Development of a new standard laboratory protocol for estimation of the field attenuation of hearing protection devices: Sample size necessary to provide acceptable reproducibility

William J. Murphy[a] and John R. Franks
Hearing Loss Prevention Section, National Institute for Occupational Safety and Health, 4676 Columbia Parkway MS C-27, Cincinnati, Ohio 45226

Elliott H. Berger
E·A·R/Aearo Company, 7911 Zionsville Road, Indianapolis, Indiana 46268-1657

Alberto Behar
University of Toronto, 45 Meadowcliffe Drive, Scarborough, Ontario M1M2X8, Canada

John G. Casali
Virginia Tech, 250 Durham Hall, Blacksburg, Virginia 24061

Christine Dixon-Ernst
Alcoa Corporate Center, 201 Isabella Street, Pittsburgh, Pennsylvania 15212-5828

Edward F. Krieg
Monitoring Research & Statistics Activity, National Institute for Occupational Safety and Health, 4676 Columbia Parkway MS C-22, Cincinnati, Ohio 45226

Ben T. Mozo
Communications & Ear Protection Inc., 303 South Ouida Street, Enterprise, Alabama, 36331

Julia D. Royster and Larry H. Royster
Environmental Noise Consultants, P.O. Box 30698, Raleigh, North Carolina 27622-0698

Stephen D. Simon
Office of Medical Research, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, Missouri 64108

Carol Stephenson
Training and Education Systems Branch, National Institute for Occupational Safety and Health, 4676 Columbia Parkway MS C-10, Cincinnati, Ohio 45226

(Received 30 March 2003; accepted for publication 20 October 2003)

The mandate of ASA Working Group S12/WG11 has been to develop ''laboratory and/or field procedure(s) that yield useful estimates of field performance'' of hearing protection devices (HPDs). A real-ear attenuation at threshold procedure was selected, devised, tested for one earmuff and three earplugs via an interlaboratory study involving five laboratories and 147 subjects, and incorporated into a new standard that was approved in 1997 [Royster *et al.*, ''Development of a new standard laboratory protocol for estimating the field attenuation of hearing protection devices. Part I. Research of Working Group 11, Accredited Standards Committee S12, Noise,'' J. Acoust. Soc. Am. **99**, 1506–1526; ANSI, S12.6-1997, ''American National Standard method for measuring real-ear attenuation of hearing protectors'' (American National Standards Institute, New York, 1997)]. The subject-fit methodology of ANSI S12.6-1997 relies upon listeners who are audiometrically proficient, but inexperienced in the use of HPDs. Whenever a new method is adopted, it is important to know the effects of variability on the power of the measurements. In evaluation of protector noise reduction determined by experimenter-fit, informed-user-fit, and subject-fit methods, interlaboratory reproducibility was found to be best for the subject-fit method. Formulas were derived for determining the minimum detectable difference between attenuation measurements and for determining the number of subjects necessary to achieve a selected level of precision. For a precision of 6 dB, the study found that the minimum number of subjects was 4 for the Bilsom UF-1 earmuff, 10 for the E·A·R Classic earplug, 31 for the Willson EP100 earplug, and 22 for the PlasMed V-51R earplug. [DOI: 10.1121/1.1633559]

PACS numbers: 43.66.Vt, 43.66.Yw [DKW] Pages: 311–323

[a]Electronic address: wmurphy@cdc.gov

I. INTRODUCTION

Royster *et al.* (1996) examined the differences between subject-fit (SF) and informed-user-fit (IUF) hearing protector testing methods. Berger *et al.* (1998) addressed the relationship between the noise reduction measured in the laboratory and the noise reduction measured in occupational settings. The relationship of the variability of real-ear attenuation at threshold (REAT) measurements as a factor controlling sample sizes necessary for adequate statistical power remains an unresolved issue. In this paper, the statistical models for estimating the within-subject repeatability and between-subject and between-laboratory reproducibility are developed and applied to the estimation of the minimum detectable difference and sample-size estimates for REAT measurements.

Hearing protector testing as prescribed by several national and international standards consists of measuring occluded and unoccluded pairs of thresholds in a diffuse sound field with at least ten subjects. The attenuations are measured for at least seven third-octave noise bands (125, 250, 500, 1000, 2000, 4000, and 8000 Hz) as the numerical difference in decibels between occluded and unoccluded threshold pairs calculated for every subject's trial. The difference between the methods studied in this paper deals with the manner in which the subjects and experimenters participated in the fitting of the hearing protection device (HPD).

Royster *et al.* (1996) reported tests on subjects with both SF and IUF methods according to the instructions upon which Method B of ANSI S12.6-1997 (ANSI, 1997) was based. In the SF method, the subjects were provided with only the manufacturer's instructions for wearing the HPD; no experimenter involvement was permitted. They were instructed to fit the device as best they could and then threshold pairs were measured. The IUF method was equivalent to the experimenter-assisted fit in ANSI S12.6-1997 (ANSI, 1997) now commonly referred to as Method A. In this method, the subjects were given the device and the manufacturer's instructions and instructed to fit the HPD. While the subject fitted the device in the presence of fitting noise, the experimenter was permitted to coach the subject to obtain a better fit. As with the SF method, threshold pairs were collected from the subjects. Detailed descriptions of the data collection and subject instructions were given in Royster *et al.* (1996).

An additional study was conducted following the study reported by Royster *et al.* (1996). Subjects were tested at NIOSH (1995) and Virginia Tech University using the subject-fit and the experimenter-fit (EF) methods. The ANSI S3.19-1974 (ANSI, 1974) EF method protocol was used and was equivalent to the test methods presently used by HPD manufacturers to produce data from which noise reduction ratings (NRRs) can be determined. In the EF method, the experimenter fitted the HPD and instructed the subject not to adjust the device during the occluded-threshold tests. From hereon, the testing reported by Royster *et al.* will be referred to as the four-lab study (E·A·RCAL Laboratory, NIOSH Hearing Protector Laboratory, US Army Aeromedical Research Laboratory, Wright Patterson Armstrong Laboratories), while the additional testing performed by NIOSH and Virginia Tech will be referred to as the two-lab study. The results of both studies are reported here.

This paper examines the REAT distributions as a function of protector, fitting procedure, and test frequency. The REATs were analyzed with a multi-level analysis of variance that determined the standard deviations for within-subject repeatability, between-subject reproducibility, and between-laboratory reproducibility. Lastly, the minimum detectable differences at each frequency were determined for each device and fitting procedure. The minimum detectable differences were used to estimate sample sizes necessary to achieve a given desired resolution between measurements.

II. METHODS

A. Data sources

For both the four- and two-lab studies, pairs of occluded and unoccluded thresholds were collected using one-third-octave band noise stimuli centered at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. The hearing protectors tested, the test methods, and studies providing the data reported here are summarized in Table I.

1. Four-lab study

The HPDs that were selected for the four-lab study are described in Sec. II C and Fig. 1 of Royster *et al.* (1996). They included the E·A·R® Classic® foam earplug, the PlasMed, Inc. V-51R premolded earplug (five sizes), the Willson Safety Products EP100 premolded earplug (two sizes), and the Bilsom UF-1 earmuff. The devices were selected based upon availability of published real-world data, market-place popularity at the commencement of the studies, and diversity of protector types. The Working Group had selected more earplugs than earmuffs because it determined that earplugs provided a greater real-world estimation problem than did earmuffs (Casali and Park, 1991).

Subjects for the four- and two-lab experiments were selected from volunteers with pure-tone air-conducted hearing thresholds less than 25 dB HL *re* ANSI S3.6-1999 (ANSI, 1999) at each test frequency and normal tympanometry. Af-

TABLE I. Hearing protection devices, numbers of subjects, and methods of fit used in the four- and two-lab studies.

Devices	Methods	
	Four-lab study	Two-lab study
	E·A·RCAL (26 subjects) USAARL (24 subjects) NIOSH (24 subjects) WPAFB (24 subjects) (Royster *et al.*, 1996)	NIOSH (25 subjects) Virginia Tech (26 subjects) (Franks *et al.*, 2000)
Bilsom UF-1 earmuff	SF[a] IUF[b]	
Willson EP100 earplug	SF[a] IUF[b]	
E·A·R Classic earplug	SF[a] IUF[b]	SF[a] EF[c]
PlasMed V-51R earplug	SF[a] IUF[b]	SF[a] EF[c]

[a]SF: Subject fit as defined in ANSI S12.6-1997, Method B (ANSI, 1997).
[b]IUF: Informed-user fit equivalent to experimenter assisted fit as defined in ANSI S12.6-1997, Method A, (ANSI, 1997).
[c]EF: Experimenter fit as defined in ANSI S3.19-1974 (ANSI, 1974).

ter an initial session for hearing screening and audiometric training, each subject made eight visits to the lab. In each visit, two different HPDs were tested in separate sessions; each included two trials pairing sets of unoccluded and occluded sound-field thresholds. In total, four unoccluded and occluded paired thresholds were measured for each device, subject, and fitting method. Subjects received a brief rest break between the two sessions of a visit. Visits were separated by a minimum of 6 h, and all eight visits were required to take place within 21 days. Each laboratory recruited 24 subjects who were naive in the use and fitting of hearing protectors. The subject population was gender-balanced and the order of device testing was randomized to minimize learning effects that might occur. REAT means and standard deviations for the four-lab study were reported in Table II of Royster *et al.* (1996). The instructions regarding fitting and size selection are given in detail by Royster *et al.* (1996). For the sized earplugs, subjects were allowed to select the size that best fit their ear canals. The experimenter provided no assistance during the subject-fit portion of the testing and provided limited coaching during the informed-user fit testing.

2. Two-lab study

The two-lab study tested the V-51R and the E·A·R Plug using SF and EF methods. The NIOSH and Virginia Tech laboratories tested REAT for 25 and 26 subjects, respectively, first with the SF method and then with the EF method. The test methods and data have been reported in detail elsewhere (Franks *et al.*, 2000). After an initial session for hearing screening and audiometric training, each subject visited the lab on separate days for SF and EF testing. In each visit, three pairs of unoccluded and occluded sound-field thresholds were collected for two different HPDs in separate sessions. Subjects were permitted a brief rest break as needed between paired trials during a visit. These data were collected in order to obtain a direct comparison of the REATs measured from the two methods on the same subjects, rather than relying solely on the SF data from the four-lab study and upon the manufacturer's report of EF data.

B. Data analyses

The analyses of the data were conducted in three phases. First, the distributions of the REAT histograms for each fitting method, device, and test frequency were tested for normality and modality. Second, the REAT distributions were analyzed with a multi-level analysis of variance (ANOVA) with trials nested within subjects and subjects within laboratories. Third, the number of subjects necessary to achieve a given resolution in the REAT data were calculated for a confidence level $1-\alpha=0.84$ and a power of $\beta=0.80$. The confidence level represents one standard deviation away from the mean. The confidence level is one minus the probability of committing a type I error, incorrectly identifying an effect. The power is one minus the probability of committing a type II error, failure to identify a real effect.

The statistical design accounts for three sources of variation: trial-to-trial (σ_{trial}), subject-to-subject (σ_{subject}), and laboratory-to-laboratory ($\sigma_{\text{laboratory}}$). From these sources, estimates for the within-subject repeatability, between-subject reproducibility, and interlaboratory reproducibility were calculated. The data from each fitting method (EF, IUF, and SF) were analyzed separately. The EF data represent the two-lab study; the IUF data represent the four-lab study and the SF data represent the pooled results from four- and two-lab studies. Learning effects due to repeated tests with the same protector (within-subject repeatability), subject effects for a given hearing protector (between-subject repeatability), or laboratory effects (between subject/between laboratory repeatability) were assessed by analysis of variance. The geographical separation of laboratories prevented using the same test subjects for directly assessing interlaboratory effects.

III. RESULTS

A. Distribution of REATs

The data for each HPD were pooled across the labs and examined by frequency and fitting method. The REAT means and standard deviations from the four-lab study have been reported for the SF and IUF methods for all four devices in Royster *et al.* (1996). Royster *et al.* did not examine the normality and modality of the REAT distributions. While the results of the two-lab study were not substantially different from the four-lab study, the additional data for SF and EF methods provided the impetus for this analysis. The REAT distributions for the pooled data from both studies are shown in Figs. 1–3. The figures depict at each test frequency histograms of the attenuation data sorted into 3-dB-wide bins. For example, if an attenuation estimate for a subject was greater than or equal to −1.5 dB and less than 1.5 dB, it was counted as one occurrence in the 0-dB bin. Similarly, the other bins were centered on multiples of 3 dB. The width of each bar represents the proportional number of occurrences at each attenuation level. In the upper left panel of Figs. 1–3, a scale bar the width of 50 occurrences is shown. The color represents a transition from minimum to maximum (blue to red), where the red bar indicates the maximum of the distribution for a given test frequency. Bins with a value of zero are not given a color or width. The diamond symbol represents the mean attenuation for a given test frequency.

Each distribution was tested for normality with the SAS univariate procedure (SAS, 1998) and was compared to three different probability distributions: normal, gumbel, and mixed-normal. In an extension to the maximum likelihood analysis in Murphy *et al.* (2002), the distribution was determined to be better estimated by one probability distribution or another. For instance, a distribution might be more likely to have been sampled from a gumbel distribution than from a normal distribution. The probability estimates of the maximum likelihood procedures were used according to Table II to classify whether a distribution was normal, gumbel, mixed-normal, or non-normal. If the SAS univariate procedure showed that the distribution was significantly different from a normal distribution, then the remainder of the tests were used to determine the class. In Figs. 1–3, the classifications of the distributions (non-normal, gumbel, and mixed-normal) are shown for each protector at each test frequency

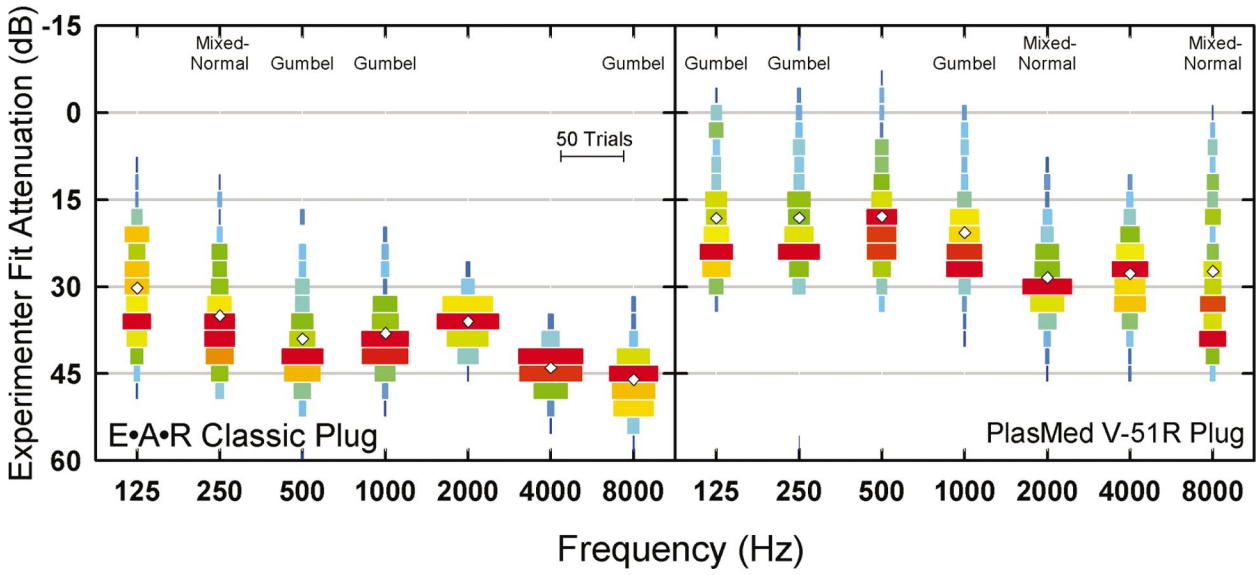

FIG. 1. Histogram-frequency plot of the attenuation measurements for the experimenter-fit method. The EF method data were collected as part of the two-lab study and represent 51 subjects with three occluded–unoccluded trials per subject per protector. The width of the bars represent the number of trials where subjects achieved an attenuation within a given 3-dB bin. A scale of 50 trials is shown in the upper left panel. The colored bars indicate the maximum (red) and minimum (blue) number of occurrences of a particular attenuation within a given frequency band. The colors were independently scaled for each frequency band distribution. The diamond symbols denote the mean of the distribution. The shape of each distribution has been tested and the classification according to Table II are listed above each histogram: ''Gumbel,'' ''Mixed-normal,'' and ''Non-normal.'' Histograms without a label are not significantly different from normal.

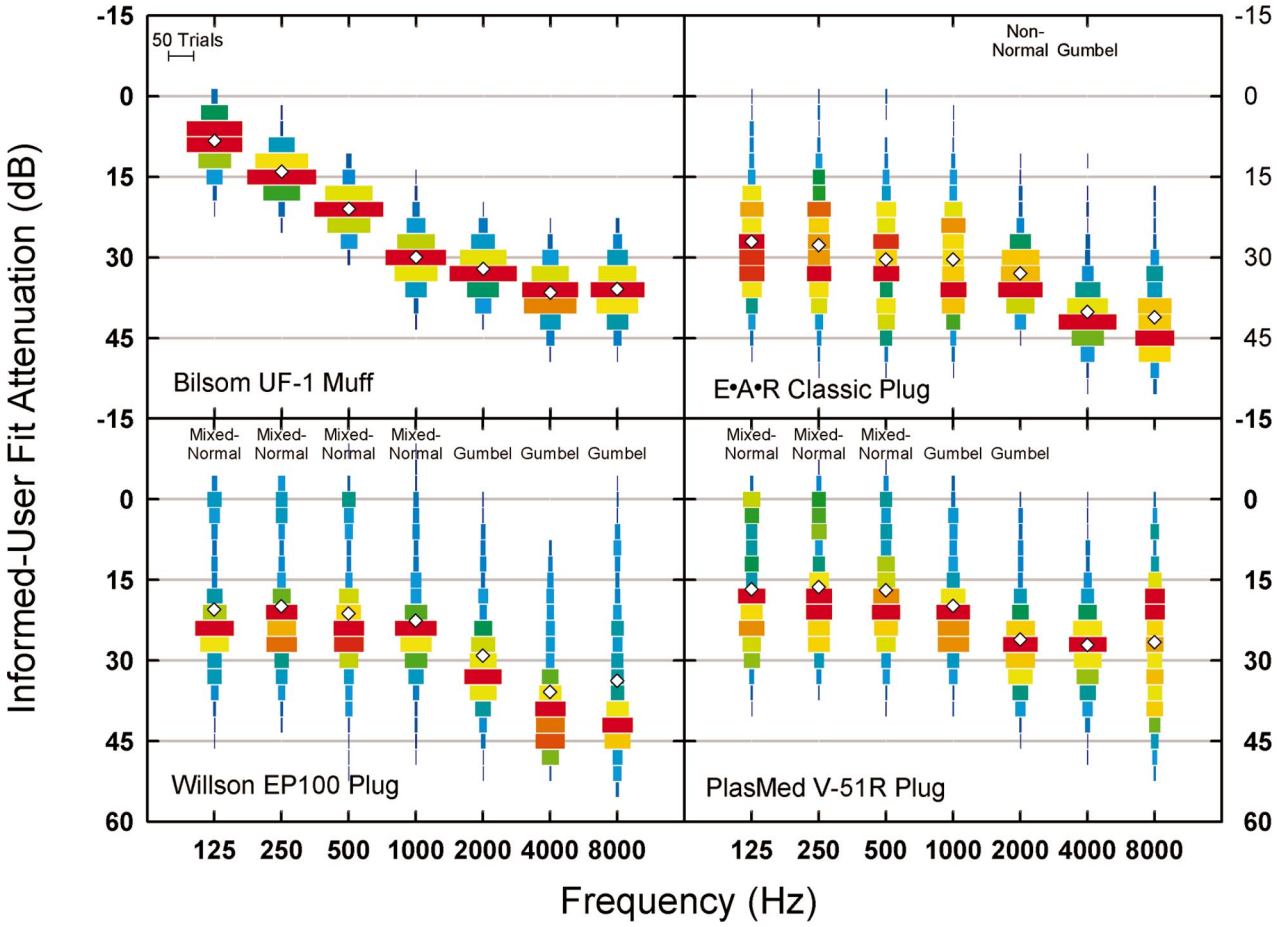

FIG. 2. Histogram-frequency plot of the attenuation measurements for the informed-user fit method. The IUF method data were collected as part of the four-lab study and represent 96 subjects with four occluded-unoccluded trials per subject per protector.

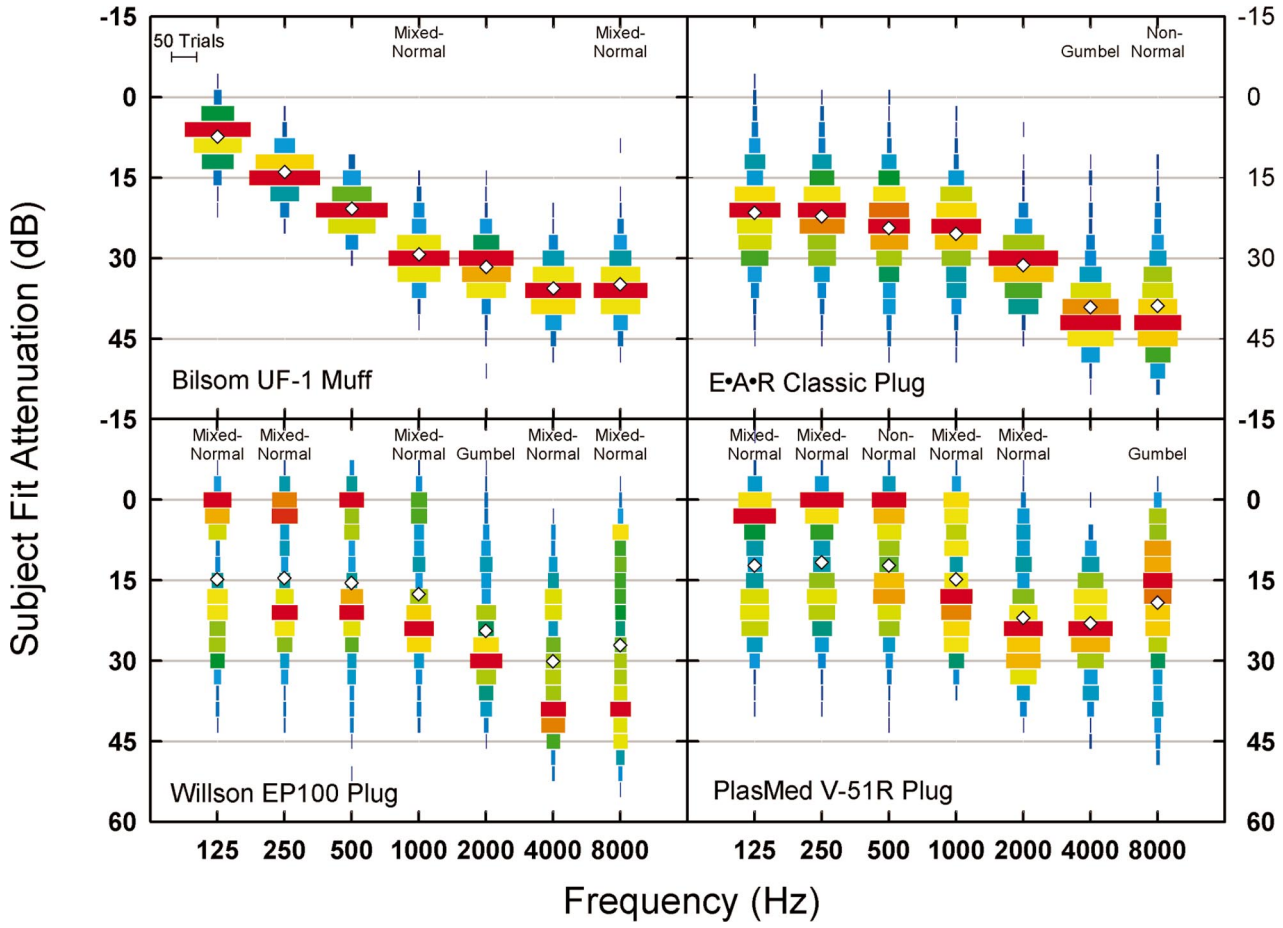

FIG. 3. Histogram-frequency plot of the attenuation measurements for the subject-fit method. The SF method data were collected as part of the four- and two-lab studies and represent 147 subjects with four occluded-unoccluded trials per subject per protector in the four-lab study and three occluded-unoccluded trials per protector in the two-lab study.

above each distribution. If a distribution was not significantly different from normal, there is no classification label. Although the figures do not display individual laboratory data, normal, gumbel, mixed-normal, and non-normal distributions are evident in the individual laboratory data as well as the pooled results of all labs.

The two-lab EF data for the E·A·R plug and V-51R earplug are shown in Fig. 1. The REAT distribution for 250 Hz was judged to be mixed-normal for the E·A·R plug. The REAT distributions for 500, 1000, and 8000 Hz were judged to be best fit by the gumbel distribution for the E·A·R plug. The EF data for the V-51R earplug were normal at 500 and 4000 Hz. The mixed-normal model yielded a better fit at 2000 and 8000 Hz. The distributions for 125, 250, and 1000 Hz were best fit by the gumbel distribution.

In Fig. 2, four-lab IUF data for the Bilsom UF-1 earmuff exhibited normal distributions for all frequencies. The IUF REAT distributions for the E·A·R plug were non-normal for 2000 Hz and best fit by the gumbel distribution for 4000 Hz. The REAT distributions for the EP100 earplug were mixed-normal for 125, 250, 500, and 1000 Hz. The distributions at 2000, 4000, and 8000 Hz were best fit with the gumbel model. The distributions at lower frequencies exhibited a clear tendency of one mode centered around 0 dB while at higher frequencies the distributions are skewed. The V-51R REAT distributions were best fit with the gumbel model at

TABLE II. Rules for classification of REAT distributions in Figs. 1–3 for a significance $p<0.05$. For instance, the probabilites for the EP100 earplug SF REAT distribution at 2000 Hz were $p_{SAS}=0.0042$, $p_{MLEGumbel\text{-}Normal}=0.015$, $p_{MLEMixedNormal\text{-}Normal}=0.0044$, and $p_{MixedNormal\text{-}Gumbel}=0.19680$. According to the rules, this distribution was classified as Gumbel.

Classification	SAS univariate normality	Maximum likelihood gumbel-normal	Maximum likelihood mixed normal-normal	Maximum likelihood mixed normal-gumbel
Normal	Not significant	Not applicable	Not applicable	Not applicable
Mixed-normal	Significant	Significant	Significant	Significant
Gumbel	Significant	Significant	Significant	Not significant
Mixed-normal	Significant	Not significant	Significant	Not applicable
Gumbel	Significant	Significant	Not significant	Not applicable
Non-normal	Significant	Not significant	Not significant	Not applicable

TABLE III. Standard deviations for the trial effects, σ_{trial}, for each device, fit, and frequency in decibels (dB).

		Frequency (Hz)						
Device	Fit	125	250	500	1000	2000	4000	8000
Bilsom	IUF	2.5	2.5	2.7	2.6	2.5	2.5	3.1
UF-1 earmuff	SF	2.2	2.5	2.7	2.5	2.6	2.7	3.1
EAR	EF	5.1	4.5	4.5	3.5	2.4	2.4	2.5
Classic earplug	IUF	4.9	4.9	4.5	4.2	3.1	3.3	3.9
	SF	5.7	5.5	5.4	4.7	3.1	3.7	4.5
Willson	IUF	5.8	6.0	5.8	5.2	4.8	5.0	6.5
EP100 earplug	SF	6.6	6.7	6.9	5.7	5.2	5.9	7.9
PlasMed	EF	5.6	5.1	4.5	4.4	4.0	4.2	5.6
V-51R earplug	IUF	6.2	5.9	6.0	5.5	4.5	5.1	6.8
	SF	5.7	5.5	5.2	5.1	4.6	3.9	5.4

1000 and 2000 Hz. The mixed-normal model provided a better fit at 125, 250, and 500 Hz. Similar trends of low attenuation at 125, 250, and 500 Hz were evident for the V-51R REAT distributions as are evident in the EP100 data.

In Fig. 3, the combined two- and four-lab SF REAT distributions for the Bilsom UF-1 muff were mixed-normal at 1000 and 8000 Hz. The REAT distribution for the E·A·R plug was best fit with the gumbel model at 4000 Hz and at 8000 Hz was non-normal. For the EP100 earplug, the REAT distributions were mixed-normal at 125, 250, 1000, 4000, and 8000 Hz. At 2000 Hz the distribution was best fit with the gumbel model. The EP100 REAT distribution at 500 Hz was not significantly different from a normal distribution, but it has the features of a bimodal distribution. The V-51R REAT distributions were non-normal at 500 Hz, mixed-normal at 125, 250, 1000, and 2000 Hz, and gumbel at 8000 Hz. The distribution at 4000 Hz was normally distributed. The low attenuation modes were apparent for the EP100 and V-51R earplugs.

B. Standard deviations

A multi-level analysis of variance (Netter, 1990) was used to estimate the standard deviations for laboratory, subject, and trial effects. The statistical model was

$$Y_{ijk} = \mu + \text{Trial}_{k(ij)} + \text{Subject}_{j(i)} + \text{Lab}_i, \quad (1)$$

where Y_{ijk} is the measured attenuation, μ is the real attenuation, and $\text{Trial}_{k(ij)}$ is the random error term for the kth trial within the jth subject and ith laboratory, $\text{Subject}_{j(i)}$ is the random error term for the jth subject within the ith laboratory, and Lab_i is the random error term for the ith laboratory. The ANOVA calculations were performed using the S-Plus software package (S-Plus, 2002). The model results are presented in Tables III–V.

Table III presents the standard deviations in decibels for the trial effects, σ_{trial}, determined from both the four- and two-laboratory studies for each of the fitting methods. Smaller standard deviations indicate less variability in the effect of the experimenter or subject on fitting the hearing protector consistently. Comparing the four sets of IUF and SF data, the Bilsom earmuff exhibited the smallest standard deviations, which is not surprising since earmuffs are easier to fit and offer fewer opportunities for improper fitting. The E·A·R plug tended to have the next smallest standard deviations. The EP100 and V-51R devices tended to have larger standard deviations. In general, as experimenter involvement decreased, σ_{trial} increased.

Table IV presents the standard deviations for the subject effects, σ_{subject}. Again, comparing the four sets of IUF and SF data, the Bilsom earmuff exhibited the smallest standard deviations. The E·A·R plug had the next smallest deviations followed by the V-51R and EP100, respectively. The small

TABLE IV. Standard deviations for the subject effects, σ_{subjects}, for each device, fit, and frequency in decibels (dB).

		Frequency (Hz)						
Device	Fit	125	250	500	1000	2000	4000	8000
Bilsom	IUF	2.9	2.0	1.9	2.9	2.4	2.8	2.9
UF-1 earmuff	SF	2.8	2.2	2.0	2.8	3.2	2.9	3.8
E·A·R	EF	5.6	6.2	5.6	4.7	2.6	2.0	3.8
Classic earplug	IUF	5.6	5.8	5.9	5.0	3.3	3.5	3.9
	SF	5.5	5.1	6.1	5.1	3.6	4.4	5.7
Willson	IUF	8.5	8.2	8.6	7.4	6.9	7.9	9.9
EP100 earplug	SF	9.6	9.4	10.4	9.4	8.7	9.4	11.5
PlasMed	EF	6.7	6.8	6.8	6.3	5.2	5.0	9.9
V-51R earplug	IUF	7.1	6.9	6.8	6.4	6.7	5.2	9.1
	SF	8.5	8.2	8.3	8.5	8.5	6.6	10.1

TABLE V. Standard deviations for the laboratory effects, $\sigma_{\text{laboratory}}$, for each device, fit, and frequency in decibels (dB). Standard deviations less than 0.05 were rounded to 0.0 dB.

		Frequency (Hz)						
Device	Fit	125	250	500	1000	2000	4000	8000
Bilsom	IUF	0.8	0.3	0.8	1.6	1.5	1.3	1.1
UF-1 earmuff	SF	0.0	0.0	0.5	1.8	1.5	1.4	0.7
E·A·R	EF	2.1	0.9	0.0	0.0	0.4	0.0	0.0
Classic earplug	IUF	4.4	5.4	5.8	5.6	2.7	2.4	3.8
	SF	1.0	1.3	1.6	2.3	2.0	2.4	2.7
Willson	IUF	2.9	2.6	3.2	2.7	3.3	3.0	4.5
EP100 earplug	SF	1.0	0.0	0.0	0.0	2.1	1.5	2.2
PlasMed	EF	0.0	0.0	0.0	0.0	1.2	0.0	0.0
V-51R earplug	IUF	2.9	3.0	2.7	1.7	2.2	2.4	4.4
	SF	0.8	1.2	1.0	1.2	1.9	1.8	0.0

variance for the Bilsom earmuff is indicative of the consistency of the fit across subjects. As with σ_{trial}, σ_{subject} increased with decreasing experimenter intervention.

Table V presents the standard deviations for the laboratory effects, $\sigma_{\text{laboratory}}$. As can be seen, the standard deviations for the SF method, where there was minimal experimenter involvement, were almost always lower than the values for the IUF method where the experimenter provided advice. However, $\sigma_{\text{laboratory}}$ for the EF method, where the experimenter was fully involved, were generally lower. For several frequencies and fitting methods, the standard deviations were less than 0.05 dB and were rounded to 0.0 dB.

C. Repeatability and reproducibility

Within-subject repeatability is a measure of the consistency of attenuation across trials for the same sample of subjects and hearing protector. Within-subject repeatability is computed with σ_{trial} listed in Table III. Between-subject reproducibility is a measure of the consistency of attenuation across subjects and trials for a hearing protector. Between-subject reproducibility is computed with the σ_{trial} and σ_{subject} from Tables III and IV. If there is no change of the testing protocol over time, then between-subject reproducibility measures consistency when two hearing protectors are tested with different subject panels. Between-laboratory reproducibility incorporates the lab-to-lab standard deviation, $\sigma_{\text{laboratory}}$, and measures the consistency of attenuations across laboratories. If there is a negligible amount of lab-to-lab variation, then between-subject reproducibility measures the consistency between two laboratories. While subjects were randomly sampled, the laboratories were not randomly selected as would be necessary to give a true estimate of laboratory effects.

Within-subject repeatability was estimated with the equation,

$$\sigma_{\text{within-subject}} = \sqrt{\frac{\sigma_{\text{trial}}^2}{(n_s n_t)}}, \tag{2}$$

where σ_{trial}^2 was the trial-to-trial variance, n_s was the number of subjects, and n_t was the number of trials per subject.

Between-subject reproducibility was estimated with the equation

$$\sigma_{\text{between-subject}} = \sqrt{\frac{\sigma_{\text{subject}}^2}{n_s} + \frac{\sigma_{\text{trial}}^2}{(n_s n_t)}}, \tag{3}$$

where $\sigma_{\text{subject}}^2$ was the subject-to-subject variance.

When $\sigma_{\text{laboratory}}$ is zero or negligible, then reproducibility $\sigma_{\text{between-subject}}$ represents consistency between laboratories. When $\sigma_{\text{laboratory}}$ is large, then comparisons of hearing protectors between laboratories are inappropriate. Between-laboratory reproducibility was estimated with the equation

$$\sigma_{\text{between-laboratory}} = \sqrt{\sigma_{\text{laboratory}}^2 + \frac{\sigma_{\text{subject}}^2}{n_s} + \frac{\sigma_{\text{trial}}^2}{(n_s n_t)}}, \tag{4}$$

where $\sigma_{\text{laboratory}}^2$ was the laboratory-to-laboratory variance. The standard deviation estimates of $\sigma_{\text{within-subject}}$, $\sigma_{\text{between-subject}}$, and $\sigma_{\text{between-laboratory}}$ calculated from Tables III–V for each device and each test frequency are shown in Figs. 4–6, respectively, using $n_s = 20$ and $n_t = 2$. These are similar to the repeatability and reproducibility concepts recommended for use in interlaboratory studies by ISO 5725-2 (ISO, 1994). Results for $n_t = 3$ or 4 are not shown, but can easily be calculated from these formulas.

1. Within-subject repeatability

The standard deviations for within-subject repeatability, $\sigma_{\text{within-subject}}$, for the Bilsom UF-1 earmuff in Fig. 4 exhibit little difference between SF and IUF methods and across test frequencies.

The standard deviations for all fit protocols are greater for the E·A·R plug than for the earmuff except at 4000 and 8000 Hz, EF fit. The standard deviations are dependent upon test frequency, being the lowest at 2000 Hz. As well, the magnitudes of the standard deviations increase above 2000 Hz as the amount of experimenter intervention decreases; the lowest standard deviations can be observed for the EF method while the highest standard deviations are for the SF method.

The standard deviations for the EP100 earplug are the highest of the four devices tested during these investigations. The standard deviations for the SF method are generally 0.1

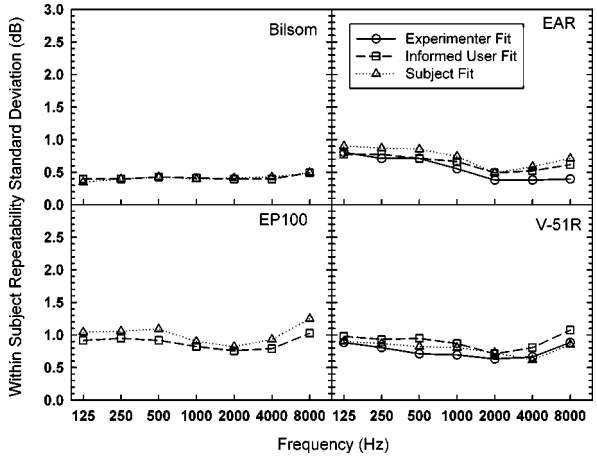

FIG. 4. The standard deviations for within-subject repeatability for each protector by fitting method and frequency. The REAT data was analyzed with a multi-level ANOVA where within-subject effects were nested within subjects and between-subject effects were nested within laboratory effects. The Bilsom earmuffs had the lowest variance and exhibited little difference between fitting methods. The other devices exhibited more variance and a trend to increased variability with decreased experimenter involvement. Experimenter-fit data are from the two-lab study, while subject-fit data are from the four-lab study.

to 0.2 dB greater than the standard deviations for the IUF method. Again, the magnitudes of the SF standard deviations are frequency dependent, being lowest at 2000 Hz.

The V-51R earplug exhibits higher standard deviations than the E·A·R plug, but lower than the EP100 earplug. As with the other earplugs, standard deviations tended to be lowest at 2000 Hz. However, for this earplug, the SF values were somewhat lower than the IUF values except at 2000 Hz where they were equal. The EF method gave the lowest standard deviations where the EF and SF methods values were virtually identical.

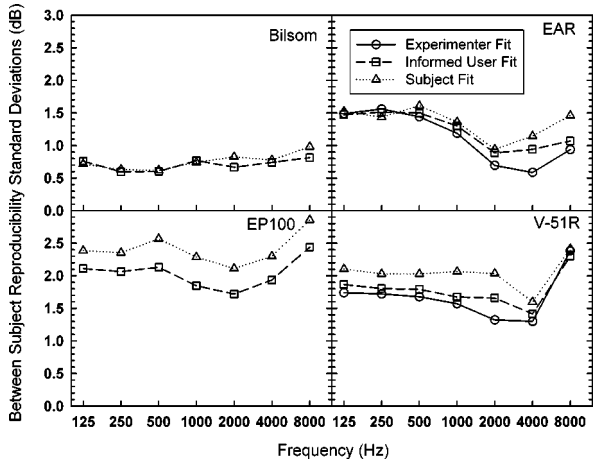

FIG. 5. The standard deviations for between-subject reproducibility for each protector by fitting method and frequency. The REAT data was analyzed with a multi-level ANOVA where within-subject effects were nested within subjects and between-subject effects were nested within laboratory effects. The Bilsom earmuffs exhibited the smallest standard deviations with almost identical results between IUF and SF fitting methods. The earplugs exhibited comparable standard deviations across devices with a slight trend for increased standard deviations with decreasing experimenter involvement. Experimenter-fit data are from the two-lab study, while subject-fit data are from the four-lab study.

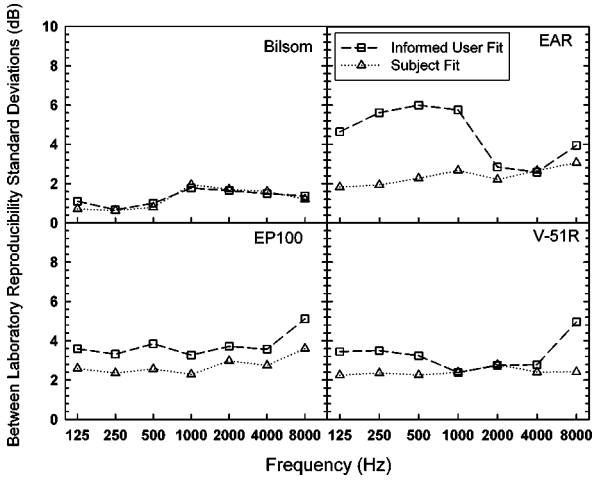

FIG. 6. The standard deviations for between-laboratory reproducibility. The magnitude of the standard deviations, greater than 2 dB, for the earplugs indicate poor reproducibility. The data from the Bilsom earmuffs indicate highly reproducible results across labs. The E·A·R Plug exhibits greater standard deviations for the IUF method. The standard deviations for the IUF method are greatest for almost all test frequencies for every device.

The results for the within-subject repeatability show consistent performance across test methods. The repeatability standard deviations depend upon the type of protector. Although slight differences exist between the standard deviations as a function of experimenter involvement, the differences are too small to be considered meaningful.

2. Between-subject reproducibility

The standard deviations for between-subject reproducibility, $\sigma_{\text{between-subject}}$, for each of the tested hearing protectors are shown in Fig. 5. In general the lowest standard deviations are for the Bilsom UF-1 earmuff and the highest are for the EP100 premolded earplug. The SF standard deviations are approximately equal to the IUF standard deviations for the Bilsom UF-1 earmuff. For the EP100 premolded earplug, and the V-51R premolded earplug, the IUF standard deviations for reproducibility are uniformly less than the SF standard deviations. For the E·A·R plug, the IUF standard deviations are also less than the SF standard deviations except at 250 Hz. For the Bilsom UF-1, differences between IUF and SF are trivial. However, these differences are not trivial for the E·A·R plug at 4000 and 8000 Hz, for the EP100 at all frequencies, and for the V-51R at all frequencies except 4000 and 8000 Hz.

In the two-lab study, the SF standard deviations are greater than the EF standard deviations for the V-51R earplug below 8000 Hz. For the E·A·R plug, the SF standard deviations are greater than the EF standard deviations except at 250 Hz.

3. Between-laboratory reproducibility

The standard deviations for between-laboratory reproducibility, $\sigma_{\text{between-laboratory}}$ are shown in Fig. 6. The smallest deviations are those of the Bilsom UF-1 earmuff and the largest are for the E·A·R plug IUF data. In contrast to the between-subject reproducibility, the SF standard deviations

are less than or approximately equal to the IUF standard deviations for nearly every frequency and protector.

D. Minimum detectable differences and sample size estimates

For a method to be useful in providing numbers for rating systems, the predictiveness of the data must be determined. In general, as sample size increases, so does the precision of the predictions of protector performance. In these cases, when the desired precision for a protector is established, the number of subjects necessary to achieve a given resolving ability can be estimated for a given confidence level, $1-\alpha$, and power, $1-\beta$ (e.g., confidence level of 84%, $\alpha=0.16$; power of 80%, $\beta=0.20$). The predictive precision, or resolution, of a hearing protector may be expressed as the smallest difference in two attenuation measurements that can be reliably detected for a given confidence level and a power level.

The confidence level was selected based upon the approach used in determining the noise reduction rating for subject fit data (NRR_{SF}); the calculation subtracts one standard deviation from the mean attenuation which yields an minimum attenuation estimate for 84% of the population at a given frequency (Franks *et al.*, 2000). The power level of 0.80 typically selected for behavioral data such as hearing thresholds assumes normality of the data.

1. Minimum detectable differences

Power calculations for testing both the within-subject repeatability and between-subjects reproducibility involve a simple modification of the formulas for power for a two-sample *t*-test (Rosner, 1990). For a hypothesis without multiple sources of variation and assuming the data are drawn from a normal distribution, the minimum detectable difference, D, with power, $1-\beta$ and confidence level, $1-\alpha$, is

$$D=(\text{Probit}(1-\alpha)+\text{Probit}(1-\beta))\frac{\sqrt{2}\sigma}{\sqrt{n}}, \tag{5}$$

where the Probit function provides the appropriate percentile value from a standard normal distribution (SAS, 1998). The minimum detectable difference is the value below which differences in real-ear attenuations at threshold are statistically insignificant.

Notice that the final term in the above equation, $\sqrt{2}\sigma/\sqrt{n}$, is simply the standard error of the difference between the means. Assuming that the between-subject reproducibility hypothesis to be tested is based on n_s subjects with n_t trials apiece, the minimum detectable difference becomes

$$D=(\text{Probit}(1-\alpha)+\text{Probit}(1-\beta))\sqrt{2}\sigma_{\text{between-subject}}, \tag{6}$$

where $\sigma_{\text{between-subject}}$ from Eq. (3) replaces $\sigma/\sqrt{n}$ in Eq. (5). As expected, increasing the number of subjects, n_s, decreases the contributions of σ_{subject} and σ_{trial} to the standard error term. Quite often, the minimum detectable difference will be smaller than the meaningful difference and it is always more influenced by the number of test subjects than the number of repetitions of test conditions (trials) per subject. Increasing the number of trials, n_t, decreases only the contribution of σ_{trial} to the standard error. The between-subject reproducibility test is valid when the between-laboratory reproducibility is small. As a rule of thumb, $\sigma_{\text{between-laboratory}}$ should be no more than 10% of the minimum detectable difference. The effect of subject sample size and number of trials is illustrated in Fig. 7. The σ_{trial} and σ_{subject} were assumed to be 5 dB. Doubling the number of subjects improves the minimum detectable difference more than doubling or tripling the number of tests per subject.

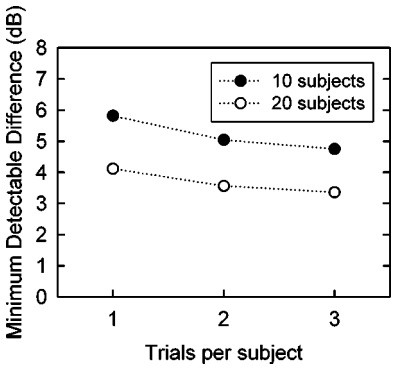

FIG. 7. Effects of number of subjects and number of trials per subject on the minimum detectable difference. The example shown is for the Bilsom UF-1 earmuff. The same relationship applies to the other devices tested—doubling the number of subjects is more effective than doubling the number of trials per subject. The parameters used to calculate the data points were $\sigma^2_{\text{trial}}=5$ dB, $\sigma^2_{\text{subject}}=5$, $n_s=10$, 20, and $n_t=1$, 2, 3.

Since the within-subject repeatability test uses the same subjects for the hearing protectors under test, the between-subject source of variation that would be nested within protectors is eliminated. For this design, the minimum detectable difference may be computed as

$$D=(\text{Probit}(1-\alpha)+\text{Probit}(1-\beta))\sqrt{2}\sigma_{\text{within-subject}}. \tag{7}$$

The within-subject repeatability test is valid so long as the same subjects are used for testing both hearing protectors. If different subjects are used, as might be required if there is a time gap between the testing of the two hearing protectors, then one should use the between-subject reproducibility hypothesis. When using this hypothesis, a laboratory needs to demonstrate that there are no significant trends over time relative to the size of the difference they are trying to detect. Table VI shows the minimum detectable differences calculated using $1-\alpha=0.84$, $1-\beta=0.80$, $n_s=20$ and $n_t=2$, and the between-subject reproducibility found in Table IV. For the Bilsom earmuff, the minimum detectable differences ranged between 1.6 at 250 and 500 Hz and 2.6 dB at 8000 Hz for the SF method. For the E·A·R plug, the SF method yielded differences ranging from 2.5 to 4.2 dB. For the EP100 and V-51R, the ranges of the minimum detectable differences for the SF method were 5.5 to 7.4 dB and 4.2 to 6.3 dB, respectively. Small minimum detectable differences for the Bilsom and EAR protectors indicate that a given protector is more likely to be uniformly fit across subjects.

TABLE VI. Minimum detectable differences for each device, fit, and frequency in dB. Note: $1-\alpha=0.84$, $1-\beta=0.8$, $n_s=20$, $n_t=2$. The between-subject standard deviations are taken from Table IV.

		Frequency (Hz)						
Device	Fit	125	250	500	1000	2000	4000	8000
Bilsom	IUF	2.0	1.6	1.6	2.0	1.7	1.9	2.1
UF-1 earmuff	SF	1.9	1.6	1.6	1.9	2.1	2.0	2.6
E·A·R	EF	3.9	4.1	3.7	3.1	1.8	1.5	2.4
Classic earplug	IUF	3.8	3.9	3.9	3.4	2.3	2.4	2.8
	SF	4.0	3.7	4.2	3.5	2.5	3.0	3.8
Willson	IUF	5.5	5.4	5.5	4.8	4.5	5.0	6.4
EP100 earplug	SF	6.2	6.1	6.7	6.0	5.5	6.0	7.4
PlasMed	EF	4.5	4.5	4.4	4.1	3.4	3.4	6.2
V-51R earplug	IUF	4.9	4.7	4.7	4.4	4.3	3.7	6.0
	SF	5.5	5.3	5.3	5.4	5.3	4.2	6.3

2. Number of subjects necessary for a desired resolution

The equation for calculating precision based on the variance for the sample size tested is

$$N_{\text{subjects}}=n_s\left(\frac{D}{R}\right)^2, \tag{8}$$

where N_{subjects} is the estimated sample size, n_s is the sample size for the tested population, D is the minimum detectable difference determined from the tested population for a given power and confidence level, and R is the target resolution in decibels. Minimum detectable difference is equivalent to the desired resolution except that the desired resolution is chosen rather than determined from the tested population. This formula has been applied to the minimum detectable differences derived from the four- and two-lab studies for a resolution of 6 dB with an $\alpha=0.16$ and $\beta=0.20$ for each protector and fitting condition. The desired resolution is the figure of merit of the ability to distinguish between two distributions of attenuations at any of the test frequencies. If, for instance, two sets of attenuations had been measured from earmuffs from separate production runs, the resolution and the minimal detectable difference could be used to determine how many subjects need to be tested to identify any manufacturing differences in the devices.

Figure 8 displays the estimated number of subjects, N_{subjects}, calculated from the minimum detectable difference in Eq. (6) for the three fitting methods using $\sigma_{\text{between-subject}}$ with $n_s=20$ subjects, $n_t=2$ trials, and $R=6$ dB. These values were selected according to the sample sizes, repetitions, and desired resolution used in the ANSI S12.6-1997 (ANSI, 1997) standard. While results for a given protector will not yield the same results presented here, these estimates represent conservative estimates from a variety of protectors. The preceding calculations should be performed for a particular data set to determine whether the minimum number of subjects has been achieved to reach the 6-dB minimum detectable difference.

The Bilsom UF-1 earmuff exhibited the smallest estimated sample sizes—4 subjects for the SF method at 8000 Hz. Likewise, the E·A·R plug exhibited a somewhat larger sample size than the earmuff (10 subjects for the SF method at 500 Hz), but considerably smaller than the two premolded earplugs. The EP100 earplug required 31 subjects for a resolution of 6 dB (8000 Hz), while the V-51R earplug needed 22 subjects for a 6-dB (the distance between distributions) resolution (also driven by 8000 Hz). The number of subjects must be rounded up to the nearest integer (e.g., 3.4 would round up to 4 subjects).

IV. DISCUSSION

A. REAT histograms

The REAT data presented in Figs. 1–3 illustrate the effect of fitting method upon the quality of fit. For an earmuff such as the Bilsom UF-1, the fit can be affected by disruption of the seal of the cushion such as by the ear pieces of safety glasses. Discounting damage to the cushions and improper placement of the muff over the pinna, earmuffs are easily fit on a subject's head. The agreement between IUF and SF REAT histograms demonstrate that additional experimenter involvement has little effect on the attenuation results for

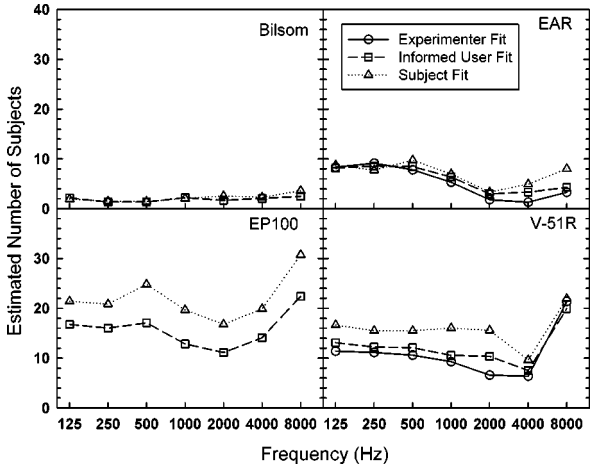

FIG. 8. Sample sizes necessary to achieve a 6-dB resolution in minimum detectable difference in attenuation determined by a power calculation based upon the repeatability and reproducibility analysis. The power calculations were based on $\alpha=0.16$ and $\beta=0.20$ which is equivalent to an NRR calculated with one standard deviation.

earmuffs. For an earplug, the data tend to suggest that the premolded devices were not well-fit while the E·A·R earplug was able to achieve a reasonable seal.

For the Bilsom UF-1 earmuff, the mixed-normal model yielded the best fit to the SF REAT distributions at 1000 and 8000 Hz. The distributions did not exhibit the low attenuation mode seen in the premolded earplugs and should not be considered to be bimodal. The mixed-normal model has a range of potential shapes besides the bimodal distribution. Murphy *et al.* (2002) examined the SF REAT data from the NIOSH laboratory and concluded that the mixed-normal model in all cases yielded an accurate fit of the distributions. The mixed-normal model can provide an accurate fit, but not significantly better than the normal or gumbel models.

The data collected for the E·A·R plug demonstrate an increase in the mean REAT as the experimenter involvement increases. The EF data exhibit a notch at 2000 Hz, which may indicate the maximum attenuation of the earplug has been reached (Berger, 1983). The REAT distributions for IUF are more evenly spread across a range of attenuation than the SF distributions below 1000 Hz. This difference may result from the interlaboratory differences observed under the IUF method. One of the laboratories achieved greater REATs for the lower frequencies than other laboratories. The broad range of REATs could reflect the effect of insertion depth and canal size on the ability of the E·A·R plug to achieve a tight seal. Presumably, a deeper insertion of the E·A·R plug yielded the greater REAT for the EF method compared to the IUF or SF methods. In the EF data, none of the REATs at any frequency were less than 7.5 dB.

For the premolded V-51R earplug, the influence of experimenter intervention was clearly evident. The mean values for the EF and IUF REAT histograms were greater than the SF mean REATs at all frequencies. As the stimulus frequency increased above 1000 Hz the differences in mean REATs between the fitting methods decreased. The V-51R earplugs yielded less average attenuation than the foam earplug at all frequencies. Depending upon the quality of fit, V-51R earplugs gave more attenuation than the Bilsom muff for some individuals in the lower frequencies.

Similarly, the premolded EP100 earplug exhibited a low attenuation mode for several frequencies in both the IUF and SF REAT distributions. For 500 Hz, the distribution was not significantly different from normal ($p = 0.052$). Otherwise, the maximum likelihood tests would have identified the distribution as mixed-normal. For the REAT distributions at 4000 and 8000 Hz, a small bulge can be seen in the histogram and is correlated with the larger bulges seen in the lower frequency distributions.

For premolded earplugs, the seal of the ear canal by the plug flanges is the critical factor influencing the bimodal character of the REAT distribution. In Royster *et al.* (1996), the earmuff REAT data were reported for the case where safety glasses were worn by the subjects. The lack of seal continuity around the stems of the glasses reduced the REAT performance. Similarly, for earplugs with an orifice, the insertion loss is about 0 dB for frequencies below 500 Hz (Hamery *et al.*, 1997). As the experimenter involvement increased, the mean REATs increased and more of the distributions were unimodal. A second factor might be the quality of the manufacturer's instructions. The mean IUF REATs were greater when subjects were coached rather than relying only on manufacturer instructions.

For foam earplugs, the attenuation characteristics of the foam and the bone-conduction limits are relatively consistent across subjects (Berger, 1983; Berger *et al.*, 2003). Consequently, varying insertion depths across subjects and the ability to seal the ear canal with a minimum of material probably contribute to the unimodal REAT distributions.

Several root causes of the problem of poor and improper fits could be investigated. The subject may have small ear canals, possess a low tolerance of discomfort, or perceive that deep insertion will damage the tympanic membrane. One may be concerned for the eardrum for a very deep insertion, but the depths typically observed for SF tests were nowhere near that deep. Instructions may be inadequate. The protector's design may not lend itself to a tight seal against the head or ear canal walls. The verification of such speculations would require further analysis of ear canal data and materials used in manufacturing the devices. The clear conclusion is that the type and design of the HPD affect the fit along with the amount of experimenter involvement.

B. Repeatability and reproducibility

The standard deviations for within-subject repeatability and between-subject reproducibility exhibited trends with increasing experimenter involvement and with the spread of data observed in the REAT histograms. For repeatability, the Bilsom muff had the smallest $\sigma_{\text{within-subject}}$ and was followed by the E·A·R plug, the V-51R and EP100 earplugs, respectively. The standard deviations for all devices at 8000 Hz were greater than the value at 4000 Hz. Some slight trend can be seen for increasing standard deviations with decreasing experimenter involvement. The spread of these values was seldom more than 0.2 dB and not more than 0.4 dB for any frequency. The lack of change between fitting methods can be understood as a metric of how subjects fit the device and perform the psychoacoustic task. The smaller standard deviations for within-subject repeatability for the Bilsom device reflect the consistency of fit for repeated tests. For the three earplugs, the occluded threshold measurements may have increased overall variance within subjects. In Franks *et al.* (2003), the variance of repeated occluded thresholds was greater than repeated measures of unoccluded thresholds. Unfortunately, the data examined in this paper were the REAT measures for each subject's trial, not the unoccluded and occluded thresholds.

The standard deviations for the between-subject reproducibility include the additional information about the consistency of fit across subjects. The Bilsom earmuff had the smallest standard deviations which again reflects the ease of fitting the device. The E·A·R plug had smaller deviations than the other plugs. Since fitting the earplugs requires some training, the standard deviations for all the earplugs reflect a trend of decreasing standard deviations with increasing experimenter intervention. The differences in the range are less than 0.5 dB for most devices and frequencies.

The standard deviations for between-laboratory reproducibility yielded results contrary to the within-subject repeatability and between-subject reproducibility. As stated before, the increase in the E·A·R plug IUF between-laboratory standard deviations for frequencies below 1000 Hz has been attributed to greater attenuation achieved by one laboratory. These effects are evident in the increased range of the REATs presented by Royster *et al.* (1996) and are reflected in the values for $\sigma_{\text{between-subject}}$ for the IUF condition in Fig. 5.

One problem with the large between-laboratory standard deviations is the inability to make comparisons of the REATs. Some of the devices and methods failed to produce any between-laboratory standard deviations, $\sigma_{\text{between-laboratory}}$, less than 2 dB. Using three standard deviations as a criterion, differences in REAT for premolded earplugs measured with the SF method only begin to be meaningful when they are greater than 6 dB. The two-lab EF data tend to have standard deviations less than 2 dB and sometimes nearer to one. When the between-laboratory variations are as small as 2 dB, it is appropriate to make comparisons of REATs measured in the different laboratories.

The increased variance at 8000 Hz needs to be addressed since it influences the number of test subjects for the two premolded earplugs and the earmuff. The effect of the acoustic leak at low frequencies correlates with poor attenuation at 8000 Hz. In Hamery *et al.* (1997), the insertion losses of a variety of earplugs were measured for many configurations of the size and position of a nonlinear orifice. Several plots exhibited a maximum insertion loss below 8000 Hz and a local minimum at the 8000-Hz frequency. If the poorly fit protector behaves similarly to the earplug with an orifice, then the correlation of poor attenuation at 8000 Hz might be explained by a resonance of the occluded volume. The protectors tend to have more attenuation at 8000 Hz. These two factors work against having small variances. The protector that can reduce the potential for leakage (earmuffs and foam plugs) will likely have smaller variance and consequently require fewer subjects to achieve a given minimum detectable difference.

C. Power calculations

Lastly, the power calculations need to be examined. The power calculations demonstrate that consistent HPD performance across subjects require fewer subjects to achieve a desired level of resolution (see Fig. 8). For the Bilsom muff, the number of subjects necessary to achieve a 6-dB resolution with the SF method is less than 4. The E·A·R plug requires fewer than 10 subjects. The other two devices, EP100 and V-51R, respectively, require 31 and 22 subjects to achieve the 6-dB resolution. The worst-case frequency is the appropriate choice upon which to estimate the number of subjects. If the confidence level and power were made more stringent (e.g., $1-\alpha=0.9$ and $\beta=0.85$), the minimum detectable difference will increase. While the minimum detectable difference is dependent upon the number of subjects tested, it is also a function of σ_{trial} and σ_{subject} for the device. With power calculation outcomes, it is possible to determine whether a rating for a protector has any relevance when a set number of subjects is used.

The current noise reduction rating provides no estimate other than the standard deviations listed on the secondary label of the error that might be associated with the rating. The definition of the power calculation and development of a method to assess the error associated with the REAT measurement and subsequent NRR rating would achieve parity among ratings. The earmuff data have smaller standard deviations across subjects and laboratories than the earplugs. This difference implies that a worker can expect to achieve attenuations near to what is measured in the laboratory when wearing earmuffs. An error estimate for the NRR of the UF-1 earmuff should be smaller than the error estimate of the earplugs. Such estimates are not currently a part of the EPA rating regulation or the ANSI S12.6-1997 (ANSI, 1997) standard.

D. Subject-fit method and the ANSI S12.6-1997 (ANSI, 1997 standard

This research as described in the Introduction and in Royster *et al.* (1996) was designed by the S12 Working Group 11 to develop a laboratory testing standard that was predictive of real-world protection received by motivated workers. Two protector fitting protocols were investigated. The data derived from the subject-fit method compared to the informed-user-fit method did not yield the expected results. The standard deviations for between-subject reproducibility and between-laboratory reproducibility were expected to increase with decreased experimenter involvement. Instead, the standard deviations for between-laboratory reproducibility fluctuated considerably for the IUF method. Therefore, the more consistent and realistic method with respect to real-world data (Berger *et al.*, 1998) proved to be the subject-fit method.

The authors of the ANSI S3.19-1974 (ANSI, 1974) standard selected a sample size of ten subjects and three test repetitions for each subject. Each repetition consisted of a pairing of occluded and unoccluded thresholds. As demonstrated from the present power calculations for determining the estimated subject sample size for a 6-dB resolution, the working group's selection of ten subjects and two trials was valid for the Bilsom UF-1 earmuff where as few as four subjects with two trials would have been sufficient. As well, ten subjects with two trials would be sufficient for the E·A·R plug for 6-dB resolution. However, testing premolded products requires more subjects and, as the V-51R and EP100 are examples of hearing protectors that are fitted ineffectively, the working group selected 20 subjects with two trials. As Eq. (8) and Fig. 7 show, increasing the panel size increases the statistical power and decreases the minimum detectable difference. As ANSI S12.6-1997 (ANSI, 1997) provides for adequate sample sizes for at least 6-dB resolution for almost any type of hearing protector, it is also possible to calculate the precision of REATs for any set of 10 subjects for earmuffs or 20 subjects for earplugs.

The ANSI S3.19-1974 (ANSI, 1974) statistical treatment of the three repetitions of REAT measurements from ten subjects as 30 statistically independent, uncorrelated data points is incorrect. The correct approach is to average the three repetitions from each subject and then use the individual

averages for calculating means and standard deviations. This statistical shortcoming has been remedied in the S12.6-1997 standard where the averages of repeated measures are used for determining means and standard deviations.

Based upon that rationale, the power calculations for the subject-fit method using 6-dB resolution yield reasonable subject sample sizes for the uniformly performing HPDs (Bilsom muff and E·A·R plug). The power calculations provide a different perspective with which to evaluate the performance of a hearing protector. Those devices which are easily fit and produce highly consistent results across frequencies will yield smaller variances and, consequently, require fewer subjects to achieve a given resolution.

V. CONCLUSIONS

The analysis of the between-subject reproducibility and between-laboratory reproducibility of the devices tested in the four- and two-lab studies has been presented. Some of the underlying causes for increased variance for various protectors have been discussed. The variance of repeatability increases with decreasing experimenter intervention. The variance for reproducibility is most consistent when the experimenter factor is removed. In a separate analysis the Working Group also identified the subject-fit method as the best estimator of field performance (Berger *et al*. 1998), and for that reason the Method-B procedure was initially selected for inclusion in the 1997 standard. The results of the analyses in this report verify that, for purposes of data reproducibility, Method B is an apt choice as well.

ACKNOWLEDGMENTS

The authors wish to acknowledge Dr. Charles Nixon's contributions for his oversight in data collection at Wright Patterson Air Force Base. The authors wish to thank Dr. Doug Ohlin for his assistance with developing the research protocol and support from the U.S Army. Portions of this work were supported by the U.S. EPA Interagency Agreement 75090527.

ANSI (**1999**). S3.6-1999, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

ANSI (**1974**). S3.19-1974, *American National Standard Method for the Measurement of Real-ear Protection of Hearing Protectors and Physical Attenuation of Earmuffs* (American National Standards Institute, New York).

ANSI (**1997**). S12.6-1997, *American National Standard Method for Measuring Real-ear Attenuation of Hearing Protectors* (American National Standards Institute, New York).

Berger, E. H. (**1983**). ''Laboratory attenuation of earmuffs and earplugs both singly and in combination,'' Am. Ind. Hyg. Assoc. J. **44**, 321–329.

Berger, E. H., Franks, J. R., Behar, A., Casali, J. G., Dixon-Ernst, C., Kieper, R. W., Merry, C. J., Mozo, B. T., Nixon, C. W., Ohlin, D., Royster, J. D., and Royster, L. H. (**1998**). ''Development of a new standard laboratory protocol for estimating the field attenuation of hearing protection devices. Part III. The validity of using subject-fit data,'' J. Acoust. Soc. Am., **103**, 665–672.

Berger, E. H., Kieper, R. W., and Gauger, D. (**2003**). ''Hearing protection: Surpassing the limits to attenuation imposed by the bone-conduction pathways,'' J. Acoust. Soc. Am. **114**, 1955–1967.

Casali, J. G., and Park, M. Y. (**1991**). ''Laboratory versus field attenuation of selected hearing protectors,'' Sound Vib. **10**, 28–38.

Franks, J. R., Murphy, W. J., Johnson, J. L., and Harris, D. A. (**2000**). ''Four earplugs in search of a rating system,'' Ear Hear. **21**, 218–226.

Franks, J. R., Murphy, W. J., Harris, D. A., Johnson, J. L., and Shaw, P. B. (**2003**). ''Alternative field methods for measuring hearing protector performance,'' Am. Ind. Hyg. Assoc. J. **64**(4), 501–509.

Hamery, P., Dancer, A., and Evrard, G. (**1997**). *Étude et réalisation de bouchons d'oreilles perforés non linéaires*, ISL R128/97 (Insitut Franco-Allemand de Recherces de Saint-Louis, Saint–Louis).

ISO 5725-2 (**1994**). *Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method* (International Organization for Standardization, Geneva).

Murphy, W. J., and Franks, J. R. (**1998**). ''Analysis of repeatability and reproducibility of hearing protector real-ear attenuation at threshold measured with three fitting methods,'' Nat. Hear. Cons. Assoc., 19–21 February, Albuquerque, NM.

Murphy, W. J., and Franks, J. R. (**2001**). ''A reevaluation of the Noise Reduction Rating,'' Meeting of the Commissioned Officers Association of the US Public Health Service, 29 May, Washington, DC.

Murphy, W. J., Franks, J. R., and Krieg, E. F. (**2002**). ''Hearing protector attenuation: Models of attenuation distributions,'' J. Acoust. Soc. Am. **111**, 2109–2116.

Netter, J., Wasserman, W., and Kutner, M. H. (**1990**). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, 3rd ed. (Irwin, Boston), pp. 970–1001.

NIOSH (**1995**). *The NIOSH Compendium of Hearing Protection Devices*, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Publication No. 95–105.

Rosner, B. (**1990**). *Fundamentals of Biostatistics*, 3rd ed. (Duxbury, Belmont).

Royster, J. D., Berger, E. H., Merry, C. J., Nixon, C. W., Franks, J. R., Behar, A., Casali, J. G., Dixon-Ernst, C., Kieper, R. W., Mozo, B. T., Ohlin, D., and Royster, L. H. (**1996**). ''Development of a new standard laboratory protocol for estimating the field attenuation of hearing protection devices. Part I. Research of Working Group 11, Accredited Standards Committee S12, Noise,'' J. Acoust. Soc. Am. **99**, 1506–1526.

SAS/STAT Software (**1998**). Version 6.12 SAS Institute Inc.

S-Plus 6.1 for Windows (**2002**). Insightful Corporation, Seattle, WA.

U.S. Environmental Protection Agency (**1978**). CFR Title 40, subchapter G, 211, subpart B—Hearing Protective Devices, U.S. EPA.