Perception & Psychophysics
1974, Vol. 15, No. 2, 238-242

Nonmetric scaling of loudness and pitch using similarity and difference estimates*

SCOTT PARKER and BRUCE SCHNEIDER
Columbia University, New York, New York 10027

In Experiment 1, nonmetric analyses of estimates of similarity and difference were used to generate a scale of loudness for 1,200-Hz tones varying in intensity. For both similarity and difference estimates, loudness was found to grow approximately as the 0.26 power of sound pressure. In Experiment 2, nonmetric analyses of estimates of similarity and difference were used to generate a scale of pitch for 83.3-dB pure tones varying in frequency. For both similarity and difference estimates, pitch was found to vary with frequency in accordance with the mel scale.

To develop subjective scales of loudness and pitch, most investigators rely on the ratio and interval scaling procedures developed by S. S. Stevens (1958). In these procedures, the S is required, in one fashion or another, to directly report ratios or intervals of sensory magnitude. These judgments are used to construct scales of the amount of a given attribute. They typically show that sensory magnitude is a power function of sound pressure and that pitch increases nonlinearly with tone frequency.

There is reason, however, to be somewhat concerned about the validity of scales established by direct estimation procedures. In a typical ratio estimation experiment, Ss are asked to assign numbers to stimuli varying along a single dimension such that the ratios among the numbers reflect the ratios of magnitudes among the stimuli. To argue from these experiments that psychological magnitude is a power function of intensity, it is necessary to assume that these numbers are multiplicatively related to psychological magnitudes. Unequivocal determination of the function relating sensory magnitude to stimulus intensity requires an experiment designed to determine a scale of sensory magnitude in which much weaker assumptions about the nature of the sensory judgments are employed. Recent work by Shepard (1962a, b) on nonmetric scaling techniques supplies the basis for such experiments.

In the experiments reported below, scales of loudness and pitch were developed using nonmetric scaling techniques. These techniques require that the S order tonal pairs with respect to how much the elements of a pair differ with respect to some attribute. In Experiment 1, the elements of a tonal pair were identical in frequency (1,200 Hz) and differed only in intensity. Ss therefore ordered loudness differences. In Experiment 2, the elements of a tonal pair were identical in intensity (83.3 dB re 0.0002 dynes/cm^2) and differed only in frequency. Ss in Experiment 2 ordered pitch differences. Two different methods were employed to obtain a rank ordering of the loudness and pitch differences. One group of Ss in each experiment was asked to estimate magnitudes of differences in loudness (Experiment 1) or pitch (Experiment 2). A second group of Ss was asked to estimate magnitudes of similarities between the elements of a pair. Difference estimates were interpreted as distances along a psychological continuum, and similarity estimates were interpreted as proximities (order-inverse with distance). The rank orders of the difference and similarity judgments were used to determine interval scales of the sensory attributes in question (loudness or pitch). A concordance of the results for difference and similarity judgments would suggest that the geometric representation for the stimuli is in fact correct, since it can be obtained via these different experimental tasks. (A discordance of results might mean that the experimental approach used here is unstable and unreliable, or it might mean that instructions to consider differences induce perceptual structures unlike those induced by instructions to consider similarity.) The interval scale representations of loudness and pitch obtained from these judgments of difference and similarity were compared to tonal intensity and frequency. In particular, two objectives of the study were to determine (1) whether Stevens's power law or Fechner's logarithmic law was a better description of the loudness function, and (2) whether Stevens and Volkmann's (1940) mel scale was appropriate for pitch.

METHOD

Experiments 1 and 2

Subjects

Nineteen of the 20 Ss were Columbia University undergraduates or graduate students in Columbia's Department

*These data are based on a dissertation by the first author, conducted under the supervision of the second author, and submitted to the Department of Psychology, Columbia University, in partial fulfillment of the requirements for the PhD degree. We would like to thank Donald Hood and Eugene Galanter for their helpful comments and suggestions. We would also like to thank Dr. Galanter for making available to us the facilities of the Psychophysics Laboratory. This research was supported by NSF Grant GB 36211. Reprints may be obtained from Scott Parker, Department of Psychology, Columbia University, New York, New York 10027.

Table 1
Stimuli Used in Experiments 1 and 2

Experiment 1 (dB re 0002 dynes/cm²)	Experiment 2 (Hz)
50	460
56*	525
60	645
68	760*
72	830
80	920
86*	1060
94	1130*
98	1290
104	1370

Note–All stimuli in Experiment 1 are at 1200 Hz. All stimuli in Experiment 2 are 83.3 dB re .0002 dynes/cm².

of Psychology. One S was a graduate student at the University of Pennsylvania Department of Linguistics. Their ages ranged from 18 to 28 years. All Ss claimed to have normal hearing. Eight of the Ss had had some musical training. Seven had previously served as Ss in magnitude estimation experiments. None was paid for participation.

Apparatus

Calibrations and listening conditions were identical to those used by Carvellas and Schneider (1972), except that the Ss sat in an Industrial Acoustics sound-resistant booth, Model 300.

Procedure

The 10 tones used in each experiment are listed in Table 1. Each S served in three experimental sessions. In the first session, the Ss were presented with the 45 pairs of unequal tones constructible from the set of 10 tones. In the second and third sessions, the 45 tone pairs were presented twice, with a 10-min break separating the presentations. In Experiment 1, five Ss estimated the loudness difference of the tones in each pair, and five Ss estimated the loudness similarity of the tones in each pair. In Experiment 2, five Ss estimated pitch similarity and five Ss estimated pitch difference.

Prior to hearing the first-session tone-pair sequence, the Ss estimating difference in Experiment 1 (loudness difference) were instructed as follows: "This is an experiment on your perception of difference. You will hear pairs of tones. The tones in a pair will differ in loudness. Your task is to decide how different the tones in a pair are and to assign a number to that difference. You will first hear a pair of tones whose difference we will assign the number 60, to give us a starting point. For any subsequent pair, if the tones in that pair sound twice as different as did those in the first pair, assign it the number 120; if in some pair, the tones sound half as different as did those in the first pair, assign it the number 30. You may use any positive number–integer, fraction, or decimal. You may not use negative numbers or zero. Are there any questions?"

The instructions were similarly constructed for all Ss, except that they concerned pitch rather than loudness in Experiment 2, and concerned similarity rather than difference for half the Ss in each experiment. Also, Ss estimating similarity had the first pair designated as having a similarity of 20 rather than a difference of 60.

Ss estimating similarity commonly claimed that they did not understand what they were to do. They were then told to assign a number to "how much the tones sound alike."

The S was then led into the booth and shown how to wear the earphones and how to operate the three-position switch and the intercom. Two minutes later, the standard pair (tones marked by * in Table 1) was presented and identified. When the S was satisfied with that pair, the 45-tone pair sequence was presented. In the second and third sessions, no instructions were given and each tone-pair sequence was preceded by an identified presentation of the standard pair. Ss listened to each pair as long as they wished and then spoke a number.

The sequences in the second and third sessions were such that each tone pair appeared twice before and twice after each other tone pair. Also, each tone appeared equally often on each operative position of the three-position switch. The second sequence of the third session was identical to that used in the first session.

RESULTS

Experiment 1–Loudness

For both similarity and difference estimation, each S's first estimate was discarded. The geometric mean of the remaining four estimates was computed for each of the 45 stimulus pairs. The geometric mean was chosen as a measure of central tendency since the variance of magnitude estimates generally increases with the mean. The geometric means were then ranked, within Ss, from 1 to 45. Kendall's coefficient of concordance, W (Siegel, 1956, pp. 229-238), for these rank orders was found to be 0.93 for the five Ss estimating loudness similarity and 0.94 for the five Ss estimating loudness difference. Thus, agreement among Ss as to rank order of loudness similarity was good, and agreement on rank order of loudness difference was equally good.

For the similarity estimates and the difference estimates, the arithmetic mean of the ranks across the five Ss was computed for each stimulus pair. These mean ranks were then themselves ranked from 1 to 45, providing an ordinal index of each group's similarity or difference estimates. These ranks were reversed ($R' = 46 - R$) for the similarity group, but not for the difference group, and the ranks were used as input to a nonmetric scaling computer program (Carvellas & Schneider, 1972). The use of these averaged ranks assumes that geometric distance is monotone increasing with increasing loudness difference estimates, and monotone decreasing with increasing similarity estimates (since these ranks were reversed). Since 1,200-Hz tones vary minimally in pitch over the range of intensities used (Stevens, 1935), the basis for all judgments was presumed to be loudness variation among the tones. Hence, the analysis was one-dimensional, and the tone intensities in decibels were used as the starting configuration in both cases.

Stress, Kruskal's (1964) measure of goodness of fit, was computed for the outputs of the nonmetric program. Stress measures the discordance between the predicted distances, ds, and a set of distances, $\hat{d}$, that are (1) monotonically related to the original distances (i.e., preserve the rank-ordering) and (2) are as much like the ds as they can be within the restrictions imposed by (1). Stress is given by $[\Sigma(d - \hat{d})^2/\Sigma d^2]^{1/2}$, often expressed as a percentage. Notice that perfect ordinal agreement produces $d = \hat{d}$, and, in this case, stress = 0. The stress values were 4.7% for difference estimates and 3.1% for similarity estimates. Kruskal (1964) states that stress values of 5% or less indicate "good" agreement between the ds and $\hat{d}$s.

The index of metric determinacy (M), which was originally developed by Shepard, was estimated from Young's (1970) nomogram. M is the squared Pearson correlation coefficient between the true distances (whose rank ordering serves as the input to the algorithm) and the ds produced by the algorithm. Hence, M varies between 0 and 1, and M = 1 means that the true distances have been perfectly reconstructed. In no empirical investigation using these techniques are the true distances known, but Young provides a nomogram for estimating M from the number of points, number of dimensions, and stress–all of which are available. The result is that, in nonmetric scaling analysis, if M is sufficiently high (above .98, say), thepoint coordinates produced by the algorithm are properly regarded as an interval scale representation of the original points. In these experiments, M was, conservatively, 0.98 for the difference estimates and 0.99 for the similarity estimates. Thus, for both experiments, the projection values achieved from the nonmetric program may be taken as an interval scale representation of stimulus loudness.[1]

One objective of this study was to determine whether loudness was related to sound pressure as described by Fechner's law, Stevens's law, or something else entirely. Since the projection values, P_i, provide interval scale representation for loudness, $L_i = aP_i + b$, where L_i is the loudness of Stimulus i, P_i is the projection value for Stimulus i (from the nonmetric scaling program), and a and b are constants. If loudness is as described by Fechner's law, L is linear with log I, where I is stimulus sound pressure. Since decibels are a logarithmic transform of sound pressures, Fechner's law states L is linear with stimulus intensity in decibels. Hence, P is linearly related to stimulus intensity in decibels, if Fechner's law is correct.

Stevens's law is $L_i = kI_i^n$. Thus, the present data follow Stevens's law if $aP_i + b = kI_i^n$. Equivalently, $P_i + (b/a) = (k/a)I_i^n$, or $P_i + b' = k'I_i^n$, where b′ is unknown. Taking logarithms on both sides of the equation, Stevens's law describes thepresent data if there is a b′ such that $\log(P_i + b') = n\log(I_i) + \log k'$, i.e., if for some b′, $\log(P_i + b')$ is linear with stimulus intensity in decibels. A value of b′ was therefore sought that would increase the squared correlation coefficient (r^2) between $\log(P_i + b')$ and decibels. The values of r^2 for numerous choices of b′ were computed, and b* designates that b′ which maximized r^2. The numbers $P_i^* = 100(P_i + b^*)$ are referred to as adjusted loudness projections.

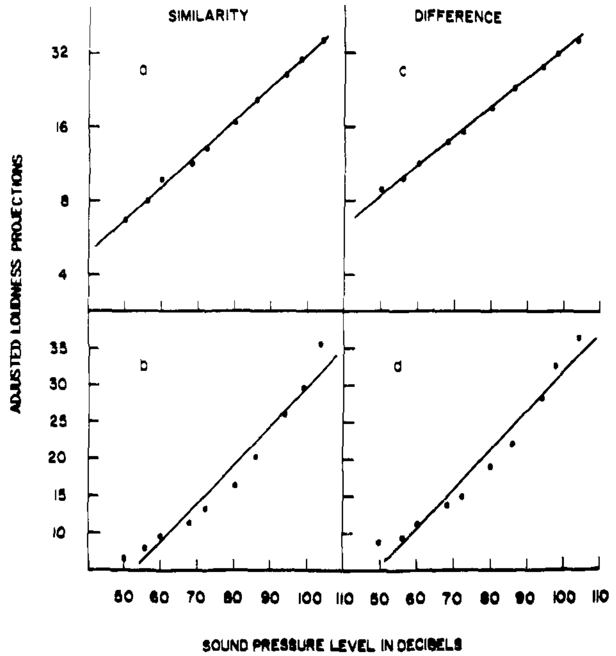

Fig. 1. Adjusted loudness projections (see text) as a function of stimulus intensity in decibels. Notice that the ordinate is spaced arithmetically in Panels b and d, and logarithmically in Panels a and c.

Figure 1 shows plots of adjusted loudness projections vs stimulus intensity in decibels for both loudness similarity estimates and loudness difference estimates. The ordinate is spaced logarithmically in Panels a and c and arithmetically in Panels b and d. Best-fitting straight lines, determined by the method of least squares, are drawn in each panel. It can be seen that there are no systematic deviations from linearity in the upper panels, whereas the point configurations in the lower panels appear to be concave upwards. Linearity in the upper panels indicates conformity of the data with Stevens's law; linearity in the lower panels would indicate conformity of the data with Fechner's law. Values of r^2 are 0.998 and 0.997 for Panels a and c, and 0.949 and 0.963 for Panels b and d. These data, then, are not in accord with Fechner's law, but do conform to Stevens's law. The best estimated for n in the formulation $P_i^* = kI_i^n$ are 0.27 for similarity estimation and 0.24 for difference estimation.

Experiment 2–Pitch

As in Experiment 1, each S's first estimate was discarded. The geometric mean of the remaining four estimates was computed for each of the 45 stimulus pairs. The geometric means were ranked within Ss from 1 to 45. Kendall's coefficient of concordance for these rank orderings was 0.92 for pitch similarity estimates and 0.95 for difference estimates. Here again, the Ss agreed both on the rank order of pitch similarities and on the rank order of pitch differences.

For the similarity estimates and for the difference estimates, the arithmetic mean of the ranks across the five Ss was computed for each stimulus pair. These mean

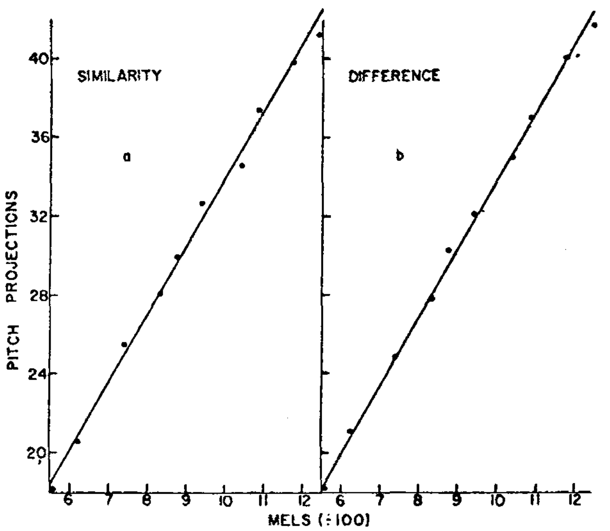

Fig. 2. Pitch projections (see text) as a function of stimulus pitch in mels/100.

ranks were then themselves ranked from 1 to 45. The ranks were reversed for the similarity estimates but not for the difference estimates, and these ranks were used as input to the nonmetric analysis. Since 83.3-dB tones vary minimally in loudness over the range of frequencies used (Fletcher & Munson, 1933), the basis for all judgments was presumed to be pitch variation among the tones. Hence, the analysis was one-dimensional, and the frequencies of the tones in hertz were used as the starting configuration. The projections, P_i, obtained from the program were multiplied by 100, $P_i^* = P_i \times 100$. The P_i^* are referred to as pitch projections.

Stress was computed for the output of the program. The value of stress for the similarity estimates was 6.0%; that for the difference estimates was 3.5%. Thus, agreement between the ds and $\hat{d}$s is good for the difference estimates, and fair for the similarity estimates [Kruskal (1964) states that a "fair" fit is indicated by a stress between 5% and 10%]. Estimates of M, the index of metric determinacy, were made from Young's (1970) nomograms. M was, conservatively, .98 for the similarity estimates and .99 for the difference estimates. Thus, the P_i in both cases may be taken as an interval scale representation of stimulus pitch.[2]

Figure 2 shows a plot of pitch projections vs stimulus pitch in mels/100 (Stevens & Volkmann, 1940, p. 336) for both the similarity and difference estimates. Best-fitting straight lines, determined by the method of least squares, are drawn in each panel. There are no systematic deviations from linearity in either panel. Values of r^2 for these plots are 0.992 for the similarity estimates and 0.996 for the difference estimates. Thus, agreement is quite good between the pitch projections and the mel scale.

DISCUSSION

The high inter-S agreement, low stress values, and high (estimated) M values indicate that interval-scale representations for loudness and pitch can be obtained from perceptual-interval scaling experiments in which Ss' responses are treated as ordinal proximity indices. In both experiments (loudness and pitch), the interval scale representations were essentially identical for two distinct proximity indices (similarity and difference).

The representation found for the tones varying in intensity (Experiment 1) is a validation of the notion that psychological loudness is a power function of stimulus intensity rather than a logarithmic one; i.e., that loudness is given by Stevens's rather than by Fechner's law. Rule et al (1970), using the rank-order properties of judgments of area difference for pairs of circles, also found that their scale values were a power function of the actual areas of the circles. Similar results were obtained from lifted weights in the same experiment. The Rule et al study and the loudness data of the present one indicate that for judgments of perceptual difference, a power function representation of the stimuli is often appropriate. Furthermore, to obtain this representation, one need only assume that Ss' magnitude estimates are monotonic with perceptual differences. This is much weaker than the assumption employed in direct estimation techniques: namely, that the numbers generated by the Ss are proportional to sensory magnitudes. The fact, however, that the nonmetric techniques also result in a power function representation of sensory magnitude lends support to the assumptions underlying the direct estimation techniques.

In the present experiment, both the similarity and the difference estimates produced spatial representations that were essentially identical power functions of stimulus intensity (exponents of 0.27 and 0.24 for similarity and difference, respectively). This suggests that the perceptual structures underlying judgments of loudness difference, loudness similarity, and direct estimates of loudness are all power functions of stimulus intensity. And, with respect to pitch, the perceptual representation is identical for all three kinds of judgments. A study by Markley et al (1969), however, suggests that this convergence to a single representation for these three kinds of judgments may not hold for all sensory continua. They had Ss rate the similarity of pairs of lines, and used the rank-order properties of these similarity judgments to determine a scale of line length. They found that their scale values were a logarithmic rather than a power function of actual line length. Thus, in this instance, a different perceptual structure appears to underlie judgments of line length similarity as compared to direct estimations of line length where a power function representation holds (exponent close to 1.0). The reasons why similarity, difference, and direct estimates converge on the same representation for loudness and pitch, but apparently not for line length, remain to be determined.

The exponents of the best-fitting power functions (0.27 and 0.24 for similarity and difference, respectively) are approximately equal to each other, but are at variance with previously reported values obtained from straightforward magnitude estimation experiments, where the exponent for monaural loudness is approximately 0.54 (e.g., Reynolds & Stevens, 1960). The sources of the discrepancy in exponent between the present results and those of the more traditional metric scaling methods are not clear.

However, there are some results on loudness that are in accord with those of the present experiment. Garner (1954), combining equisection and fractionation data for 1,000-Hz tones ranging from 50 to 110 dB re .0002 dynes/cm^2 developed the lambda scale for monaural loudness. Loudness in lambda units grows as the 0.26 power of sound pressure. Garner shows, in addition, that the lambda scale provides a better account of previous bisection and equisection experiments than does Stevens's (1956) sone scale. The data of Beck and Shaw (1961), who also worked with loudness differences, agreed more closely with the lambda scale than with the sone scale.

The results of Experiment 2 (pitch) are in accord with Stevens and Volkmann's (1940) revised mel scale. The recovery of the mel scale is also in agreement with the results of Carvellas and Schneider (1972). These experiments all involve the Ss' responding on the basis of pitch intervals.

An interesting aspect of the results of Experiment 2 (pitch) is the absence of any phenomenon like octave generalization. Two pairs of tones in the stimulus array had a frequency ratio of 2:1 (460 and 920 Hz : 645 and 1,290 Hz), and three of the Ss (two estimating difference and one estimating similarity) affirmed upon inquiry that they had heard tone pairs with octave separation. Yet, in neither the similarity nor the difference portion of the experiment was there any suggestion in the data that octaves sounded alike for these Ss. Octave generalization has been found in conditioning studies with rats (Blackwell & Schlosberg, 1943), pigeons (Gerry, 1971), and humans (Humphreys, 1939; Bersh, Notterman, & Schoenfeld, 1956). There may well be some set of instructions for experiments of the present type that would produce evidence of octave similarity, but it is disheartening that "how much the tones sound alike" is inadequate to the task.

REFERENCES

Beck, J., & Shaw, W. A. Ratio-estimations of loudness intervals. American Journal of Psychology, 1967, 80, 59-65.

Bersh, P., Notterman, J., & Schoenfeld, W. N. Generalization to varying tone frequencies as a function of intensity of unconditional stimulus. Publication 56-79, School of Aviation Medicine, USAF, Randolph AFB, Texas, 1956.

Blackwell, H. R., & Schlosberg, H. Octave generalization, pitch discrimination, and loudness thresholds in the white rat. Journal of Experimental Psychology, 1943, 33, 407-419.

Carvellas, T., & Schneider, B. Direct estimation of multidimensional tonal dissimilarity. Journal of the Acoustical Society of America, 1972, 51, 1839-1848.

Fletcher, H., & Munson, W. A. Loudness, its definition, measurement, and calculation. Journal of the Acoustical Society of America, 1933, 5, 82-108.

Garner, W. R. A technique and a scale for loudness measurement. Journal of the Acoustical Society of America, 1954, 26, 73-88.

Gerry, J. E. Peak shift on the tonal frequency continuum: The effects of extinction and punishment. Psychonomic Science, 1971, 23, 33-34.

Humphreys, L. G. Generalization as a function of method of reinforcement. Journal of Experimental Psychology, 1939, 25, 361-372.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 1964, 29, 1-27.

Markley, R. P., Ayers, D., & Rule, S. J. Similarity judgments of line length. Perception & Psychophysics, 1969, 6, 58-60.

Reynolds, G. S., & Stevens, S. S. Binaural summation of loudness. Journal of the Acoustical Society of America, 1960, 32, 1337-1344.

Rule, S. J., Curtis, D. W., & Markley, R. P. Input and output transformations from magnitude estimation. Journal of Experimental Psychology, 1970, 86, 343-349.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika, 1962a, 27, 125-140.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. Psychometrika, 1962b, 27, 219-246.

Siegel, S. *Non-parametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956. Pp. 229-238.

Stevens, S. S. The relation of pitch to intensity. Journal of the Acoustical Society of America, 1935, 6, 150-154.

Stevens, S. S. The direct estimation of sensory magnitude-loudness. American Journal of Psychology, 1956, 69, 1-25.

Stevens, S. S. Problems and methods of psychophysics. Psychological Bulletin, 1958, 54, 177-196.

Stevens, S. S., & Volkmann, J. The relation of pitch to frequency: A revised scale. American Journal of Psychology, 1940, 53, 329-353.

Young, F. W. Nonmetric multidimensional scaling: Recovery of metric information. Psychometrika, 1970, 35, 455-473.

NOTES

1. The data on difference estimation were reanalyzed using Young's (1968) Torsca 9, nonmetric scaling procedure, which arrives at a set of point coordinates via a different algorithm from that used in the present analysis. The value of r^2 between Torsca's projections and those reported here was 0.999–essentially perfect interval scale agreement.

2. Pitch difference estimation data were reanalyzed using Torsca 9. Again, r^2 for the projections produced by the two nonmetric analyses was 0.999.

(Received for publication July 26, 1973;
revision received September 20, 1973.)