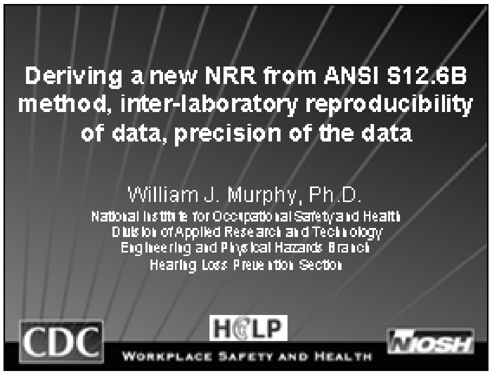
Deriving a new NRR from ANSI S12.6B method, interlaboratory reproducibility of data, precision of the data

William J. Murphy, Ph.D.

National Institute for Occupational Safety and Health
Division of Applied Research and Technology
Engineering and Physical Hazard Branch
Hearing Loss Prevention Section

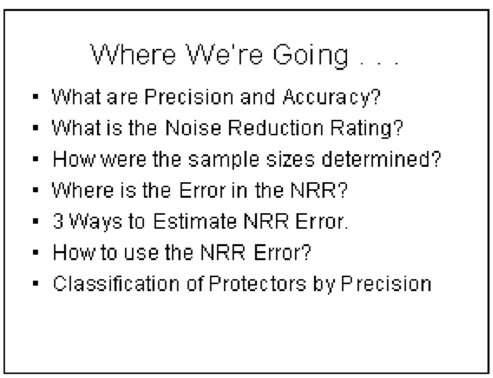
Oral Text of a Presentation at the
Workshop on Hearing Protection Devices
United States Environmental Protection Agency
Washington, D.C.
March 27, 2003



Good Morning. I am privileged to be speaking this morning about the ANSI S12.6 Method B interlaboratory reproducibility of data and the precision of that data. This talk will review the analysis of interlaboratory reproducibility and will extend those concepts to other methods for estimating the error in the Noise Reduction Rating. Three methods for estimating the NRR error will be developed and motivated. Some of this presentation will be technical, however, with proper motivation and explanation, the essence of the concepts should become apparent.

Before continuing, the title of the presentation deserves some translation. In fact, this presentation seeks to answer the question, "When have enough subjects been tested?"

(Reference ANSI S12.6-1997; Royster et al., 1996; Murphy et al., 2003)



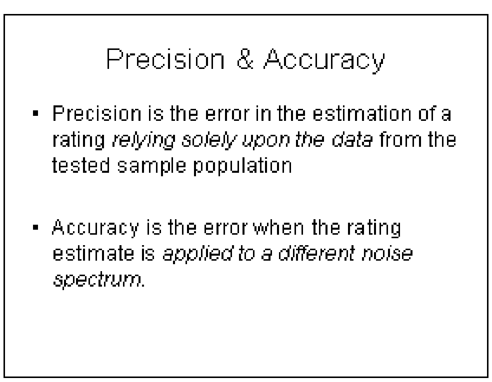
Several topics will be covered. First the definitions of Precision and accuracy will be given and their relationship to hearing protection will be motivated.

Next, A brief review of hearing protector ratings and testing procedures will be given.

Then the development of the subject sample size requirements in ANSI S12.6 method B will be explained.

Following that, three methods for estimating the error in the noise reduction rating will be developed and applied to the interlaboratory data.

Finally, a classification scheme for hearing protection ratings based upon precision will be presented.

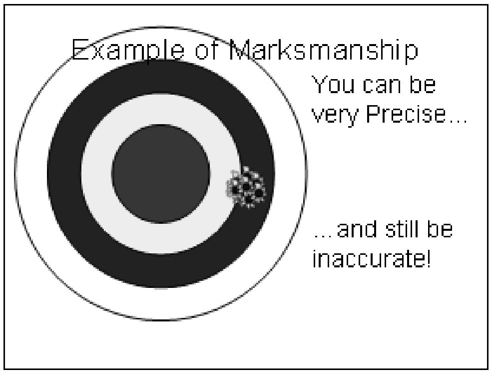


First, the meanings of precision and accuracy as they apply to hearing protector ratings need to be set forth.

The Precision of a hearing protector rating is the error in the estimation of a rating that is derived solely from the tested sample population.

The Accuracy of a hearing protector rating is the error in applying the rating estimate to a different noise spectrum.

(Reference Berger, 2003)



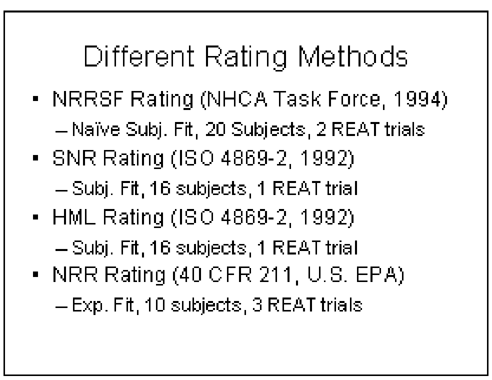
These definitions can be motivated with an example from marksmanship. Imagine clamping a rifle to a bench rest and shooting a set of 10 shots. The spread of those ten shots about the center of the group is a measure of the precision of that rifle. A heavier barrel with better rifling and stiffening ribs will make the barrel less prone to vibration.

The grouping will be come smaller and the rifle more precise.

Unfortunately, if you miss the target, Precision without accuracy is useless.

Similarly, a protector can have very tight attenuation distributions, indicative of a highly controlled testing protocol or a well-designed protector. The rank-ordered comparison of real-world attenuation of hearing protectors, subject-fit and laboratory REAT data demonstrate that laboratory data tend to be very precise but way off target. If the test procedure and rating method are not accurate, then ultimately the rating is meaningless. This presentation will focus on how to estimate the precision. Other speakers will be addressing the issue of accuracy for different noise spectra.

(Suter, 2003; Berger, 2003)



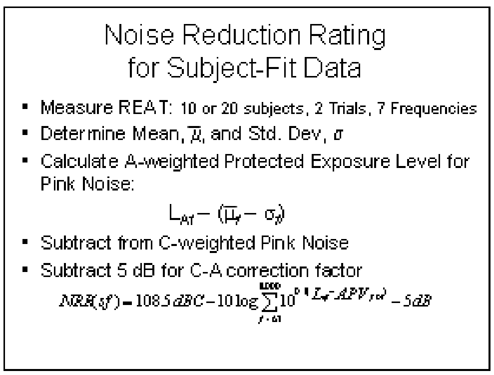
A brief review of selected hearing protector rating methods is necessary.

The Noise Reduction Rating Subject Fit was developed by the NHCA task force on hearing protector effectiveness. The task force developed a rating based upon testing 20 subjects twice to estimate the real ear attenuation at threshold for a protector. The subjects were to be naïve with respect to protector use and testing. This method incorporated a mean minus one standard deviation to estimate the protection of 84% of the users that would wear a device.

The next two methods, Single Number Rating and the High-Middle-Low ratings are the European methods approved by the ISO in 1994. The SNR method provides one number that is subtracted from a C-weighted noise to estimate the A-weighted exposure level of a person wearing a protector. Similarly, the HML method requires the user to know the difference between the C-weighted and A-weighted sound pressure levels before applying the rating. HML is more accurate than the SNR method when applied to a variety of noise spectra for the purpose of estimating the protected exposure level. Both methods are calculated from 16 subjects performing one REAT trial and allowing them to have some level of experience with the use of protectors.

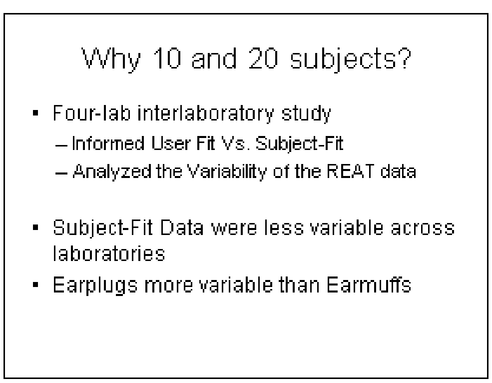
The NRR has been the subject of criticism almost from its inception. The NRR uses 10 experienced subjects, has the experimenter fit the protector and measures the REAT for three trials. The NRR must be derated before applying it to the problem of estimating a worker's exposure level. The experimenter-fit results are representative of the best possible performance of a hearing protector.

(EPA, 1978; ISO 4869-1, 1990; ISO 4869-2, 1994; Royster, 1995; OSHA, 1999)



For those in the audience that are unfamiliar with hearing protector ratings, we will review the Noise Reduction Rating Subject Fit. The manufacturer sends a product to a testing laboratory. The lab must recruit a panel 10 or 20 subjects who have no experience with protector use and testing. The unoccluded and occluded hearing thresholds are measured for each subject after they have been qualified for testing. Each subject is measured twice at seven frequencies. After the panel is completely tested, the lab must calculate the means and standard deviations at each frequency of the real ear attenuation at threshold. From these values, the overall A-weighted protected exposure level is determined and subtracted from the C-weighted Pink noise. For the NRRSF, a correction factor of 5 dB is subtracted. When it is all said and done, this formula describes the process. Most protector rating schemes utilize a similar formula. This is an important point since this talk examines the error in using this formula. The C- and A- weighted reference spectrum, 108.5 dBC and L_{Af} , and the C-A correction factor are the components that are varied when assessing the accuracy of the rating metric.

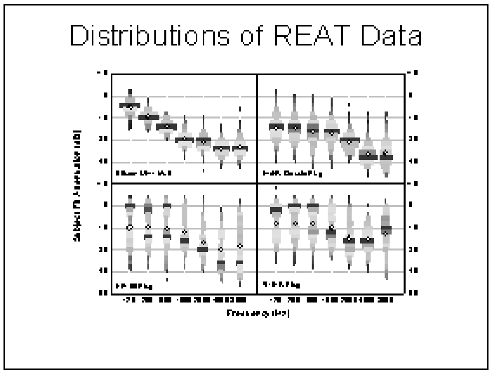
(Royster, 1995; Franks et al., 2000)



How did ANSI S12.6-1997 arrive at the threshold of testing 10 subjects for earmuffs and 20 subjects for all other devices? In the early 1990s, an interlaboratory study was conducted between four labs, NIOSH, EARCal, WPAFB, and USAARL. The study tested four hearing protectors with two protocols: Informed User Fit, and Subject Fit.

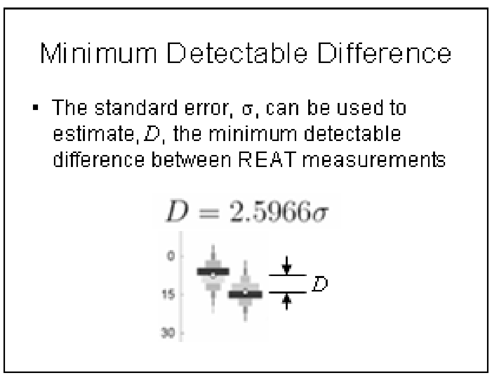
Statistical analysis demonstrated that Subject-Fit data were less variable across laboratories than Informed User Fit data. The analysis also showed that Earplugs were more variable than earmuffs.

(Royster et al. 1996; ANSI S12.6-1997; Murphy et al. 2003)



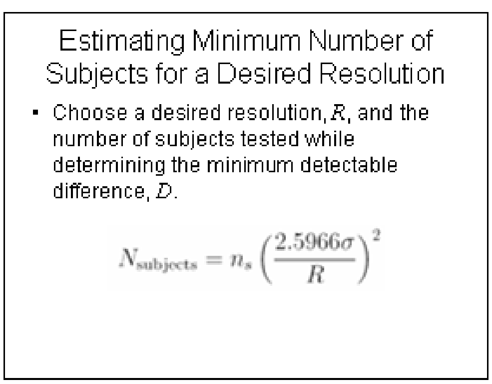
This figure shows the distributions of REAT data for the subject fit trials from the interlaboratory study. Two points are evident. The REAT distributions for the Bilsom and EAR Classic are unimodal and for the most part symmetric about the mean value, the diamond symbol. For the V-51R and EP100 premolded earplugs, the distributions are bimodal in the low frequencies and widely spread at the higher frequencies.

(Murphy et al., 2002; Murphy et al., 2003)



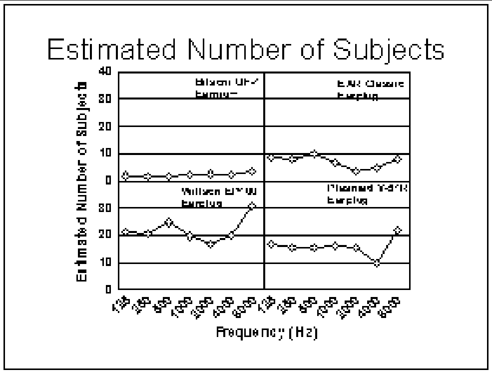
From the statistical analysis of the data across laboratories, subjects and trials, an error term, sigma, was estimated at each frequency for the four protectors. From these error terms, assumptions of statistical certainty were made for the purpose of determining the minimum detectable difference between two distributions of data. The minimum detectable difference is the distance between the centers of the distributions.

(Murphy et al., 2003)



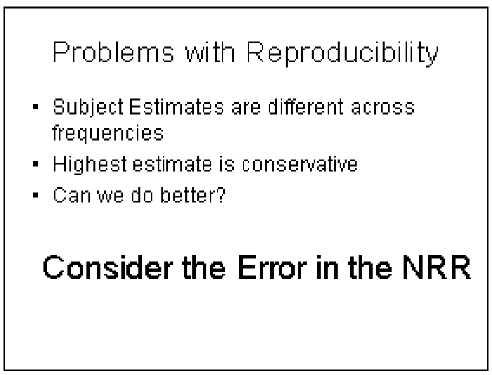
Once the minimum detectable difference has been determined and the desired resolution chosen, the minimum number of subjects for testing can be calculated. Appendix C of ANSI S12.6-1997 standard used a desired resolution of 6 dB to estimate the sample sizes for testing different types of hearing protectors. The n_s in this formula is the number of subjects actually tested and the N_subjects is the estimated number of subjects to achieve a desired resolution of R.

(ANSI S12.6-1997; Murphy et al., 2003)



In this figure the interlaboratory subject-fit data have been analyzed and the estimated numbers of subjects have been plotted for each protector and frequency for a desired resolution of 6 dB, number of tested subjects is equal to 20 and number of trials is equal to 2. For the UF-1, the estimated number of subjects was less than 4. For the Classic earplug, the estimated number of subjects was less than 10. For the V-51R and EP100, the estimated numbers of subjects were about 23 and 32, respectively. Some problems with this analysis exist.

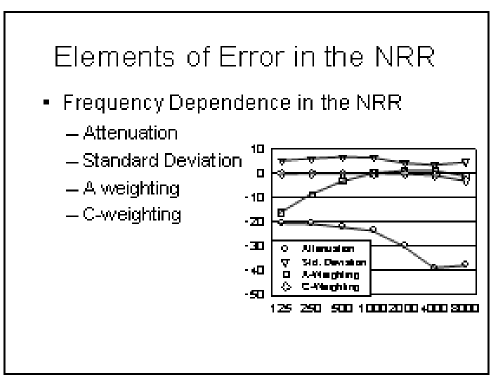
(Murphy et al., 2003)



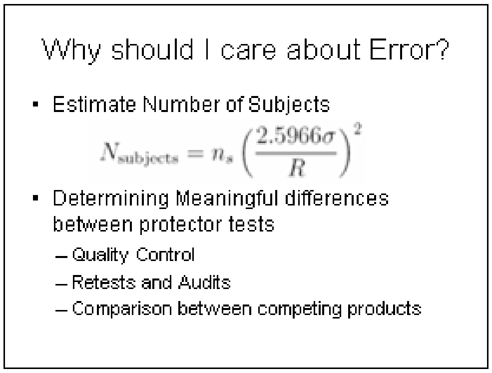
Using the reproducibility, the estimated numbers of subjects were plotted for each frequency in the previous slide. The analysis does not show how to combine the subject estimates across frequencies. Without that information, the most conservative estimate was the maximum number of subjects at any frequency. For a resolution of 6 dB, at least 32 subjects need to be tested for the EP100.

Can a better estimate be developed?

Yes, if one considers the error in the NRR.



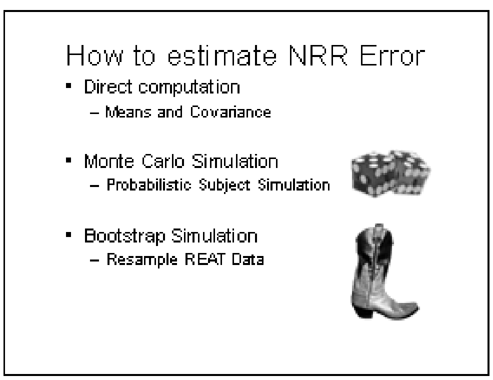
The NRR calculation involves summing energy and attenuations across frequencies and distills down to four components that are frequency dependent. The protector's attenuation typically increases with frequency. The standard deviations tend to be constant within a few decibels. The A-weighting curve deemphasizes the lower frequency bands of the reference spectrum and the C-weighting of the reference spectrum is relatively constant except at the higher frequencies. These terms will jointly influence the error contribution.



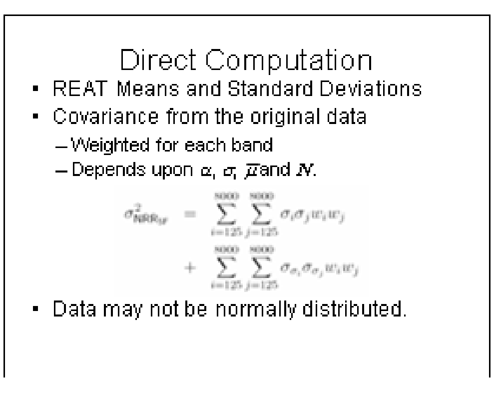
So why should one care about the error?
First as has been shown previously, the number of subjects necessary to achieve a desired resolution can be estimated. Using the prior formula, if one knows the error, sigma, then the desired resolution need only be chosen to know whether sufficient subjects have been tested.

More importantly, the error can be used to determine meaningful differences between protector tests. The applications might include quality control within a manufacturing facility, retesting the product for labeling and audit purposes, and making comparisons between competing products on the market. The current mode of comparison is usually performed on the basis of the NRR magnitude. If product A has an NRR of 21, then it must do a better job than product B which has an NRR of 20. No thought has been given to characterizing the protector based upon the precision of the rating. An intelligent consumer might look at the standard deviations provided on the secondary label and be able to make some sense of them. And if the user is an acoustician, they will know how to take that rating and perform the octave band calculation to get their exposure level and they will consider the comfort factor for an extended period of wearing the protection.

Sadly this is rarely the case.



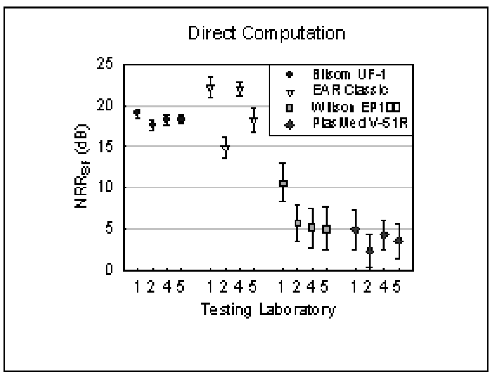
Recently, NIOSH has developed three methods to estimate the error in the NRR. The first is a direct computation using the means and covariance of the REAT data. The second is a Monte Carlo method that simulates data based upon the means and covariance of the REAT data. And finally a bootstrap method in which one samples the original REAT data to form new data sets that are used to estimate the NRR multiple times. Each of the methods has good and bad points that are a function of the assumptions used in their calculation.



The REAT data can be characterized by the mean attenuation at each frequency and the covariance matrix for the entire set of measurements for the tested subject sample. The covariance is simply the variation of the attenuation at one frequency with the attenuation measured at another frequency. When a subject achieves an excellent fit, the attenuations will generally be greater across frequencies than a subject who achieves a poor fit. Thus, the covariance matrix can be used to better assess the error. From a derivation of the variance of the NRRSF, we find an equation of the following form. What is interesting about this result is that the individual frequencies are weighted according to their contribution to the overall protected sound pressure level.

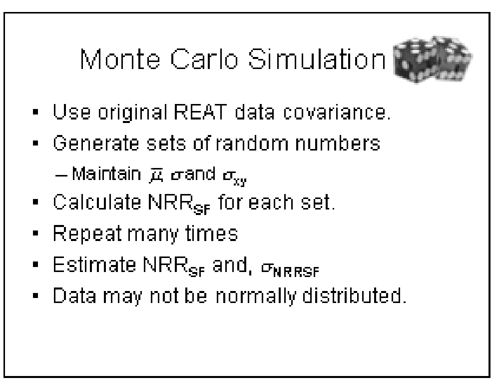
The primary shortcomings of this derivation are that it assumed the REAT data are normally distributed and that it must be derived for each particular rating method.

(Bevington, 1969)



The subject fit data from the interlaboratory study were analyzed using the means and covariance to estimate the error bars. The error bars about the NRRSF for the UF-1 earmuff are small, about 0.5 dB. The errors about the EAR Classic ratings are about 0.9 to 1.5 dB. The error bars for the EP100 range from 2.1 to 2.6 dB and the errors for the V-51R are 1.8 to 2.4 dB.

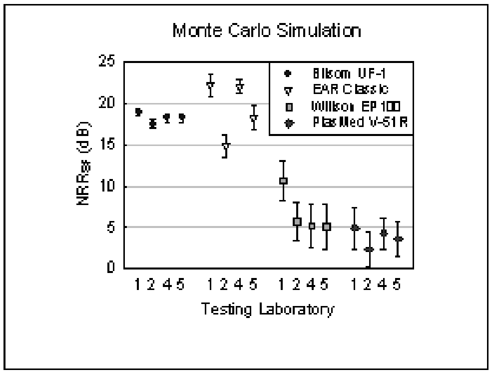
Further analysis of the data was performed to determine whether or not the differences in the NRRSF measurements in the different labs were statistically significant. Only for the EAR Classic were these data different from one another. Lab 2 was not significantly different from Labs 1 and 4, but not Lab 5. The remainder of the protectors exhibited no significant difference across labs. Please note that even though the EP100 exhibited a difference of 6 decibels between Labs 1 and 5, the difference is not significant.



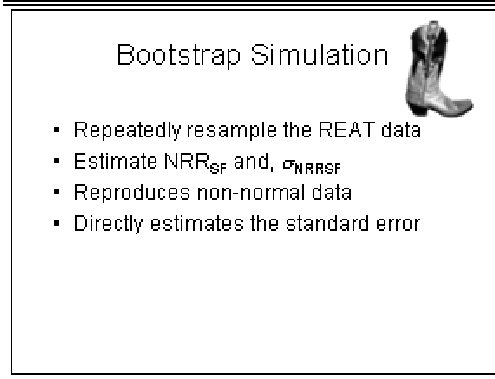
For the Monte Carlo simulation, a set of random numbers is generated that has the same mean and standard deviation as the original REAT data. The NRRSF is computed for that set of data and the result is stored. The process is repeated several thousand times to guarantee convergence of the mean and the standard deviation of the NRRSF.

The method makes an assumption that the subjects are randomly drawn from a normally distributed population. For some protectors, the REAT distributions were not normal but bimodal. Bimodality has a small, unpredictable affect on the NRRSF calculation. A better model of the distribution of the data is the topic for continued research.

(Press et al., 1986)

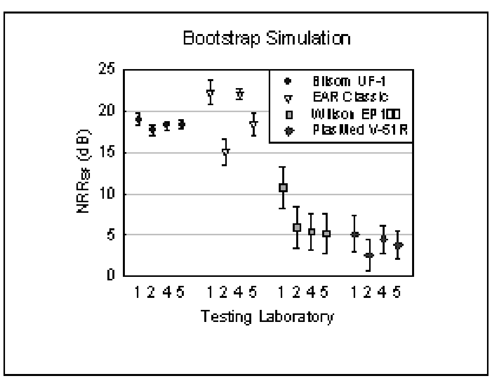


As we examine the errors for the Monte Carlo method, they are approximately 5% larger than the errors for the direct method. The NRRSF calculations are the same, and there are no discernable differences in the results.

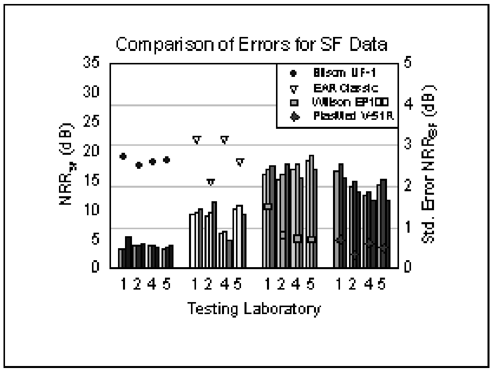


The Bootstrap simulation is a unique approach to both model the REAT data and estimate the error inherent in the hearing protector rating. One assumes that the subjects can be randomly sampled such that they have an equal probability of being selected for each throw of the dice. This sampling strategy is called Sampling with Replacement. The number of subjects drawn is the same as in the original sample. In the case of the Interlaboratory study, each lab tested 24 subjects, so each random sample will select 24 subjects.

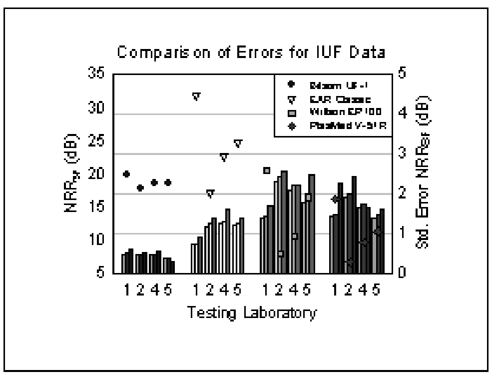
(Efron and Tibshirani, 1993)



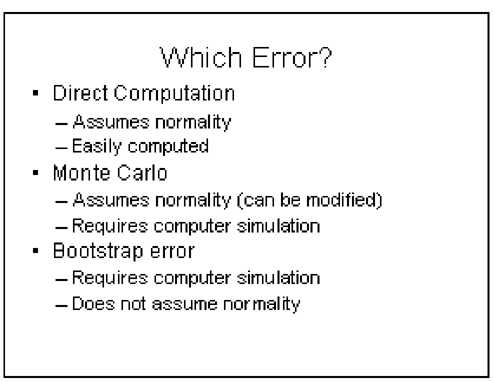
The Bootstrap errors are slightly greater than the direct and sometime the Monte Carlo method errors. The results incorporate the bimodal character of the data because the actual data are used in the calculation.



In this figure, the NRRSF calculations (left axis) have been combined with a bar chart for the errors shown on a different scale on the right hand axis. The errors for the Direct method are the lightly shaded bars; the errors for the Monte Carlo method are the medium shaded bars and the Bootstrap errors are the darkly shaded bars. One should recognize that the errors from each method are comparable. The UF-1 earmuff errors are less than 1 decibel. The Classic errors were less than 2 dB. The EP100 earplug errors were all above 2 dB and less than 3 dB. The V-51R errors were above 1.5 dB and less than 2.5 dB.



The same analysis was performed on the Informed User Fit data from the four-lab study. The errors in this case tend to show the bootstrap slightly greater than the other methods. For the earmuff, the Informed User Fit errors were comparable to the Subject Fit errors. For the earplugs, the errors overall were less than those for the subject-fit data.

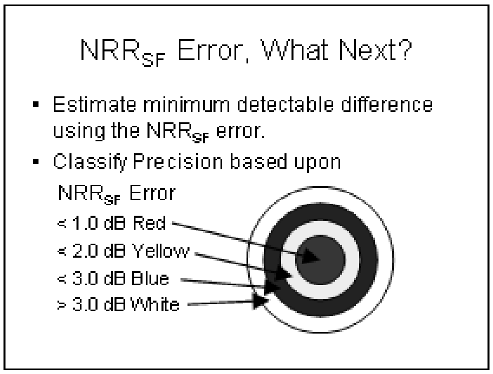


After looking at the errors, which method should be used? At this point, the results are comparable for the different methods. The direct method assumes normality of the data and may be incorrect for non-normal data. Its advantage is that it can easily be computed and can be programmed into a spreadsheet.

The Monte Carlo method also assumes normality in the data, but could be modified for non-normal distributions. It required computer simulation using a high level language.

The Bootstrap method does not assume any structure in the data because it uses the original data to generate its results. The bootstrap required a computer simulation using a high level language.

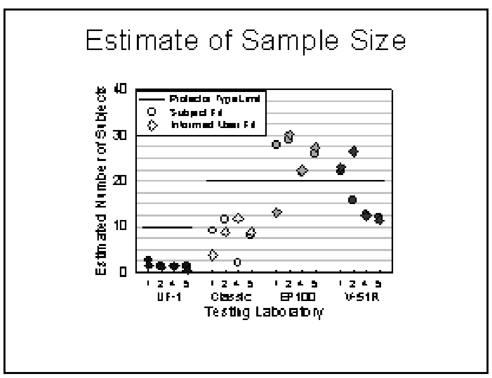
At this point, the bootstrap seems to be the best method for estimating the error. The other methods work, but may need further development to assure the results are always accurate.



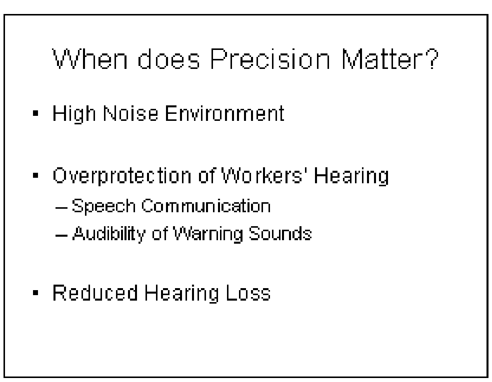
Now that the effects of the standard error on the hearing protector rating have been examined, how might the precision be used?

From the earlier formula, the minimum detectable difference can be determined and the number of subjects to test can be estimated.

The precision of the protector could be classified. The highest precision protectors with errors less that 1 dB could be classified as red. Those protectors with errors greater than 1 but less than 2 dB could be yellow. Errors greater than 2 and less than 3 dB would be blue and any device greater than 3 dB would receive a white classification. The class scheme could easily be Type 1 through Type 4.



This figure presents the estimates of the sample sizes for the interlaboratory study based upon the bootstrap errors and the minimum detectable difference of 6 decibels. Please remember that this is the distance between two distributions to be able to distinguish them. For the Bilsom UF-1 earmuff the number of subjects was less than three for both the Subject and Informed User Fit data. The EAR Classic required less than 12 subjects for all the labs. The EP100 exhibited the poorest results with Lab 2 requiring 30 subjects to achieve a 6 dB minimum detectable difference. Finally, the V-51R earplug required less than 27 subjects for Lab 2. Several of its measurements were less than the suggested 20 subjects.

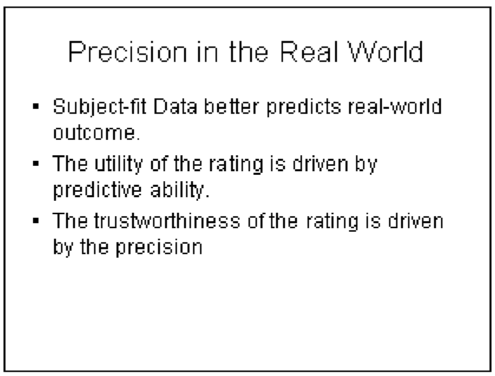


So, when does precision matter?

In high noise environments, the hearing protection must be matched to the worker's noise exposure. If the protection is inadequate, the worker will be at an increased risk of developing a hearing loss. Current practices utilize double protection which pairs an earmuff with an earplug. The muff typically will have higher precision than the earplug. If both devices were high precision, then the worker has greater assurance of adequate protection.

Two aspects of overprotection in a noisy environment must be considered: the ability to communicate and the audibility of warning sounds. If workers are unable to communicate due to overprotection, they are likely to remove or defeat the attenuation of the protector, which increases their noise exposure. Increased noise exposure means increased risk of hearing loss. Similarly, if workers cannot hear warning sounds such as backup alarms, they put their lives instead of their hearing at risk.

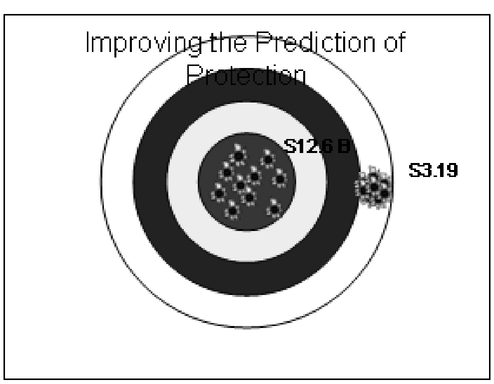
The bottom line for the employer is that they need to better characterize the noise exposure profiles of their workers to best match protection with exposure. If employers choose low precision protectors, then their workers are at greater risk of developing hearing loss.



Precision can applied to the difficult issues of hearing loss prevention. From this talk and others, the subject-fit data have proven to better predict real-world attenuation measurements than have the ANSI S3.19 Experimenter-Fit data.

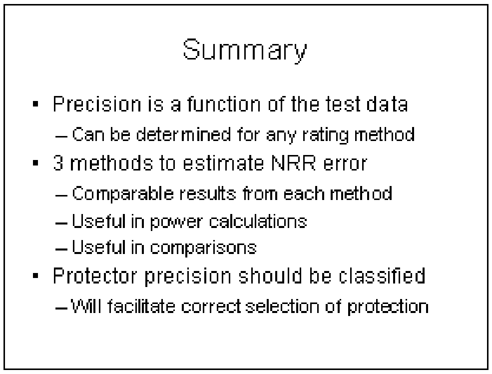
The utility of the rating is driven by is predictive ability. That OSHA requires and NIOSH recommends derating the current Noise Reduction Rating, should be evidence that Experimenter-Fit data do a poor job of predicting real-world performance. Moving to Subject-Fit data should improve the ability to predict the protected exposure noise levels for workers.

The precision of the data drives the trustworthiness of the rating. Some precision will be sacrificed when using subject-fit data, especially for earplugs and semi-aural devices. Testing a larger pool of subjects will improve the precision of the rating, in effect tightening the confidence limits for the rating and decreasing the minimum detectable difference.



If the target of a hearing protector rating is to predict how well-protected a worker might be, then consider this revision to the earlier example. Currently, ANSI S3.19 laboratory data are poor predictors of real-world performance. The data are very precise but way off target. If the United States shifts its regulations to using subject-fit data, then some precision is sacrificed for the sake of accuracy.

(Berger et al. 1998)



In summary, the precision is a function of the original REAT data measured for the sample pool of subjects. The precision is an inherent property of the data and can be determined for any method. The accuracy of a hearing protector rating method depends upon the noise spectrum where the protector will be used and its ability to describe real-world performance.

Statistical analysis has been developed to estimate the numbers of subjects necessary to achieve a level of statistical certainty. That analysis was limited by its inability to combine results across frequencies. The formulas would continue to be useful if we knew the error in the protector rating.

Three methods have been briefly presented to estimate the error in the rating: the Direct, Monte Carlo, and Bootstrap methods. Each method yielded comparable results, but currently the bootstrap has the most potential to be applied to any rating method. The error in the protector rating can be useful in power calculations to predict how many subjects need to be tested. The error will also permit meaningful comparisons between tests and devices.

Finally, some applications of the precision to the practical problem of hearing loss prevention have been presented. Precision is function of the actual REAT testing data rather than the color of its plastic or the type of foam from which it was manufactured.

Thank you again for the privilege of speaking to you this morning.

References

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. and Kieper R.W. (**2000**). "Ongoing experience with laboratory-based subject-fit REAT methods for measuring hearing protector attenuation" J. Acoust. Soc. Am. 108(5, Pt.2), p. 2619

Berger E.H. (**2003**). "Rating Metrics for Hearing Protectors," Presentation at U.S. Environmental Protection Agency Workshop on Hearing Protection Devices, Washington DC, March 27, 2003.

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.

Bevington, P.R. (1968). *Data Reduction and Error Analysis for the Physical Sciences,* pp. 56-65 (McGraw-Hill Inc., New York).

Efron, B. and Tibshirani R. (**1993**). *An Introduction to the Bootstrap,* (Chapman and Hall, New York).

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. (2003). "Comparison of the regulatory Noise Reduction Rating (NRR) and the required ANSI S3.19 test method with real-world outcomes and results from testing with the new ANSI S12.6B method," Presentation at U.S. Environmental Protection Agency Workshop on Hearing Protection Devices, Washington DC, March 27, 2003.

ISO 4869-1 (**1990**). "*Acoustics—Hearing Protectors—Part 1: Subjective method for the measurement of sound attenuation,*" (International Organization for Standardization, Geneva).

ISO 4869-2 (**1994)**. "*Acoustics—Hearing Protectors—Part 2: Estimation of Effective A-Weighted Sound Pressure Levels When Hearing Protectors are Worn,*" (International Organization for Standardization, Geneva).

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.

Murphy, W.J., Franks, J.R., Berger, E.H., Behar, A., Casali, J.G., Dixon-Ernst, C., Krieg, E.F., Mozo, B.T., Ohlin, D.W., Royster, J.D., Royster, L.H., Simon, S.D., Stephenson, C.

(**2003**). “Development of a new standard laboratory protocol for estimation of the field attenuation of hearing protection devices: Sample size necessary to provide acceptable reproducibility,” Submitted for publication to the J. Acoust. Soc. Am.

OSHA (**1999**). *OSHA Technical Manual, Section III, Health Hazards, Chapter 5, Noise Measurement*, U.S. Department of Labor, Occupational Safety and Health Administration, OSHA Instruction TED 1-0.15A.

Press, W.H., Flannery, B.P. Teukolsky, S.A., Vetterling, W.T., (1986). *Numerical Recipes 1st Ed.*, pp. 200-209. (Cambridge University Press, Cambridge).

Royster, L.H., (1995). Recommendations for the labeling of Hearing Protecors,” Sound Vib., July:16-19.

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.

Suter, A. (**2003**). “History and Use of EPA’s Hearing Protector Labeling Regulation,” Presentation at U.S. Environmental Protection Agency Workshop on Hearing Protection Devices, Washington DC, March 27, 2003.

U.S. Environmental Protection Agency (**1978**). CFR Title 40, sub chapter G, 211, subpart B-Hearing Protective Devices.