated with the effects of time of testing and their significance. The *F* for the effect of time is simply the ratio obtained by dividing the value in Row 2 by

Figure A1. Examples of tests with differing generalizability coefficients. In both plots, each of three subjects' scores on four test sessions is represented by a line. The main effect of subjects represents the variance in the average difference between subjects. This is large when the profiles are parallel. The interaction between subjects and sessions represents the extent to which the ordering of subjects varies across sessions; it is large when the profiles are nonparallel. In Panel A, one can see that the three subjects performed relatively consistently across the four test sessions. As a result, the main effect of subjects is large relative to the interaction between subjects and test sessions. In Panel B, one can see that the three subjects performed inconsistently across the four test sessions. As a result, the main effect of subjects is small relative to the interaction between subjects and test sessions.

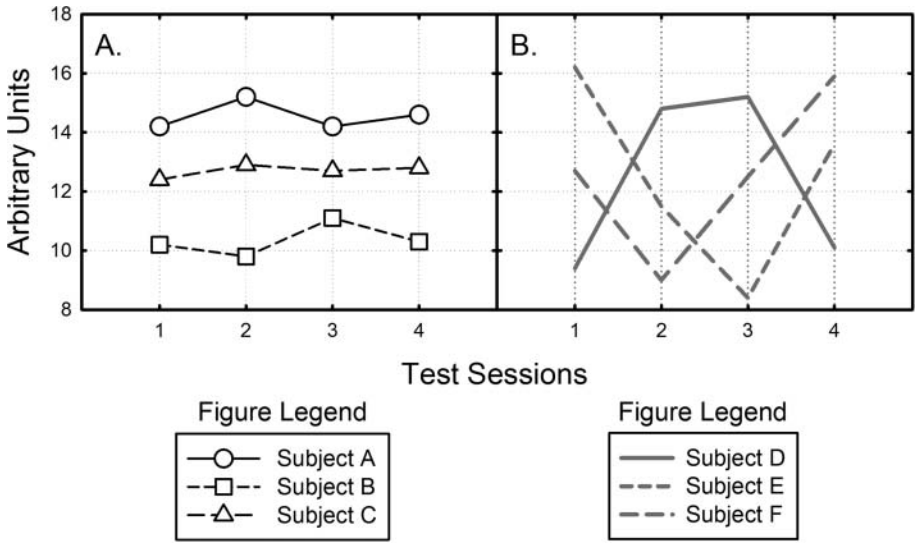


Table A1. Summary of example "thought experiment" of power reflectance at 1500 Hz in the right ear for 56 subjects tested on two occasions.

Source	*df*	SS	MS	EMS	Subjects pooled	Subjects single test
Time	1	13.99	13.99			
Subjects	55	14,794.53	268.99	$\sigma^2_{err} + k\,\sigma^2_{subj}$	0.917	0.847
Time × Subjects	55	1,224.78	22.27	σ^2_{err}		

Note. SS = sum of squares; MS = mean square; EMS = expected mean square.

that in Row 1 for the mean-square value. In this example, the p value is not significant ($p > .05$; result not shown). The expected mean square (*EMS*) for the effects of subjects is

$$EMS_{subj} = k\sigma_p^2 + \sigma_{err}^2, \tag{A2}$$

after Crocker and Algina (1986). The mean square (*MS*) for subjects represents the variance summed over all testing occasions and associated error. The *EMS* for the error is $MS_{time \times subj}$ and represents the nonadditive effect of subjects over time (i.e., the effect of test sessions), often computed as the residual in this design:

$$\sigma_{subj}^2 = (MS_{subj} - MS_{err})/k. \tag{A3}$$

A pooled estimate of the proportion of the variance in subject scores that is consistent across time (additive) can be obtained by the following equation:

$$\rho_{pooled}^2 = (MS_{subj} - MS_{err})/(MS_{subj} - MS_{err}) + MS_{err}. \tag{A4}$$

This represents the reliability of a composite of the test scores summed across time and is reported under "Subjects pooled" in Table A1. Because one is usually interested in the reliability of single tests, this is computed as

$$\rho^2 = [(MS_{subj} - MS_{err})/k]/[(MS_{subj} - MS_{err})/k + MS_{err}], \tag{A5}$$

and takes into account the estimate of single test variance from Equation A3 and is reported under "Subjects single test" in Table A1.

More complex models that include additional facets that interact with subjects can also be constructed. For example, if separate measurements were taken for each ear on several occasions, then an Ear × Subjects interaction could be computed. In this case, the model would be

$$\rho^2 = [(MS_{ear \times subj} - MS_{err})/ek]/[(MS_{ear \times subj} - MS_{err})/ek + MS_{err}], \tag{A6}$$

where e represents the number of ears and k represents the number of test sessions. The MS_{err} is now the value of $MS_{time \times ear \times subj}$. Thus, interactions between subjects and various measures repeated across subjects can be generated by substituting these interaction terms for the main effect terms in Equation A5. Conceptually, in a model including subjects, ears, and occasions, the generalizability coefficient associated with the main effects of subjects (MS_{subj}) represents the reliability of a score averaged across ears, whereas the generalizability coefficient associated with the Ear × Subjects interaction represents the reliability of a score as the difference between the ears. In fact, a wide variety of models can be generated following the logic of the general linear model (see Laenen et al., 2006).

Although statistical packages such as SAS, SPSS, or Statistica do not offer explicit tools for computing generalizability coefficients, the individual *MS* values are provided in the standard ANOVA summary table for a model including time as a repeated measure (i.e., in within-subject designs). Thus, generalizability coefficients can be readily computed with just a few simple operations. To aid in these computations, Mushquash and O'Connor (2006) provided a guide for users of SAS, SPSS, or MATLAB.