

S. S. STEVENS Psycho-Acoustic Laboratory, Harvard University, Cambridge, Massachusetts (Received June 17, 1955)

This paper reviews the available evidence (published and unpublished) on the relation between loudness and stimulus intensity. The evidence suggests that for the typical listener the loudness L of a 1000-cycle tone can be approximated by a power function of intensity I, of which the exponent is $\log_{10}2$. The equation is: $L = kI^{0.3}$. Intensity here is assumed to be proportional to the square of the sound pressure.

In terms of *sones*, where 1 sone is the loudness produced by a tone at 40 db above the standard reference level, the equation for loudness L as a function of the number of decibels N becomes: $\log L = 0.03N - 1.2$. Otherwise said, a loudness ratio of 2:1 is produced by a pair of stimuli that differ by 10 db, and this re-

lation appears to hold over the entire range of audible intensities.

At low levels of intensity, the loudness of white noise grows more rapidly than the loudness of a 1000-cycle tone, but above the level of approximately 50 db the two loudnesses remain more nearly proportional. The suggestion is made that for all levels greater than 50 db the loudness of continuous noises may be calculated from the equation: $\log L = 0.03N + S$, where S is a spectrum parameter to be determined empirically.

THE purpose of this review is to examine the available data on the measurement of subjective loudness. It is hoped that by assembling the relevant information in one place we may be able to reach a reasonable conclusion concerning the relation between loudness and intensity. The various results obtained by workers in this field make it plain that the scale relating loudness to intensity is not something that can be determined with high precision, but these efforts also make it plain that people are able to make quantitative estimates of loudness and that it is not unreasonable to try to determine a loudness scale that will be representative of the typical listener.

Despite its many pitfalls, I think we can probably make sense of this problem provided we are sufficiently modest in our demands. Not only must we renounce the hope for high precision, but also we must keep in mind the nature of the central issue. What we want to know is how loud various stimuli sound to people. By this we mean, perforce, what do people say when they try to describe loudness in quantitative terms? In asking this question we are not at the outset trying to solve the problem of how the ear works or how the nervous system performs its integrations. We are not trying to count nerve impulses nor prove a theory. We are merely looking for the empirical answer to a very empirical question: How do people describe sounds when we ask them to use a numerical language instead of adjectives?

Our interest in asking this question may be academic or it may be practical. The academic side of the issue has a long history full of side-taking and polemics.¹ The practical side of the problem had its origin in acoustical engineering. Not long after they had developed the conventional decibel scale for measuring sound intensity, the engineers noted that equal steps on the decibel scale do not *sound* like equal steps, and that a level of

¹ E. G. Boring, Am. J. Psychol. 32, 449-471 (1921).

^{*} This work was carried out under Contract N5ori-76 between Harvard University and the Office of Naval Research, U. S. Navy (Project NR142-201, Report PNR-168). Reproduction for any purpose of the U. S. Government is permitted.

50 db does not sound like half of 100 db. Since the engineer often faces the problem of communicating with a customer, it was soon realized that there was a need for a scale whose numbers would make more sense to the customer than do the numbers on the decibel scale.

The generation of a loudness scale is in principle quite simple. All we need to do is produce an array of sounds and ask a group of listeners to assign numbers to these sounds in such a way that the numbers reflect the perceived loudness of the sounds. In practice, of course, it turns out that many alternative techniques are possible, and that subtle differences in experimental procedure sometimes influence what the listener says or does. When the listener tries to tell us about the relative loudness of two sounds, he is subject to a host of potentially biasing factors. Some of these factors are built into the listener; some are supplied by the experimenter. Some are easy to discover; others are as elusive as foxes in a forest. The parameters that affect experiments on loudness are numerous and sometimes so conflicting that it may be impossible under any one procedure to optimize all of them. Consequently, the definitive experiment in this field, if we may hope for such, will probably need to involve a multiple attack of the sort that might balance out the various sources of distortion and bias.

Nevertheless, under a wide variety of conditions rather similar results continue to be obtained. Provided we are willing to be content with modest precision, we can predict pretty well what the average listener will say about loudness in most ordinary situations. It is my own belief that this predictability is sufficiently high to justify the standardization of a loudness scale to represent for the "standard observer" the relation between loudness and intensity.

METHODS

Three principal methods have been used to obtain direct estimates of the relation between loudness and intensity: bisection (or equisection), ratio determination (including fractionation), and direct magnitude estimation. Methods that involve the comparison of one-vstwo tones and one-vs-two ears are not included here because they rest on assumptions that can be tested only by the direct comparison techniques that are required to establish a loudness scale in the first place.

The method of bisection, involving the determination of the loudness that lies midway between two given loudnesses, has its limitations, owing principally to the fact that it leads only to an interval scale² (analogous to the temperature scales, Fahrenheit and Celsius). Its results do not allow us to set up a ratio scale (analogous to scales of weight and length). Ratio scales can, in principle, be determined by the methods of fractionation and magnitude estimation, and such ratio scales can be used to predict what ought to happen in a bisection experiment. Thus bisection might provide a test of the generality of a ratio scale, but by itself bisection cannot generate a ratio scale.

In the tabulations that follow only the results obtained by magnitude estimation and ratio determination will be considered.

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.

Ratio determination means those procedures that aim to discover what intensity produces a loudness bearing a prescribed ratio to a given loudness. The ratio may be a fraction or a multiple of the standard.

We can further divide the method of ratio determination into two classes: (1) The subject is allowed to adjust the intensity to produce the desired ratio (method of adjustment). (2) The experimenter sets the intensity and asks the subject to say whether it is too high or too low (method of constant stimuli). Of course, from a series of magnitude estimations a ratio determination can be made by a process of interpolation. Consequently, it is possible to combine the results of experiments that use these two different procedures.

Each of these methods has its advantages and its disadvantages. And, of course, each has many subvarieties, some of which are better or worse than others.

Merits of the Methods

Let me try to list some of the assets and liabilities of these methods as they seem to be disclosed in the studies of other experimenters and in our work at the Psycho-Acoustic Laboratory.

Method of Magnitude Estimation (ME)

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.

² S. S. Stevens, editor, *Handbook of Experimental Psychology* (John Wiley and Sons, Inc., New York, 1951), Chap. 1.

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 estimates 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 Constant Stimuli (CS)

This method differs from magnitude estimation in that the subject merely indicates whether the comparison stimulus is greater than, less than, or equal to, some criterion, such as "half as loud." Some experimenters have changed the comparison in the direction indicated by the subject's response and recorded only the value the subject said was equal to the criterion. Others have limited the subject's responses to "greater" and "less" only, and presented comparisons at random until sufficient judgments have been made to determine a "psychometric function" from which the "equal" point could be found. In this more rigorous form, the method of constant stimuli is relatively time-consuming and laborious.

The principal source of bias in the method of constant stimuli seems to stem from the fact that the range, level, order, and spacing of the comparison stimuli may influence the subject's judgments.

There is also an order effect called the "time error" that seems to operate when stimuli are presented in succession. This effect is evidenced by the fact that when two successive stimuli are physically equal the second tends to be judged louder than the first. The magnitude of the effect is usually not large, and its role in loudness ratio judgments is not easy to determine. It could mean that when a comparison tone following a standard is set to half loudness the setting is slightly biased in the direction of being too low.

Method of Adjustment (adj)

This method puts the task of finding the criterion level under the control of the subject. He turns a dial to adjust the comparison tone. A potential source of bias in the method of adjustment derives from the relation between loudness and the angular turning of the dial. Most experimenters have used logarithmic attenuators on which the scale (decibels) is very nonlinearly related to loudness. The use of this nonlinear control apparently leads the subject to set the dial too low when adjusting for half loudness. It is possible, of course, to arrange a potentiometer so that the position of its knob is approximately proportional to loudness. We find that the results obtained with such a "sone potentiometer" are different from, and presumably superior to, the results obtained with decibel attenuators.⁸

The design of a sone potentiometer can be achieved, of course, only after we have some notion of the form of the loudness function. In principle we might proceed by successive approximations. Beginning with an arbitrary potentiometer (e.g., logarithmic) we would determine a first approximation to the loudness function, which we would then use to modify the characteristic of the potentiometer. With this modified potentiometer we would redetermine the loudness function, and again modify the potentiometer, and so on. In practice, however, the bias due to a decibel attenuator can probably be effectively eliminated by substituting a potentiometer whose angular turn is roughly proportional to the loudness function first determined with the decibel attenuator.

The intensity of the comparison tone when the subject first hears it at the beginning of his adjustment also seems to make a difference. This bias can presumably be avoided if the experimenter sets the comparison at random about the criterion point.

Whatever method is used, it probably makes a lot of difference how the subjects are treated. Their attention flags quickly, and whatever biasing forces are at work can produce more formidable distortions when the subject's attention has wandered. Judging loudness is at best a delicate, difficult business.

Another source of bias arises from the common practice of combining results by averaging decibels. This would be a proper average if loudness were linearly related to decibels, which it is not. Since it is loudness we want to average, it might be better, before we average, to turn all the decibel values into loudness values (sones) by means of a reasonably approximate loudness scale. Then the averaged sones may be converted back to decibels via the same loudness scale.⁴ When we are concerned with the decibel reduction needed to produce half loudness, the difference between averaging decibels and averaging sones may amount to a decibel or more, depending on the original variability. Evidence of this fact is seen in Table I where the median decibel reductions reported by Geiger and Firestone⁵ are smaller than the decibel averages. The median is a less efficient statistic than the mean, but in experiments on fractionation the median of the decibel values usually comes closer to what we would get if we averaged sones instead of decibels.

As a matter of fact, experimenters seem generally to have paid too little attention to the problem of

³ S. S. Stevens and E. C. Poulton, "The estimation of loudness by unpractised observers," J. Exptl. Psychol. (to be published). ⁴ S. S. Stevens, Science 121, 113-116 (1955).

⁶ P. H. Geiger and F. A. Firestone, J. Acoust. Soc. Am. 5, 25–30 (1933).

averaging their data. It is a difficult problem and there is no universal solution to it. There are many potential sources of skewness and no one averaging procedure can undo all of them.

For experiments on loudness a defensible rule might be this: When a conversion of the decibel values to sone values eliminates the skewness in the data, it is advisable to average in sones. When this conversion to sones does not eliminate the skewness in the data, medians should be computed. Medians must, of course, be used with caution when the number of cases is small.

Another general source of distortion in loudness judgments, shared more or less by all the methods, arises from the preference the subjects show for listening to moderate levels of loudness. They seem to avoid listening to sounds that are too loud or too weak by adjusting the comparison stimulus a little too high when it is in a low range and a little too low when it is in a high range. Consistent with this they underrate the loudness of a faint sound relative to a faint standard, and they overestimate the loudness of an intense sound relative to an intense standard. The center of the range, between 60 and 80 db above threshold, seems to be the preferred level for listening⁶ and it is as though departures from this region were "resisted" by the subjects.

The consequences of this tendency show up in almost every experiment that requires a subject to judge one stimulus relative to another. Not only is the variable stimulus judged relative to the standard, but also to a slight extent it appears to be judged relative to the "comfortable listening level." It is as though the effective standard stimulus, from the point of view of the subject, were systematically shifted toward the comfortable level. The resulting constant errors can probably be ascertained and eliminated by a balanced procedure that makes each stimulus serve once as the standard and once as the variable. In ratio determination, this balancing procedure would call for both fractional and multiple loudness judgments, and an averaging of the two results.

Many of the foregoing problems relating to loudness judgments are further discussed and illustrated in the Appendix.

DATA

Tables I to IV record the decibel values that correspond to a loudness ratio of 2:1. Other ratios have not been recorded in these tables, but whenever an experiment has determined a ratio of 4:1 it has been used to estimate an additional entry of 2:1 by taking half the decibel ratio corresponding to 4:1. These entries are italicized. Ratios greater than 4:1 have not been used for this purpose, despite the fact that higher ratios often give results quite consistent with those obtained for 2:1 itself. Brief descriptions of the experiments on which these tables are based are given in the Appendix.

TABLE I. Tones-decibel reductions required for half loudness.

Yea	ar and investigator	Frequency	Method subject	l and cts	i 1020	2030	30-40	40–50	Decibels 50–60	re 0.0002 60-70	dyne/ci 70-80	n² 8090	90–100	100–110	110-
1930	Richardson and Ross		ME	11							12				
1932	Laird, Taylor, and Wille	1024	CS	10	5.0	13.5	15.1	19.0	19.0	19.5	20.5	20.0	22.0	24.0	
1932	Ham and Parkinson	350 1000 2500	ME 53 ME ME 54	-54 51 -55				7.9	10.0 7.2 9.0	11.0 8.2	10.0 9.4	10.0 7.8 9.6	7.8		
	Warble ton Warble ton Warble ton Warble ton Warble ton	e 500-1000 e 260500 e 500-1000 e 260500 e 500-1000	ME CS 35 CS 13 CS 32 CS 13	18 37 16 39 16			9.0	10.8	7.0 8.8 9.2	9.0 9.4 6.8 <i>9.1</i> 9.0	9.4 11.3 9.5 9.3 8.9 7.8	8.8 7.6 6.7 8.6	7.6		
1933	Geiger and Firestone	60 1000 1000	adj adj adj	31 44 44	(mean) (median)	2.1 3.6 3.1	4.1 4.3 3.5		3.5 7.5 7.0	5.9 9.4 8.8	5.8 9.8 9.5	7.5 11.7 10.0			
1934	Churcher, King, and Davies	800 800	adj adj	34 30				8.0 7.0	8.5	10.0 <i>8.0</i>	10.0	12.0 15.0		16.0	
1936	Rschevkin and Rabinovitch	1000 1000	CS CS	11 11				9.8	12.0	12.8 <i>13.5</i>	12.3 <i>13,2</i>	11.4 <i>14.0</i>	10.1 <i>11.0</i>	10.5	
1951	Pollack	1000	adj	7	3.2	2.5	3.0	4.1	6.2	6.4	7.9	8.0	8.1	8.9	
1952	Garner	1000	adj	18	6.3	9.4	13.0	16.4	17.9	18.0	17.3	17.4	16.6	14.6	
1953	Robinson	1000	cs	25	5.0		10.3		11.6	12.8	11.6	10.6	10.8	10.3	
1954	Garner	1000	adj	18		7.4	10.2	13.2	15.6	16.7	17.8	17.9	17.7	16.3	17.5
P.A.L.	Stevens and	1024	CS	20				6.9	10.2	7.5	9.4	9.8	8.5	9.3	
	Stevens Stevens and Poulton	1000 1000 1000	000 ME 14-17 000 adj 22, 33 000 adj 22. 33		(mean db) (mean sones)					9.0	9.5	9.0 7.8	9.5 9.5 8.7	9.0	13.0
Poul Poul	Poulton Poulton	1000 1000 1000 1000	adj 22, ME ME ME ME	, 33 8 8 32	(median) (numerical es (fractional est	timates) timates))				13.0	7.4 13.5	8.1 10.8 11.1 11.2	12.0 12.0 10.0	

⁶ I. Pollack, J. Acoust. Soc. Am. 24, 158-162 (1952); T. Somerville, BBC Quart. 3, 11-16 (1949).

	Year and investigator	Frequency	Metho subj	od and ects	2030	30-40	40-50	Decib 50–60	els <i>re</i> 0.0 60–70	002 dyn 70-80	e/cm² 80–90	90100	100-110 110-
1932	Ham and Parkinson Warble to Warble to	ne 500–1000 ne 250—500	CS CS 4	19 0–47			8.5 12.8	9.5 12.8	9.8 7.2	6.1			
1932	Geiger and Firestone	60 1000 1000	adj adj adj	31 44 (n 44 (n	nean) nedian)	5.6 10.6 9.5	5.3 8.4 8.0		5.3 7.0 6.1	5.3 7.1 6.5	6.6 9.1 8.2	4.8 7.5 7.5	
1936	Rschevkin and Rabinovitch	1000 1000	CS CS	11 11		18.5	16.3	17.2	15.5 16.7	10.8 <i>14.0</i>	9.0 11.3	10.8 9.5	7.3
1951	Pollack	1000	adj	7	3.5	6.3	9.9	9.7	9.3	10.8	7.5	6.5	
1953	Robinson	1000	CS	25	17.5	17.2	17.0	17.0	14.6		10.3	10.3	
P.A.L	. Poulton	1000	ME	8					12.0	11.7			
P.A.L	. Poulton	1000	ME	32			11.0	12.0	7.5	8.5			

TABLE II. Tones-decibel increase required for twice loudness.

In the tables the values recorded are the decibel reductions required to produce half loudness or double loudness. Since not all experimenters used the same standard levels, the various columns in the tables are made to represent intervals of 10 db, and if the standard used fell anywhere within a given 10-db interval it is recorded in the appropriate column. The justification for this grouping of results rests on the fact that a small change in the level of the standard stimulus produces a negligible change in the decibel difference required for a loudness ratio of 2:1.

These tables, then, provide an inventory of most of what has been done about loudness. The next problem is, what do we conclude from these data?

Treatment of the Data

The entries of Tables I and II have been plotted in Figs. 1 and 2. Each point represents a table entry, except that the mean values have not been plotted when medians were available and the data for 60 cps have been omitted.

These plots are encouraging or discouraging, depending on one's level of aspiration in these matters. Nevertheless, these are the data, and we must now try to decide what they mean. Many possible schemes could be used to strike some sort of average and thereby determine the decibel change corresponding to a loudness ratio of 2:1. If we could decide which data are most trustworthy we could weight them heaviest. Or we could use some other weighting, based perhaps on the number of subjects used. Of course, after working on this problem for a couple of years, and after using the various methods, I find it difficult not to form strong opinions on this question of which data to trust, but opinions are always liable to be wrong.

Since the data in Figs. 1 and 2 are obviously skewed, perhaps the safest objective procedure is to consider the medians of the columns as the most representative values. The median is usually a good statistic to use on skewed data, provided the number of cases is not too small.

An important question that must be decided at the outset is whether to combine the data for halving and doubling. At low intensities, half loudness seems to require a smaller decibel difference than twice loudness. At high intensities the reverse is true. But as we have already noted, there are biasing forces that tend to produce this kind of result. If we assume that the effects of these forces are roughly equal and opposite, we can argue that data on halving and doubling should be

			Method and				Deci	bels re 0.	.0002 dy	ne/cm ²			
7	ear and investigator	Kind of noise	subjects	20–30	30–40	40-50	50-60	60–70	70–80	80–90	90-100	100-110	110-
1932	Laird, Taylor, and Wille	audiometer buzz	CS 10	9.3	13.0	15.5	17.5	19.0	18.5	20.5	21.5	26.0	
1932	Ham and Parkinson	recorded noise	ME 54				6.5	8.5	8.0	10.1			
1933	Geiger and Firestone	40 tone complex	adj 31 (mean) (median)	4.3 3.9	4.9 5.0		7.0 7.0	10.4 10.0	9,3 9.0	11.0 12.0			
1951	Pollack	white noise	adj 7	3.2	3.2	4.7	5.5	6.7	8.7	8.7	8.7	7.4	8.3
1953	Robinson	white noise	CS 25	7.3		9.6	10.0		12.5		12.3		
		white noise									12.4		
P.A.L	. J. C. Stevens		ME 10							10.0	10.0		
P.A.L	. Poulton and Stevens	white noise	adj 16 (m	ean db)		6.7	7.1	8.6	9.4	10.0	9.1	9.4	10.0
		white noise	adj 16 (me	edian)		6.0	6.5	7.7	8.8 8.6 7.9	8.6	8.9	8.5	10.0

TABLE III. Noise-decibel reduction required for half loudness.

TABLE IV. Noise-decibel increase required for twice loudness.

Yea	ar and investigator	Kind of noise	Metho subje	d an ects	id 10–20	20–30	30-40	D 40-50	ecibels <i>re</i> 50–60	0.0002 60-70	dyne/cm [;] 70–80	80-90	90-100	100-110	110-
1932	Laird, Taylor, and Wille	audiometer buzz	CS	10	14.3	16.0	20.0	19.5	17.5	19.0	15.0	14.5			
1933	Geiger and Firestone	40 tone complex	adj adj	31 31	(mean) (median)		6.2 5.5	5.8 5.0		5.7 5.0	6.0 5.5	8.5 7.3	7.8 7.5		
1951	Pollack	white noise	adj	7		2.7	2.3	4.7	7.7	9.6	8.5	7.9	6.2	4.9	4.8
1953	Robinson	white noise	cs	25	12.7		13.6	14.5	12.3	11.3		8.1			
		engine noise	CS	25				13.1			12.0		8.0		
P.A.L	. J. C. Stevens	white noise	ME ME	10 10	(mean) (median)			9.0 11.0	9.0 11.0						
P.A.L	. Poulton and	white noise	adj	16	(mean db)		8.4	9.0	8.7	9.5	8.1	7.7	7.0	6.9	
	Stevens		adj	16	(median)		7.4	8.9	8.8	9.4 9.4 9.4	7.9	7.8	6.1	6.5	

combined.⁷ It then turns out that to a fair approximation the decibel difference corresponding to a loudness ratio of 2:1 is constant throughout the whole intensity range.

820

If we accept this notion, it follows that loudness can be described by a power function of intensity, and we can write a simple formula for it. Furthermore, if we know that loudness is proportional to intensity raised to a (fractional) power we can, in principle, use bisection data to determine the exponent. For if loudnesses have been adjusted so that $L_3-L_2=L_2-L_1$, and if $L=kI^n$, then, canceling the k's, $I_3^n-I_2^n=I_2^n-I_1^n$.

Since we assume the three intensities I_1 , I_2 , and I_3 were measured in the experiment, the value of n is determinable—at least by a process of cut and try. Some of the difficulties that attend bisection experiments will be discussed later.

Our problem now is to decide what decibel difference corresponds to a ratio of 2:1 in loudness for tones in the vicinity of 1000 cps. The medians of the values in the tables suggest that this decibel difference is in the vicinity of 10 db. Actually the median of all the 178 values in Tables I and II taken together turns out to be exactly 10.0 db. The skewness of these data is shown by



FIG. 1. The decibel reductions required to produce half loudness. Each point represents an entry from Table I, except that mean values are omitted when medians are available. The data for 60 cps are omitted, and the two rows of entries for Poulton involving eight subjects have been combined. The open circle to the right of each column indicates the median of the column.

⁷ E. C. Poulton and S. S. Stevens, J. Acoust. Soc. Am. 27, 329-331 (1955). the fact that the arithmetic mean is 10.9 db, which is higher than the median. The standard deviation is 3.9 db. The range on either side of the median that includes 50% of the cases is 2.4 db (the semi-interquartile range).

Figure 3 shows a histogram of all the data for the halving and doubling of tones. The values are grouped by intervals of 0.5 db.

Since in Tables I and II some experimenters are represented by more entries than are other experimenters, it might be instructive to determine one representative value for each experimenter by taking the median of all his values. When this is done, it turns out that the median of these medians is 10.3 db. If we omit the single and somewhat questionable value entered for Richardson and Ross, this median becomes 10.0 db.

Many other procedures for treating these data can readily be thought of, but against any given procedure reasonable objections can probably be urged. There appears to be no escape from the fact that a certain degree of arbitrary judgment must attend our choice of a loudness scale for our typical "standard observer."

With this arbitrariness fully in mind, I should like to propose that for the 1000-cycle tone we accept the value of 10.0 db as the intensity ratio corresponding to a loudness ratio of 2:1. The exponent of the power function then becomes $\log_{10} 2$ and the formula for loudness becomes $L = kI^{0.3}$, where I refers to energy flux density and is assumed to be proportional to the square of the sound pressure. In terms of sound pressure p the formula is $L = k'p^{0.6}$. If we measure I in units equal to the reference level of 10^{-16} watt per cm² and define L=1 sone when I=40 db—then k becomes 0.06 and we have $L=0.06I^{0.3}$. Or, if we measure the stimulus as the number of decibels N above the reference level we can write

 $\log L = 0.03N - 1.2$.

This is a simple formula to use, but for most practical work it will probably suffice to remember that the loudness of a 1000-cycle tone goes up by a factor of two for each 10 db-increase in the stimulus.

It will be noted that the chief difference between this

revised sone scale and those previously constructed concerns the slope of the function at levels below about 40 db. Over this range the old scales were steeper, and their steepness increased as the threshold was approached. This steepness is consistent with the results of experiments on the halving of loudness, but if we confine our attention to halving alone we are viewing only one side of the coin. The outcome of experiments on the *doubling* of loudness suggests that the loudness function becomes less steep near threshold, rather than more steep. Since halving and doubling appear to be subject to a bias that affects them in opposite directions, it seems reasonable to let the loudness scale be determined by a compromise between the results of these two procedures. The simple power-function loudness scale is such a compromise.

LOUDNESS AT THRESHOLD

From the foregoing equation we see that at the threshold of hearing, loudness is not zero. If we assume that the threshold is equivalent to the reference level, then at threshold L=0.06 sone. That the threshold loudness



FIG. 2. The decibel increases required to produce twice loudness. Each point represents an entry from Table II, except that mean values are omitted when medians are available, and the data for 60 cps are omitted. The open circle to the right of each column represents the median of the column.

should turn out to be a small fraction of a sone is not entirely unreasonable, provided loudness is fundamentally quantal in nature-provided, in other words, loudness is not infinitely divisible. What this means is simply that if a sound is heard at all, its loudness is finite. As Robinson⁸ suggests, when intensity crosses the threshold, loudness comes on with a jump.

Perhaps it will prove possible to obtain a more direct estimate of the sone value of a just audible tone, but until that is done the threshold loudness given by the formula is probably as good as any.

SOME FURTHER EXPERIMENTS

After the foregoing summary of the loudness experiments was prepared, I undertook some additional experiments on the magnitude estimation of the loudness of 1000-cycle tones.⁹ These experiments give additional



FIG. 3. Histogram of the results for the halving and doubling of the loudness of tones. The data of Figs. 1 and 2 are combined, making 178 entries.

support to the conclusion that a loudness ratio of 2:1 corresponds to an intensity ratio of 10 db. I used a variety of procedures, including the making of direct loudness estimates over a range as great as 90 db at a single sitting.

Figure 4 shows some of the results obtained when the estimates were made relative to standard intensities of medium level. The median of the estimates of 18 subjects shows that the loudness ratio corresponding to intensities separated by 90 db is approximately 1000 to 1.

The slopes of the straight lines in Fig. 4 are such that 10 db corresponds to a 2:1 loudness ratio. The fact that the data in Fig. 4 fall slightly below this loudness function at the low intensities and above at the high intensities has been shown to be due to the level of the standard chosen. In other words, the slope of the function obtained is altered when the standard is shifted, so that slopes can be obtained that are both greater and less than that of the straight lines in Fig. 4.



FIG. 4. Magnitude estimations relative to a fixed standard stimulus. Upper curve: the loudness of the standard (80 db) was called 10. Lower curve: the loudness of the standard (90 db) was called 10. The points are the medians of 36 loudness judgmentstwo by each of 18 subjects. The vertical lines mark the interquartile ranges.

⁸ D. W. Robinson, Acustica 3, 344–358 (1953). ⁹ S. S. Stevens, "The direct estimation of sensory magni-tudes: loudness," Am. J. Psychol. (to be published).

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 in-



FIG. 5. Magnitude estimates of loudness made with no fixed standard. The eight different intensities were presented in irregular order, and the subject estimated the loudness by numbers of his own choosing. These numbers were multiplied by appropriate factors to make each subject's estimates at 80 db average 10. The points represent medians of the resulting values and the vertical lines represent the interquartile ranges.

tensities, 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.

NOISE

As shown in Tables III and IV, the data for noise are not as numerous as those for tones. Although these tables include results on other types of noise, in trying to analyze the data we shall consider only those for white noise. From these values for white noise, plotted in Fig. 6, it would appear that the decibel change required for a loudness ratio of 2:1 may be somewhat less than 10 db. The median of all the values for white noise is 8.5 db and the mean is 8.4 db. I might point out, parenthetically, that these values are strikingly similar to the decibel change I have found to be necessary to produce a 2:1 difference in the apparent brightness of a white light.

On the basis of the evidence in Fig. 6 we might well conclude that, with white noise, the decibel difference required for a 2:1 ratio in loudness is relatively invariant with intensity. This is the conclusion we reached for pure tones. As we have seen, it is an attractive conclu-



FIG. 6. The median decibel values for the halving (filled circles) and doubling (open circles) of white noise. The data are from Tables III and IV.

sion, because it entails a simple power-function relation between loudness and intensity. Furthermore, if the loudness of tones and the loudness of noise are both power functions of intensity, then the ratio of these two loudnesses is itself a power function of intensity.

But here we run into trouble, for this last conclusion does not agree with the results of experiments in which subjects have been asked to equate the loudness of a tone to the loudness of a white noise. These experiments¹⁰ are not themselves in particularly good agreement, mainly because the task of equating the loudness of a pure tone to the loudness of a white noise is difficult. The difficulties are not unlike those encountered in the frustrating art of heterochromatic photometry. Subjects seem to disagree more when they try to match the loudness of a tone to that of a noise then when they try to

 ¹⁰ F. H. Brittain, J. Acoust. Soc. Am. 11, 113-117 (1939); H.
Fletcher and W. A. Munson, J. Acoust Soc. Am. 9, 1-10 (1937);
G. A. Miller, J. Acoust. Soc. Am. 19, 609-619 (1947); D. W.
Robinson⁸; I. Pollack.²⁶

set one tone to half the loudness of another. But all the experiments relating tones to noise seem to concur on one point, namely, that the ratio between the loudness of tones and noise is not a power function of intensity.

Figure 7 shows the data as I am able to read them from published graphs. The curves represent the difference in decibels between the level of a 1000-cycle tone and that of a white noise that matches the tone in loudness. Why are these curves so far apart? Is it owing to differences in the methods used to measure the noise, or to the large variability among subjects, or to both? Whatever the reason for these discrepancies, all five of the curves are somewhat similar in that they are more or less concave downward. Three of the curves show a



FIG. 7. The results of five experients in which the loudness of a tone was matched to the loudness of a noise. All but Fletcher and Munson used an approximately white noise. In these experiments it appears that the variable stimulus was the tone.

definite maximum. The curves are not the straight lines they would have to be if, for both tones and noise, loudness were a power function of intensity.

It is clear from Fig. 5 that, at low levels, a given increase in the stimulus produces a greater change in the loudness of a white noise than in the loudness of a tone. This fact seems certain. Apparently, when we start from below threshold and increase the intensity of a white noise, the amount of energy surpassing the threshold increases very rapidly for a time. Since all the energy that exceeds the threshold presumably contributes to loudness, it is not unreasonable that the loudness of white noise should grow rapidly over the first few decibels above threshold. This rapid growth of loudness is not inconsistent with at least some of the data in Tables III and IV—those of Pollack and of Poulton and Stevens.

On the other hand, the fact that some of the curves in Fig. 7 turn down at the higher intensities suggests that over the upper range the loudness of white noise grows *less* rapidly with intensity than does the loudness of a 1000-cycle tone. But in Tables I to IV there is no evidence for a less rapid growth in the loudness of noise. Obviously we face a discrepancy that needs explaining.

A possible explanation can be found in the fact that in the experiments underlying the curves in Fig. 7 the variable stimulus seems to have been the tone and not the noise. As we have previously noted, it appears to be



FIG. 8. Results when the subject adjusts the loudness of a white noise to equal that of a 1000-cycle tone. Each point is the median of 24 adjustments—two by each of twelve subjects. The vertical lines mark the interquartile ranges.

a general rule that subjects overestimate an intense variable relative to an intense standard—and that they do the opposite at low levels. Consistent with this general rule, when the subject adjusts a stimulus to be equal to an intense standard he tends to undershoot, and he tends to overshoot when the standard is weak. Thus, in the extreme case, when the standard borders on painful, the subject does not err in the direction of setting the variable too high. As we should expect, he errs in the direction of setting it too low.

We can predict, therefore, that it makes a difference whether the variable stimulus is the tone or the noise. A convincing demonstration of this difference was obtained in a preliminary experiment on a dozen subjects using the method of adjustment. One time the subject adjusted the tone, and another time he adjusted the noise. His two types of settings were inconsistent with each other at the high intensity levels, and they were equally inconsistent, in the opposite direction, at the low intensity levels. The discrepancies at 30 and at 100 db were of the order of 5 db.

Following this encouraging outcome, I improved the apparatus and undertook to test another dozen subjects, using the noise as the variable. The subject used a "sone potentiometer" to adjust the intensity of a 10-kc band of white noise fed through a pair of PDR-10 "extended range" earphones. The response of these earphones was approximately "flat" up to 7500 cps and down about 10 db at 10 kc. A rotary switch presented the tone and the noise alternately. Each stimulus presentation lasted about 1.3 sec, with a silent interval of about 0.5 sec. The levels of tone and noise were measured with a Ballantine electronic voltmeter (model 300), and for our present purpose the sound-pressure level is considered the same when the voltage reading is the same. (Actually, when measuring white noise the Ballantine read about 1 db lower than a thermocouple.) The onset and decline of the stimuli were governed by an electronic switch set for a 5-msec rise-time.

The results are shown in Fig. 8. We see that when the noise is varied there is no tendency for the curve to turn down at the high intensities. In this respect, the data in Fig. 8 are more consistent with the evidence in

Tables I to IV than are the data in Fig. 7. This is an encouraging fact.

Nevertheless, the problem of deciding on the exact form of a standard loudness scale for white noise can probably not be settled on the basis of our present knowledge. At any rate, my own impression is that there are still too many loose ends and inconsistencies. It appears, however, that at the low intensities the scale for noise will prove to be decidedly steeper than the scale for the 1000-cycle tone, and that at the higher intensities the loudness scale for noise will have approximately the same slope as the loudness scale for the 1000-cycle tone.

Plotted against a common scale of sound-pressure level, the two loudness scales will cross each other, probably in the vicinity of 30 or 40 db. This crossing point is one of the important items that need to be pinned down.

The principle mentioned above regarding the necessity of using both the tone and the noise as the variable in a loudness balance appears to have wide generality. An experiment on loudness balance should itself be balanced. This principle suggests that, in the determination of equal loudness contours, each of the frequencies to be compared should be made to serve as both the fixed stimulus and the variable stimulus. In measurements in which this balanced procedure has not been followed we have reason to expect that systematic errors may be present—errors that are presumably minimal near the middle of the loudness range and that increase as we go to very faint or very loud tones.

BISECTION EXPERIMENTS

Let us now return to the problem of the bisection (or equisection) of loudness intervals in order to see how well the experimental results agree with those of experiments on ratio determination. Two major studies of bisection have been made, one by Garner¹¹ and the other by me (unpublished). In both experiments the subject sat before a set of five keys which he pressed in order to produce the tones. The intensities of the tones produced by the two end keys were fixed, say, 40 db apart, and the subject adjusted the intensities produced by the intermediate keys in order to divide the 40-db interval into four equal-appearing steps in loudness.

Garner's subjects adjusted the intermediate tones by throwing switches that controlled motor-driven attenuators (2 sec/db). Before the beginning of a given run, all the variable tones were set at one or the other end of the 40-db range. In most of my experiments the subject simply turned attenuators (no visible dials) and before each run the attenuators were randomly positioned. My three attenuators had different sized steps: 1.5 db for the lower quarter point, 1 db for the midpoint, and 0.5 db for the upper (three quarter) point. Actually, I now believe it would have been wiser to have used sone potentiometers instead of decibel attenuators for this purpose.

It is a distressing fact that my results do not agree with those obtained by Garner. The intervals produced by Garner's 18 subjects are more nearly equal in *decibel size* than are the intervals produced by my 49 subjects. I cannot prove it, but I rather suspect that Garner's data are somewhat biased by the fact that the subjects could judge the time it took for the motor to drive the attenuator through the 40-db range (80 seconds) and that their fractional judgments were perhaps unconsciously influenced by the judgment of fractional values of this time interval. An additional, but slight, source of bias lies in the fact that Garner averaged decibels. The bisection experiment is one in which averaging by sones has a demonstrable advantage.⁴

In another experiment, I tried to check the validity of the results by having the experimenter reset the variable intensities to various levels and then requiring 25 subjects to estimate the relative spacing between the resulting loudnesses by means of a set of adjustable markers on a steel bar. The subject set the spacing of the markers to represent his perception of the spacing of the loudnesses. The data obtained agree quite precisely with the sone averages of the data obtained under the method of adjustment. The bar and markers constitute a rather useful device for this purpose.

One of the first things I discovered in these studies was the disconcerting fact that the point at which a subject bisects a 40-db interval depends on the order in which he listens to the tones. For example, when he hears them in the ascending order from faintest to loudest he may judge the midpoint to be about 32 db from the bottom of the range. When he hears them in descending order he may judge the midpoint to be about 27 db from the bottom. In an appendix to his main experiment, Garner verified this order effect, but in what order or orders his subjects pressed the keys in the main experiment is not known.

Because of this order effect, I found it necessary to control the order in which the stimuli were presented. For the part of the experiment that we shall consider here, the subject was required to use *both* the ascending and descending order and to try to adjust the loudnesses so that the intervals sounded equal in both directions.

TABLE V. Sample results obtained by subjects who set one tone to the midpoint of the loudness interval between two other tones. The predicted values were calculated on the assumption that the decibel change corresponding to a 2.1 loudness ratio is 10 db. All values are in decibels *re* standard reference level.

Interval bisected	Midpoint obtained	Midpoint predicted	Number of subjects
16-56	42.5	46.9	7
36-76	65.3	66.8	12
56-96	86.2	86.8	45
90-120	109.6	111.7	12
101-130	119.3	121.8	10
111.5140	130.6	131.8	12

¹¹ W. R. Garner, J. Acoust. Soc. Am. 26, 73-88 (1954).

It is possible, of course, to control the order of presentation of the tones, but it is not possible, unfortunately, to control the degree to which the subject pays attention to the two orders. He may or may not choose to ignore one of them.

Table V gives some of the results obtained, together with the values that would be predicted on the assumption that a loudness ratio of 2:1 corresponds to a decibel change of 10 db. It turns out that the midpoint to which the subjects adjusted the attenuator is slightly lower than the predicted midpoint. This same effect was found by Garner when he compared his bisection data with the fractionation data obtained from the same subjects.

Since bisection and ratio determination do not quite agree, we must either choose which method we will trust, or else reject them both. Garner decided to accept the verdict of bisection and to correct the ratio determinations accordingly. I would prefer to do the opposite. My mistrust of the bisection data is based on two considerations. One is the rather large differences obtained as a function of the order in which the tones are presented. The other is the old question¹² whether the subject is able to bisect an interval without being influenced by the ratios as well as the differences between the loudnesses. Does he set the middle tone b so that a-b=b-c, or does he set if so that a/b=b/c? Or, does he compromise between these two criteria and thereby produce a setting of b that is closer to the half-way point between a and c in decibels than it would otherwise be?

What looks like a tendency among some subjects to slip over from an interval toward a ratio judgment seems to be particularly strong when the lower tone of the interval is within about 20 db of threshold.¹³

In view of the difficulties that attend bisection judgments, the agreement shown in Table V between the obtained and the predicted bisections might be considered reasonably satisfactory. There is a similarly good agreement between the bisections we would predict and those recently reported by Black14 who used a noise with a spectrum that sloped 6 db per octave. Since Black's subjects listened to the noises only in ascending order, the midpoint obtained tends to be higher than the midpoint predicted. This is just what we should expect.

The data in Table V also demonstrate another important fact, namely, that loudness continues to grow as a power function of intensity up to at least 140 db. At this level the stimuli, under binaural listening, are truly painful. One of a pair of earphones burned out at this level! If, as some have supposed, loudness reaches a ceiling above which it ceases to grow, this upper limit must be beyond 140 db. It appears that over a range of

at least 140 db the loudness of a 1000-cycle tone grows as a power function of intensity.

POSSIBLE ENGINEERING RULE FOR MEASURING THE LOUDNESS OF NOISE

The determination of a scale to express the loudness iudgments of a typical listener confronted with a 1000cycle tone is only a first step toward the development of a procedure for measuring the loudness of sounds in general. In principle we can use the 1000-cycle tone as a vardstick and evaluate an unknown noise by finding the level of the 1000-cycle tone that matches the unknown noise in loudness. But such loudness balances between tone and noise are difficult to perform in the field. For one thing, in order to avoid systematic biases we need to make the noise as well as the tone serve as the variable stimulus.

What we need, therefore, is a reasonable, workable procedure for transforming sound level measurements into sones. The procedures¹⁵ now in vogue give fair results but they are not particularly simple. In addition to assuming a loudness function for the 1000-cycle tone that is significantly steeper than the present evidence indicates it ought to be, they make the further assumption that the total loudness of a wide-band noise can be found by adding together the loudnesses of certain of its component bands (octave bands or mel bands). This assumed additivity certainly needs a more thorough validation than it has vet received.

Perhaps a better way to approach the problem of the loudness of noise is first of all to break the problem into two separate parts. We may then ask (1) how does the loudness of noise depend on the over-all level, and (2) how does it depend on the frequency composition? It appears not unlikely that the first of these questions can be given a simple answer. As more and more results accumulate, I become increasingly convinced that all continuous noises of engineering interest follow to a good approximation the same simple rule. This rule says that, for all levels above about 50 db, the loudness Lchanges by a factor of two when the over-all level Nchanges by 10 db. In other words, for a constant spectrum and for N > 50 db

$\log L = 0.03N + S$

where S is the spectrum parameter.

The value of S will depend upon the make-up of the spectrum (including its phase relations) and will therefore need to be evaluated empirically. As we have seen, the value of S for a 1000-cycle tone is -1.2. In Figs. 7 and 8, we see that above 50 db a white noise is as loud as a 1000-cycle tone when the noise is roughly 10 db lower than the tone. Hence for white noise the value of the spectrum parameter S is approximately -0.9. We obtain this result by a formula that gives the value of S from the results of matching the 1000-cycle tone to

 ¹² S. S. Stevens and H. Davis, *Hearing, Its Psychology and Physiology* (John Wiley and Sons, Inc., New York, 1938).
¹⁸ Newman, Volkmann, and Stevens, Am. J. Psychol. 49, 134–

^{137 (1937).}

¹⁴ John W. Black, "Control of the sound pressure level of voice," Joint Report NM 001 104 500.42, U. S. Naval School of Aviation Medicine, Pensacola, Florida (February 15, 1955).

¹⁵ Beranek, Marshall, Cudworth, and Peterson, J. Acoust. Soc. Am. 23, 261–269 (1951); F. Mintz and F. G. Tyzzer, J. Acoust. Soc. Am. 24, 80–82 (1952).

the noise in loudness. If N_{1000} is the level of the tone in decibels and N_n is the level of the noise in decibels, the formula becomes $S=0.03(N_{1000}-N_n)-1.2$. The values of S and N will depend of course upon the characteristics of the meter. Provided, however, the meter is properly calibrated to read the sound-pressure level of the 1000-cycle tone, the loudness L computed by these formulas is invariant with meter characteristics (expect for certain nonlinearities).

In principle the values of S for other spectra can be obtained by a similar procedure: a loudness balance made against a 1000-cycle tone. The optimal level for making this loudness balance is in the vicinity of 70 db (the comfortable level for listening) because in this region the bias in loudness matching is minimal. By following this procedure we could readily determine the value of S for a variety of representative spectra of engineering interest.

Alternatively, we could perhaps work out a method for evaluating S from an octave- or a mel-band analysis of the noise. But we would first have to determine the proper rule for combining the loudnesses of the separate bands. This may prove more difficult than an evaluation of S for representative spectra by a direct loudness balance. In any case, I believe that the equation here proposed will lead to a more valid assessment of the loudness of noise than the procedures now in use.

For many purposes we do not need to know the value of S. In order to determine by what ratio we have changed the loudness of a noise we need to measure only the sound-pressure level before and after, provided the spectrum remains constant. For example, a 20-db reduction in the over-all level of a noise will decrease its loudness to 25% of the original value. Only when a significant change is made in the spectrum would we need to worry about S, and then we would face the empirical problem of evaluating S either by means of a loudness balance, or by a proper rule for the combining of loudnesses in separate frequency bands.

This scheme does not depend, of course, upon the use of the 1000-cycle tone as a reference standard, and it may ultimately prove better to choose as the reference sound a spectrum that can be more easily balanced against other noises. In the meantime, the potentiality of the simplified procedure based on the power-function relation between loudness and intensity invites serious consideration. The procedure has the advantage that it explicitly separates the problem of the dependence of loudness on intensity from the problem of its dependence on frequency composition.

As this paper goes to press there appears an interesting new study by Quietzsch¹⁶ which is relevant to our present concern. Quietzsch carried out experiments on the halving and doubling of loudness (1000 cps, 30 to 80 db) and on various procedures for measuring the loudness of different kinds of noise.

¹⁶ G. Quietzsch, Acustica 5, 49-66 (1955).

Over the range of levels explored the results from 21 subjects using the method of adjustment for halving and doubling give a loudness function very much like the one I have proposed. Above about 50 db the slope is approximately 10 db for a 2:1 loudness ratio, and below about 50 db the results for halving and doubling show the usual divergence: halving gives a steeper slope; doubling gives a flatter slope.

Quietzsch also studied 37 widely different types of noise by means of (1) loudness balances against a 1000cycle tone, (2) sound level measurements with two types of meters, and (3) measurements based on the summation of the loudness in separate octave bands. Among other things, Quietzsch proposes certain corrections to make the summation technique [procedure (3)] agree more closely with the results of direct loudness balance [procedure (1)].

From the results of Quietzsch's procedures (1) and (2) we can determine the value of what I have called the spectrum parameter S for his various noises and for his two types of meters. Unfortunately, in balancing the tone against the noise he varied only the tone, but despite this possible source of bias we learn many instructive facts from his experiments. Applying our formula for S, we note that when he uses a standard sound level meter (DIN-Lautstärkemesser) the value of S ranges from very near -1.2 (the value of S for the 1000-cycle tone) to about -0.5, with an average value of about -0.85. When he uses a particular peak-reading meter (Geräuschspannungsmesser) with a frequency characteristic resembling the human threshold, the value of S ranges from -1.3 to -0.74, with an average value of about -1.05. These figures suggest that the peak-reading meter has an advantage in that it indicates a more nearly invariant spectrum parameter.

Ideally, what we would like to have is a meter whose readings would always lead to a value of S=-1.2, which would mean that it would give the same reading for a 1000-cycle tone as for a noise, provided the noise and the tone were equally loud. If we were to limit our concern to levels above 50 or 60 db, could we combine in a single device a set of characteristics (frequency response, peak indication, time constant, etc.) that would allow us to measure the loudness level of noise with reasonable accuracy? This is an old question, I realize, but one that may need re-examination. If such a device could be constructed it would then be a simple matter to calibrate its scale to read directly in sones.

In the meantime it would seem advisable to use the equation $\log L = 0.03N + S$, and to evaluate S either by direct loudness balance or by an improved "summation method."

APPENDIX. DESCRIPTION OF THE EXPERIMENTS

Although Merkel¹⁷ (1889) appears to have been the first to perform an experiment in which the subject was

¹⁷ J. Merkel, Phil. Stud. 5, 499-557 (1889).

required to estimate the ratio between two loudnesses, it was not until the development of the modern electronic art that systematic studies could be made to determine the dependence of loudness on stimulus intensity. Merkel dropped balls from various heights onto a sounding block in order to produce different sound intensities.

Richardson and Ross¹⁸ (1930) used electrical equipment to produce tones for this purpose. The observer heard tones of two different intensities, and he was required to rate the loudness of one of the tones as a fraction or multiple of the loudness of the other. Although their results were not reported in acoustical quantities, from their stated values of telephone current it can be determined that the average reduction in intensity needed to reduce the loudness by a factor of two was about 12 db. This is the value entered in Table I.

Laird, Taylor, and Wille¹⁹ (1932) were among the first to make fractional estimates of loudness over a wide range of intensities. For their tonal stimuli they used a 2-A audiometer which provides stimulus intensities in 5-db steps. Their procedure was a simplified method of constant stimuli. Unfortunately, their results are so far out of line with those of most other experimenters that their data have been deliberately ignored by those who have tried to construct loudness scales. An attempt was made by Stevens and Herrnstein to repeat these measurements with a 2-A audiometer, using presumably the same procedure. Their results, shown in Table I, are quite at variance with those of Laird, Taylor, and Wille. Later on, Stevens, Rogers, and Herrnstein²⁰ tried deliberately to alter the subjects' judgments in the direction of larger decibel reductions for half loudness. They presented comparison tones at fainter levels than those used by Stevens and Herrnstein and they altered the schedule by which the successive comparison tones were chosen. This procedure was fairly successful with subjects who had never judged loudness before, but it made much less difference to subjects who had served previously in a loudness experiment. Apparently, by presenting first a standard and then a batch of properly chosen comparison stimuli well below the standard, the subject can be induced to choose one of these low values as being half as loud. Garner²¹ has also demonstrated this effect. The comparison tones presented by Stevens and Herrnstein were limited to those 5, 10, 15, 20, and 25 db below the standard, because these reductions were all that turned out to be necessary. Of course, if the subject had said that a 25-db reduction was louder than half, the comparison would have been lowered still further. In the later experiment the range of comparison levels was extended to 40 db below that of the standard.

Ham and Parkinson²² produced tones by means of a loudspeaker and had groups of untrained subjects (college students) estimate the percentage of the loudness that remained after a tone had been reduced by a given number of decibels. This is one form of the method we have called magnitude estimation (ME). From the curves plotted from these percentage estimates it is possible to determine the decibel reduction required for half loudness. For this interpolation procedure, the logarithm of the estimate was plotted against decibels. The resulting plot is reasonably rectilinear, and the fitting of a curve to the data is not difficult. This curve was used to estimate the half loudness and the quarter loudness reductions. Some arbitrary judgment is necessarily involved in this process, but it probably provides a fair estimate of the half-loudness point.

In the experiments of Ham and Parkinson some degree of bias was probably introduced by the fact that the comparison stimuli were always presented in descending order, instead of in a random order.

Ham and Parkinson also used another method to determine the decibel difference for ratios of 2:1-a method of constant stimuli (CS). The subject was given a standard and eight comparison levels and was required to pick the one of the eight whose loudness was nearest to half (or double) loudness. They also used ratios of $\frac{1}{3}$, $\frac{1}{5}$, 3, and 5 and obtained judgments thoroughly consistent with those obtained for $\frac{1}{2}$ and 2. The judgments for half and double loudness are entered directly in the tables. An interpolation between the values for $\frac{1}{3}$ and $\frac{1}{5}$ and for 3 and 5 has been made to give a value for $\frac{1}{4}$ and for 4, from which additional estimates of $\frac{1}{2}$ and 2 have been obtained.

Geiger and Firestone⁵ used a method of adjustment in which the subjects were allowed to set the levels of the comparison tones by means of a potentiometer (characteristic not stated). Their 44 subjects set the comparison tone to various fractions and multiples: $\frac{1}{2}$, $\frac{1}{4}, \frac{1}{10}, 1/100, 2, 4, 10, 100$. They computed both means and medians for their data, so that we have here a chance to compare these two measures. Except when the adjustments are being made to the more extreme ratios, the medians are nearly always less than the means.

Churcher, King, and Davies²³ allowed their subjects to adjust an attenuator to reduce the comparison tone to half loudness. Having found the half-loudness point, the subject proceeded from there to make a second fractionation, and so on. Thus a curve was obtained for each subject, and finally the decibel values corresponding to the various fractions were averaged. They repeated the procedure for judgments of $\frac{1}{4}$ loudness and obtained quite consistent results.

Rschevkin and Rabinovitch²⁴ used tones lasting only

827

¹⁸ L. F. Richardson and J. S. Ross, J. Gen. Psychol. 3, 288-306 (1930).

¹⁹ Laird, Taylor, and Wille, Jr., J. Acoust. Soc. Am. 3, 393-401 (1932). 20 Stevens, Rogers, and Herrnstein, J. Acoust. Soc. Am. 27,

^{326-328 (1955).} ²¹ W. R. Garner, J. Exptl. Psychol. 48, 218-224 (1954).

²² L. B. Ham and J. S. Parkinson, J. Acoust. Soc. Am. 3, 511-534 (1932).

²³ Churcher, King, and Davies, J. Inst. Elec. Engr. (London) 75, 401-446 (1934)

S. N. Rschevkin and A. V. Rabinovitch, Rev. Acoust. 5, 183-200 (1936).

0.4 second and determined the ratios $\frac{1}{2}$, $\frac{1}{4}$, 2, and 4. They argue that the use of longer tones fatigues the ear and leads to spurious results. Otherwise their procedure was not unlike that of Laird, Taylor, and Wille. They say that their results agree with those of Laird, Taylor, and Wille, but this seems hardly to be the case.

When Rschevkin and Rabinovitch lengthened the tones to 4 seconds, the average reduction needed to produce half loudness decreased from 11.0 db to 9.5 db. But since the authors prefer the shorter presentation, the values recorded in Table I are the larger ones they obtained with the shorter tones.

Pollack²⁵ used a method of adjustment with a relatively small number of subjects. His are the smallest values yet reported for the decibel reduction required for half loudness. Since it is difficult to read values from the graphs published in his paper, the values entered in the tables are those kindly supplied me by Dr. Pollack.

Garner²⁶ used the method of adjustment to obtain 20 settings at each level by each of 18 subjects. His subsubjects adjusted a decibel attenuator (1-db steps). His is the most thorough statistical analysis of an experiment of this sort. Unfortunately, the data, as he says, "disagree markedly with data previously reported," and we are left wondering why. His second experiment,¹¹ employing 18 other subjects who made 10 settings at each level, also used the method of adjustment, but here the subject threw a switch to control a motor-driven attenuator which moved at the rate of 1 db every 2 seconds. In this second experiment the decibel reductions required for half loudness are smaller at the low levels and larger at the high levels than in the first experiment. The data recorded in Table I were kindly supplied me by Dr. Garner, since they were not tabled in his paper.

The large decibel reductions reported by Garner are probably due to a number of factors. His subjects adjusted a decibel attenuator instead of a sone potentiometer whose attenuation, as a function of angular turn, is more nearly proportional to loudness. And instead of averaging in sones, or computing a median, Garner averaged the decibel values themselves. If he had used a sone potentiometer and averaged in sones, the decibel reductions would probably have turned out smaller. Also, the fact that the subjects worked for fairly long periods at a stretch probably accentuated the distortion due to whatever biasing factors were present.

Robinson's⁸ experiment was perhaps the most thorough study yet made with the method of constant stimuli. The stimulus was presented by a loudspeaker to a subject seated 1 meter in front of it. Tones and noises were used, and results were obtained for the ratios $\frac{1}{2}$, $\frac{1}{10}$, 2, and 10. These results are internally quite consistent, except for the usual discrepancy between the data for halving and doubling at low loudness levels.

²⁵ I. Pollack, J. Acoust. Soc. Am. 23, 654–657 (1951).
²⁶ W. R. Garner, J. Acoust. Soc. Am. 24, 153–157 (1952).

A possible systematic bias may be present owing to the fact that the spacing of the comparison stimuli was in decibel steps instead of in equal-loudness steps. And since it has been demonstrated that the outcome of the method of constant stimuli is particularly sensitive to the range of the comparison stimuli presented, we must consider the possibility that this "context" factor may have played a role in Robinson's experiment.

The results of certain experiments, either finished or in progress at the Psycho-Acoustic Laboratory, are included in the tables as P.A.L. entries. These experiments have been undertaken to explore the effects of various procedures on loudness judgments and the results obtained have been incidental to this main purpose. Nevertheless, these results need to be considered in any decision relating to the choice of a loudness scale.

The results of Stevens and Herrnstein with the 2-A audiometer have already been referred to.

The experiment by Stevens⁹ using magnitude estimation was an early attempt to see how far it might be possible to go toward getting the subjects to make absolute numerical judgments of loudness. Three standards were used: 70, 100, and 120 db sound-pressure level (binaural earphone listening). The standard was presented only once and then 8 to 10 comparison tones spaced 5 db apart were presented in random order. The subject was asked simply, "If the standard were called 100, what would you call each of the comparison tones?" This procedure was repeated twice at each level. The logarithms of the median estimates give fairly rectilinear functions when plotted against decibels and from these plots the $\frac{1}{2}$ and $\frac{1}{4}$ loudness points have been determined.

The experiment of J. C. Stevens (reported in reference 9) with white noise was similar to the foregoing except that the standard was presented before each of the comparison stimuli. In part of this experiment the stimuli were spaced 5 db apart below the standard; in another part they were spaced more nearly proportional to equal-loudness intervals. These different spacings made no significant difference in the resulting magnitude estimations (Table III).

Another variation involved the use of a faint standard (45 db) which was assigned the value 1. The subjects then assigned numbers to a random series of louder comparison noises (Table IV). The functions determined by this ascending procedure are not significantly different (in the middle of the intensity range) from the functions obtained when the standard is called 100. Other evidence indicates that at the extremes of the intensity range these two procedures give different results—in the direction one would predict. At high intensities the ascending procedure (standard called 1) gives a steeper loudness function than does the descending procedure (standard called 100). The reverse is true for low intensities.

In the experiment by Stevens and Poulton³ using 33

828

subjects, an attempt was made to see what people who had never before made quantitative loudness judgments would do when asked to adjust one tone to a fraction of the loudness of a standard (90 db). The control used was a ganged potentiometer, which reduced the level about 15 db when turned from full-on to half-on and about 12 db when turned from half-on to quarter-on. In other words, it was made to approximate a sone potentiometer. The level was so adjusted that when this potentiometer was full on, the comparison tone was at 94 db, i.e., 4 db more than the standard.

On their first attempts, eleven subjects set to $\frac{1}{4}$ loudness, eleven to $\frac{1}{2}$ loudness, and eleven to $\frac{3}{4}$ loudness. Those who set to $\frac{1}{4}$ and $\frac{3}{4}$ then went on to set to the other fractions. These other settings were quite consistent with the first. On the first judgments the mean reductions were: $\frac{3}{4} = 4.7$ db; $\frac{1}{2} = 9.3$ db; $\frac{1}{4} = 19.0$ db.

In Table I are entered the decibel means, the sone means, and the medians of the half-loudness judgments and of half the values of the quarter-loudness judgments.

In this same experiment a study was made of the influence of the characteristics of the control device used by the subjects. One group of eleven subjects made their first settings with a decibel attenuator and, a week later, switched to a sone potentiometer. Another group of eleven subjects did the reverse. For both groups combined, the effect of using the sone potentiometer was to reduce the decibel difference required for half loudness by 2.7 db.

Using the method of magnitude estimation, Poulton did a variety of experiments. The purpose of one of them (reported in reference 9) was to explore the effect of the particular numbers used by the subjects to report their judgments. Eight experienced subjects sometimes were told that the standard (100 db) was 100 and that the comparisons were to be judged accordingly. Sometimes they were told the standard was 1 and that they were to say what fractional part of 1 seemed to represent the comparison tone. These two procedures gave very similar results, as can be seen from the interpolated estimates of half loudness and quarter loudness entered in Table I. These subjects also rated louder tones relative to a standard (60 db) called 1, and these results were used to estimate twice and four times loudness (Table II).

In another test with magnitude estimation, 32 inexperienced subjects heard a 100-db standard which was called 100. Groups of 8 subjects each judged, as their first attempt ever made, a different comparison level: 6, 10, 20, or 40 db below the standard. The medians of these first estimates were 62.5, 50.0, 29.5, and 12.0, respectively. From the pooled estimates of the group ratings on all seven comparison levels ultimately used, half-loudness and quarter-loudness values were determined, as shown in Table I. Also shown are values for similar tests with a standard of 80 db. Since these latter values were out of line with expectations, Poulton satisfied his curiosity a month later by repeating the test on eight of the subjects. Seven of the eight then gave lower estimates, but these have not been entered in Table I. What went wrong the first time is still a mystery—as it often remains in these experiments.

The same 32 subjects also made multiple estimates relative to standards (40 and 60 db) which were called 1. The results are shown in Table II.

Poulton and Stevens³ had a group of 16 previously unpracticed subjects adjust the sone potentiometer so that the loudness of a white noise was half and twice that of various standards. The means and medians of the results are shown in Tables III and IV. (Twenty other subjects were later added to the group making 36 subjects in all.)

A word should perhaps be said about another P.A.L. experiment that is only about half finished. It began as a rather ambitious project, employing the method of adjustment for the ratios $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{10}$, 1/100, 2, 4, 10, 100, and also the method of magnitude estimation both ascending and descending. Sixteen subjects were to go through this array of tasks in different orders, all counterbalanced in accordance with the precepts of modern statistical design. It looked nice on paper, but we have run into two difficulties. We are finding in practice that a few of the subjects "can't take it." They are required to shift their criterion so drastically and so often that consistency becomes almost impossible. Some of them complain of losing confidence, and their results show it. The second difficulty has arisen from the fact that we unwisely set some of the levels so that the subject had to use the extreme low end of the sone potentiometer where a very small turn of the knob makes a large change in the loudness ratio. Apparently, if we are to determine how loud different tones sound to people, we must be careful to set for them simple, unconfusing tasks, and we must stop our experiments before the listeners become uncritical in their judgments.