Moore, H. B. 1961. "Effect of System Parameters on the Stereophonic Effect." *Journal of the Audio Engineering Society* 9(1):7–12.

Stromeyer, C. F. III,k and P. Greenspun. 1987. "Blind Tonearm Comparison." Unpublished.

Appendix: Statistical Methods Used

We assume that each of our trials is an independent event and that we can therefore consider our data to be a sequence of *Bernoulli trials.* Given the high variability of performance among subjects being tested simultaneously, we feel it is reasonable to make the assumption of independence. If this assumption is false, it is more probable that we could have obtained our results if the subjects were guessing.

In order to determine the probability of our results being due to chance (i.e., guessing), we use the *binomial formula.* This gives the probability of k successes (correct answers) in n trials, given a probability p of success and $q = (1 - p)$ of failure.

$$P(S_n = k) = P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k q^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

and

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

This formula is verified by considering the fact that p^k is the probability of getting k correct answers in succession and q^{n-k} is the probability of getting $n - k$ incorrect answers in succession. The probability of a particular sequence of k correct and $n - k$ incorrect answers is thus $p^k q^{n-k}$. The number of possible sequences of n answers with k correct is $\binom{n}{k}$ ("n choose k").

Given the hypothesis that the subjects were not able to hear any differences and merely guessed, we assume that they have a fifty percent chance of getting any trial correct so $p = q = 1/2$. Thinking of the result of each trial as a *Bernoulli random variable* with an expected value of 1/2 (1 if correct, 0 if incorrect), the number of successes out of n trials, S_n will be the sum of the random variables for each trial. S_n itself is a random variable and the function $p_{S_n}(k)$, which gives the probability that $S_n = k$, is called the *probability mass function* (PMF).

As a result of the *central limit theorem,* the PMF of a sum of Bernoulli random variables will converge to a Gaussian curve. Had we conducted 100 trials and had the subjects been guessing, the envelope of the PMF of the total number correct would be a roughly Gaussian curve centered at 50. The probability that the total number correct due to guessing would be between a and b is simply the sum of all values of the PMF between a and b (the sum of all values of any PMF is 1).

In order to determine the probability that our data was due to guessing, we want to find the sum of values of the guessing PMF for S_n between k (number of correct answers in our particular experiment) and n (number of trials). This is the probability that, had the subjects been guessing, they would have done as well or better than they actually did. We compute this by evaluating the expression

$$P(k \leq S_n \leq n) = \sum_{i=k}^{n} P(S_n = i) = \sum_{i=k}^{n} \binom{n}{i} p^i q^{n-i}.$$

A more severe significance test asks the question "What is the probability of getting results that far from the mean had the subjects been guessing?" In only evaluating "one tail of the Gaussian," we are assuming that if the subjects are able to hear a difference then they will do better than chance. Consider the situation where subjects are always able to identify X, but don't understand the experimental procedure and consistently circle the wrong choice on the ballot. The total correct would be 0 and it is highly unlikely that such a result would occur if the subjects were guessing. In the event that it is possible for subjects to consistently "push the wrong button," one must consider the probability that they would do very poorly due to guessing, by evaluating

$$\sum_{i=0}^{n-k} P(S_n = i) = \sum_{i=0}^{n-k} \binom{n}{i} p^i q^{n-i}.$$

In fact, by symmetry, this is always equal to $\sum_{i=k}^{n} P(S_n = i)$ so we can just double the result from our previous computation. We employed this method when calculating the significance of the preference results, since we had no *a priori* way of knowing which cable is the better sounding product.

If we consider the first identification trial as a training period, we find that listeners identified the mystery presentation correctly 58 times out of 94. Had the subjects guessed, they would have gotten 58 or more correct with probability 1.5%.

Including the data from the first identification trial in our analysis, the listeners got 71 correct out of 122. Subjects who were guessing would get 71 or more correct with probability 4.2%.

Computing the above numbers directly is facilitated by a programming language such as Lisp with arbitrary precision integers because the "n choose k" terms get quite large ($\binom{122}{71} \approx 7^{34}$) and the $(1/2)^k$ terms get quite small. In the words of Alvin Drake, "Should this quantity [$P(k \leq S_n \leq n)$] be of interest, it would generally require a very unpleasant calculation." Many people therefore use the central limit theorem to obtain an approximation that can be easily computed with a hand calculator.

The central limit theorem implies that the sum of a large number of Bernoulli random variables is approximately a variable with a Gaussian distribution. A special case of the central limit theorem, known as the *DeMoivre-Laplace limit theorem*, is useful when p (probability of success in one trial) is close to 1/2.

$$P(a \leq S_n \leq b) \approx \Phi\left[\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$$

where $\Phi(x)$ is a function that gives the area from $-\infty$ to x under the unit normal Gaussian ($\Phi(0) = 0.5$). Values of Φ can be obtained from standard tables or by numerically integrating the formula for the Gaussian probability density function

$$f_x(x_0) = \frac{1}{\sqrt{2\pi}} e^{-x_0^2/2}.$$

See (Drake 1967) for a readable introduction to the above material. Upon request, the authors will be glad to supply Common Lisp software that calculates significance using both methods.