Loudness of 1000 Hz Tones

Adapted from S. S. Stevens (1955):

Method of Magnitude Estimation (ME) with a Standard

Magnitude estimation means the following: A standard tone is presented and an arbitrary number is assigned to its loudness, e.g., 1 or 100. Then a comparison tone is presented, and the subject decides what number he thinks should be assigned to the loudness of the comparison stimulus. Under another version of the method of magnitude estimation, the standard is omitted entirely. The subject hears a series of intensities presented in random order, and to these intensities he assigns numbers proportioned to their apparent loudness.

This is the most direct and, in some ways, the most efficient method. Each presentation of the stimulus is rated numerically, and no information need be thrown away. Like all procedures, the method of magnitude estimation is susceptible to various biases, some of which can probably be avoided or counterbalanced. One bias arises from the fact that the subject’s estimates are influenced by the order in which the stimuli are presented. Since the subject usually tries to be self-consistent, what he says about a given comparison stimulus depends to some extent on what he has said about the preceding ones. This particular bias does not affect the first judgment he makes, however, and it is therefore instructive to compare these first judgments with the later ones. Actually, these first judgments are usually consistent with the later ones.

Another bias arises from the fact that, although the subject tries to judge the comparison relative to the standard, he may be slightly influenced by the absolute level of the comparison tone. He may overestimate the relative loudness of high intensities and underestimate the relative loudness of low intensities. Since in estimating subjective magnitudes the subject is completely unconstrained in the choice of the numbers he assigns, it turns out that an occasional subject may give estimate that are far out of line with those of the group. When this happens, the distribution of the estimates is considerably skewed. And, since the arithmetic mean is not a good measure to use with skewed distributions, it is usually advisable to compute medians rather than means.

Method of Magnitude Estimation (ME) without a Standard

In another experiment I dispensed with a standard stimulus and merely presented to each of 26 subjects a series of eight intensities spaced 10 dB apart from 40 to 110 dB. The order of the tones was different for each subject and each tone was presented twice. All but one of the subjects had previously made loudness judgments in other types of experiments. The instructions were as follows:
I am going to give you a series of tones of different intensities. Your task is to tell me how loud they sound by assigning numbers to them. To turn on the tone you simply press the key. You may press it as often as you like. When you hear the first tone, give its loudness a number—any number you think appropriate. I will then tell you when to turn on the next tone, to which you will also give a number. Try to make the ratios between the numbers you assign to the different tones correspond to the ratios between the loudnesses of the tones. In other words, try to make the number proportional to the loudness, as you hear it.

In order to combine the results for the different subjects it was necessary to bring the estimates into coincidence at a given level (80 db) by multiplying by an appropriate factor—a different factor for each subject. The median estimates were then computed for each level.

The results are shown in Fig. 5. The straight line in Fig. 5 has the slope of the loudness function (10 db = 2:1 loudness).

This experiment goes about as far as it is possible to go toward getting subjects to make direct quantitative judgments of the loudness of sounds. Here we have done nothing more than present a “random” series of intensities, and to each intensity the subject has assigned a number representing his perception of the loudness. The variability is fairly large, but the median judgments fall close to the loudness function and confirm our conclusion that loudness is a power function of intensity.

Figure 1: Results from magnitude estimation experiments Stevens (1955).
Garner and the Equal Discriminability Loudness Scale

Figure 2: A Thurstonian model of an identification experiment.

There is also a group of techniques based on the Thurstone procedures which have been used for loudness scaling more recently. These techniques all make use of direct ratings (absolute judgments) as the experimental technique, but the loudness scale values are based on the dispersions and overlaps of the distributions of responses. Garner and Hake (1951) describe one such scaling procedure which results in an “equal discriminability” scale based on data obtained from the absolute judgment procedure. The one factor which all of these techniques have in common is that they do not depend on the assumed validity of the direct response of the observer, but rather make use of the statistical characteristics of entire distributions of responses. Normally some measure of dispersion is used as the unit of scale value.
Figure 3: The results of Garner (1956).

As a starting point, we are in exactly the same situation with these procedures as we are with the direct response procedures. We start with an arbitrary assumption (taking face validity again) that measures of discriminability or judgmental dispersion are legitimate measures of sensory magnitude. There is no problem here of whether Fechner’s original assumption has been proved invalid by comparison with direct-response methods, since we have no more than face validity for these methods either. However, loudness scales based on the discriminability criterion show much less variability between observers than do scales based on fractionation procedures, as was true also for scales based on the equisection procedure.

In addition, and more importantly, loudness scales based on the discriminability criterion show much less effect of specific experimental procedures and different experimental conditions than do the direct response scales. For example, Garner (1954) showed that an equal discriminability scale based on dispersions of absolute judgments has essentially the same shape as a scale based on the integration of jnd’s, when Riesz’ data were used for the jnd’s. And, of course, there is a vast literature on jnd’s for auditory intensity showing that only to a slight extent is the relation between auditory sensitivity and intensity a function of the
particular experimental method. These various techniques all give loudness functions which are roughly linear with log intensity, although showing some positive acceleration. The equal discriminability scale shows, in addition, a greater slope at the two ends of the continuum presented to the observer for judgment—an effect which seems to be due to subjective anchoring. At any rate, with the relative stability of the loudness function shown with these various techniques, we can feel a bit more certain that we do not have loudness scales unique to a highly specific procedure.

One of the most compelling arguments in favor of the discriminability procedures, however, is the relative imperviousness of these techniques to the context effect. In one experiment of mine, for example, judgments on a 21-point scale were obtained over a wide range of intensities with two quite different distribution of stimuli. In one distribution, the stimuli were spaced every 5 rib. In the other distribution, the stimuli covered the same range of intensities, but the distribution of them was severely skewed, with stimuli crowded into the upper half of the intensity range. This set of stimuli was actually distributed to provide equal spacing on a loudness function based on the observer’s own fractionation judgments, i.e., the stimuli were deliberately distributed to correspond to direct responses of the fractionation type. The discriminability scales based on these two distributions of stimuli are shown in Fig. 5, with the open circles representing stimuli spaced by log intensity, and the closed circles representing stimuli spaced by the fractionation loudness function. Clearly the two sets of points fall on the same function with remarkable closeness.

And equally clearly we are dealing with a relatively invariant and stable property of the human perceptual system when we use a discriminability function. There are, of course, individual differences with these methods, and differences due to experimental method and procedure. But these differences are slight compared to those found with the direct response methods. Certainly by contrast we can feel that the discriminability procedures are measuring some aspect of the perceptual system which is a limiting property—a property little affected by factors which easily affect the ratio scaling procedures.
Mutual Information

Consider a set of stimuli $S = \{S_1, S_2, \ldots, S_N\}$ that an observer is trained to identify with the responses $R = \{R_1, R_2, \ldots, R_M\}$. The stimuli are presented with the a priori probabilities $P(S_1), \ldots, P(S_N)$, and the observer is assumed to assign response $R_j$ to stimulus $S_i$ with probability $P(R_j|S_i)$.

The uncertainty (or entropy) of the stimulus set, $H$, is defined as

$$H(S) = \sum_{i=1}^{N} P(S_i) \log_2 \frac{1}{P(S_i)}. \tag{1}$$

Thus defined, $H$ (in bits) measures the number of yes/no questions that must be asked, on average, by someone aware of the probabilities $\{P(S_i)\}$ in order to determine the identity of the stimulus. It can be shown that

1. $0 \leq H(S) \leq \log_2 (N)$.
2. $H(S) = 0$ when all but one of the $P(S_i) = 0$.
3. $H(S) = \log_2 (N)$ when all $P(S_i) = 1/N$.

The mutual information between the stimuli and the responses, $I(S; R)$, defined as

$$I(S; R) = \sum_{i=1}^{N} \sum_{j=1}^{N} P(S_i, R_j) \times \log_2 \frac{P(S_i, R_j)}{P(S_i)P(R_j)} \tag{2}$$

$$= \sum_{i=1}^{N} P(S_i) \sum_{j=1}^{N} P(R_j|S_i) \times \log_2 \frac{P(R_j|S_i)}{P(R_j)}. \tag{3}$$

measures the average reduction in uncertainty about the stimulus provided by knowledge of the response. It can be shown that

1. $0 \leq I(S; R) \leq H(S) \leq \log_2 (N)$.
2. $I(S; R) = 0$ when $(R_j|S_i) = P(R_j)$.
3. If $N = M$, $I(S; R) = H(S)$ if and only if for each stimulus $S_i$ there is a unique response $R_j$ such that $P(R_j|S_i) = 1$ and $P(R_k|S_i) = 0$ for $k \neq j$.
4. $I(S; R) \leq \log_2 (M)$. 

6
Garner and Mutual Information

Figure 4: The results of Garner (1956).

Garner had 6 subjects identify 1 sec, 1000 Hz tone bursts that ranged in intensity from 15 to 110 dB SPL and were presented ever 7s. Intensities were spaced equally on an equal-discriminability scale. Each stimulus was presented 100 times ($4 \times 100 \times 6 = 2400$ total presentations of 4 intensities, $20 \times 100 \times 6 = 12000$ presentations of 20 intensities.

Mutual information was computed separately for each subject and the results were averaged (Fig. 4.

1. Mutual information is maximum when identifying 5 stimuli spaced by roughly 24 dB).

2. Mutual information is relatively constant (2.1 bits) when identifying more than 5 stimuli corresponding to perfect identification of 4.4 stimuli.

3. Mutual information is nearly 2 (1.90 bits when 4 stimuli are identified, indicating near-perfect identification.

Fig. 3.4. Information transmission for absolute judgments of auditory loudness (After Garner, 1953.)
Ten (10) tone bursts of frequency 1000 Hz and duration 100 ms (10 ms rise/fall time), but differing in level were the stimuli in the identification and discrimination experiments. Successive tonal bursts in the discrimination tests were spaced 1.2 s apart.

Discrimination and identification tests were compared over a number of stimulus ranges. Each range of sound levels consisted of ten equally-spaced steps (on a decibel scale). The spacing between successive steps on the scale varied from zero to 8 db, and, therefore, the extent of the total range of sound levels varied from zero to 72 db. To insure that each range covered a comparable set of sound levels, each range (except the widest) was examined over a “soft” (bottom step at 41 dB SPL), “middle” (center step at 77 dB SPL), and “loud” (top step at 113 dB SPL), set of conditions. For a given range of sound levels, each of the 10 steps served as the reference signal in discrimination tests and, as the to-be-identified signal,
in identification tests. The order of presentation of each of the 10 steps was random; each of the 10 steps occurred an equal number of times in a 100-item test.

In the discrimination (D) tests, each reference signal was paired with an intensity differential. The task of the listener was to indicate whether the second member of the pair was louder or softer than the first. In the identification (I) tests, the task of the listener was to assign correctly a numeral from one through ten to each tonal burst.

A single test was made up of 100 response items and successive responses were spaced at 3.9 sec intervals. Each test was about 6.5 min duration. All test under a given sound level range were carried out before proceeding to the next range. Six subjects, five male college students plus the experimenter, were used. All possessed Clinically normal hearing. The experience of the subjects prior to testing varied from one week (Subjects A and J) to over a year (Subjects Mc and P).

Fig. 1 indicates that the detection threshold increases as the range of variation of the reference signal increases. Specifically as the scale interval between successive steps is increased from 0.5 db to 8.0 dB, the various detection thresholds increase, on the average, by a fact of about sevenfold, in terms of the decibel measure, for all detection criteria.

The results of the identification tests are presented in Fig. 3, in terms of the average information transmitted, as a function of the range of sound levels available for identification (abscissa). Fig. 3 indicates that the information transmitted in the identification tests increases as the interval between successive steps increases. However, only in rare instances, do the transmission scores exceed 2 bits per stimulus presentation.

The computation of equivalent transmission scores in the discrimination experiment involved 3 stages: First–over the range of conditions examined, and for the specific discrimination criterion chosen, the number of discriminably different steps was determined. This stage requires the usually employed assumption that discrimination thresholds may be obtained independently over the stimulus range examined. Second–for the specific criterion chosen, the equivalent information transmission associated with that criterion of discrimination in a two-choice situation was calculated. At one level, this stage may simply be regarded as an alternative statistical statement of a set of response probabilities. More precisely, however, this stage requires the assumption that the statistical requirements of the informational model have been met. Third–the product of the number of discriminable steps over the stimulus range examined by the equivalent transmission per step yielded the equivalent transmission over the entire stimulus range examined.

For a stimulus range of 9 dB, the average detection threshold meeting the criterion of 70% correct discrimination for Subject P was about 2 dB. If we assume that successive thresholds may be obtained independently over the range, there are 4.5 discriminable steps over the 9 dB range for the 70% correct detection criterion. Associated with the 70% correct detection criterion in a two-choice situation is an information transmission of about 0.119 bit per discrimination. Thus, 4.5 steps at 0.119 bit per step yields a total equivalent transmission of 0.534 bit.

This model is incorrect. Consider the case where the subject is performing at 99% correct in discrimination, so that the mutual information is 0.92 bits per stimulus presentation.
According to Pollack, the equivalent transmission would be 0.92 times the number of discrim- 
inanle steps. But it is well known that in this case, the mutual information is proportional 
to the log of the number of essentially perfectly discriminable steps.
Fig. 3 Identification performance as a function of the extent of the range of the to-be-identified signal. Each section is associated with the results of a separate listener. The abscissa is the interval between successive steps along the range of sound levels examined. The entire stimulus range is 9 times the corresponding scale interval. The ordinate is the average information transmitted, in bits per stimulus presentation. The crosses at 0 db represent estimates of sampling bias based on evaluation of test results with a random key. The results associated with the soft, middle, and loud sections of the stimulus range are represented by open, half-filled, and closed figures, respectively. The results associated with the I, I-D, and I and D series are represented by circles, triangles, and squares, respectively. The number of determinations associated with the three series are: four-five, two-three, and eight-ten 100-item tests, respectively.

Figure 6: The results of Pollack (1956).
Fig. 4. Comparison of identification and discrimination tests on the basis of an informational evaluation of performance. The difference between successive steps of each stimulus scale is presented on the abscissa. The parameter on the curves is the criterion of detection arbitrarily selected in the discrimination tests. The ordinate represents the ratio of: (1) the information associated with discrimination responses for the indicated threshold criterion to (2) the information associated with identification responses for the same range of sound levels. Each section represents the results for a single listener.
Kinchla and Smyzer

Figure 8: The variance as a function of time according to the data of Kinchla and Smyzer (1967). The circles, crosses, and squares are data points from three different subjects. Analysis indicates that the variance would double over its \( T = 0 \) value in 3.9, 5.5, and 2.3 s respectively.

Kinchla and Smyzer (1967) proposed that an observer’s sensitivity is discriminating two bursts of different intensity and separated by time \( T \) was given by

\[
d' (I, I + \Delta I) = 10 \log_{10} \left( \frac{I + \Delta I}{I} \right) \frac{1}{\sqrt{(A + BT)}}
\]

so that

\[
\left[ \frac{10 \log_{10} \left( \frac{I + \Delta I}{I} \right)}{d' (I, I + \Delta I)} \right]^2 = A + BT
\]

Their data are shown in Fig. 8. The data indicate that as \( T \) increases, the variance increases roughly linearly, doubling from its \( T = 0 \) value in 3.89 s, on average.

Bibliography

