Perception & Psychophysics
1990, 48 (5), 409-418

Does stimulus context affect loudness or only loudness judgments?

BRUCE SCHNEIDER
University of Toronto, Mississauga, Ontario, Canada

and

SCOTT PARKER
The American University, Washington, D.C.

Marks (1988) reported that when equal-loudness matches were inferred from magnitude estimates of loudness for tones of two different frequencies, the matches were affected by changes in the stimulus intensity range at both frequencies. Marks interpreted these results as reflecting the operation of response biases in the subjects' estimates; that is, the effect of range was to alter subjects' judgments but not necessarily the perception of loudness itself. We investigated this effect by having subjects choose which of two tone pairs defined the larger loudness interval. By using tones of two frequencies, and varying their respective intensity ranges, we reproduced Marks' result in a procedure devoid of numerical responses. When the tones at one frequency are all soft, but the tones at the other frequency are not all soft, cross-frequency loudness matches are different from those obtained with other intensity range combinations. This suggests that stimulus range affects the perception of loudness in addition to whatever effects it may have on numerical judgments of loudness.

The form of the psychophysical function relating loudness to sound intensity has been under investigation for over 100 years, ever since Fechner's original work. Many different methods have been employed in attempts to measure the loudness of sounds (for reviews, see Marks, 1974, 1979), but investigators in the field have not yet arrived at a general consensus about the exact form of the function (see Krueger, 1989, for a recent attempt to form such a consensus). There is general agreement that a power law ($L = kI^p$) relates loudness (L) to sound intensity (I) over a large range of intensity values, but there is lack of consensus about the proper value of the exponent (p). Because experimental values of p depend on features of the experiment which, in theory, should not affect it, investigators have differed about which set of conditions reveals the "true" value of the exponent. For example, experimental values of p depend on the method employed (e.g., magnitude versus interval estimation procedures; Stevens, 1971), and on the range of physical intensities explored (Teghtsoonian, 1973). Even the form of the function can be locally disturbed by factors such as stimulus spacing (Stevens & Galanter, 1957) and the inclusion of intensity values near threshold (Scharf & Stevens, 1961). Instructional variables such as choice of standard stimulus (Engen & Levy, 1955) and specification of its loudness value in magnitude estimation (Robinson, 1976) can also affect the form of the function relating loudness judgments to sound intensity. Thus, the apparent loudness function can be perturbed by manipulating features of the experiment which, in theory, ought not to affect its form or the value of its exponent.[1]

How, then, are we to explain this apparent malleability of the loudness function? One approach has been to invoke the notion of response bias. According to this approach, the "true" sensory representation is difficult to uncover, because of the operation of a number of "psychological" factors that bias the subjects' reports of sensory magnitude. This approach has been used by S. S. Stevens (1957, 1971) to explain why interval scaling methods do not produce the true loudness function (which he believed was provided by magnitude estimation techniques), and by ourselves to explain why magnitude estimation cannot reveal the true loudness function whereas nonmetric scaling techniques can (Schneider, Parker, Valenti, Farrell, & Kanow, 1978). Response biases, which have been invoked to explain why varying the range (Teghtsoonian, 1973), stimulus spacing (Stevens & Galanter, 1957), and instructional parameters (Robinson, 1976) perturb the form of the loudness function, have led to several schemes designed to eliminate, control, or counterbalance these response biases (Zwislocki & Goodman, 1980).

A second approach, which has been used by Marks (1979), is to argue that different types of comparisons among stimuli generate different loudness scales. According to Marks, some of the variation in observed exponents represents underlying variation in the manner in

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Jane Carey for conducting these experiments. Requests for reprints should be sent to Bruce Schneider, Department of Psychology, Erindale Campus, University of Toronto, Mississauga Road, Mississauga, Ontario L5L 1C6, Canada.

which sound is assessed. Therefore, interval estimation techniques, which involve comparisons of loudness differences, reveal a loudness scale different from that revealed by direct estimation techniques. This approach abandons the notion of a single underlying loudness scale and introduces the concept that loudness is not simply an immediate sensory response but rather that it is task-dependent. The notion of task-dependent loudness functions that are themselves subject to response biases, at the very least, complicates the search for the loudness function.

In spite of the difficulties in scaling loudness, certain consistencies have been observed. For example, when subjects are asked to give magnitude estimates of the loudness of tones that vary in frequency and intensity (Schneider, Wright, Edelheit, Hock, & Humphrey, 1972), equal loudness contours constructed from these judgments are consistent with equal loudness contours constructed from direct sensory matches of tones at different frequencies (Molino, 1973; Ross, 1967). Moreover, both magnitude estimation techniques and nonmetric scaling of loudness differences agree in this regard (Schneider & Bissett, 1987). Thus, three different techniques give consistent results with respect to equal loudness.

Recently, however, Marks (1988) has reported some data that seem to violate this consistency. Marks had subjects estimate the loudnesses of 16 tones, 8 at 500 Hz and 8 at 2500 Hz. In one condition, the SPLs of the 500-Hz tones ranged from 35 to 75 dB and the 2500-Hz tones ranged from 50 to 85 dB. In the other condition, the 500-Hz tones ranged from 55 to 90 dB and the 2500-Hz tones ranged from 30 to 65 dB. All 16 tones were presented within a single session. Marks found that the sound pressure level of the 2500-Hz stimulus that received the same loudness rating as a 65-dB 500-Hz tone was 70 dB in the first condition but only 53 dB in the other. Given this 17-dB change in the matching relationship, it is hard to maintain that equal magnitude estimates signify equal loudness unless the loudness of a tone is determined, in part, by the context provided by the rest of the stimulus set.

Numerical biases could provide a mechanism whereby equal magnitude estimates need not reflect equal loudness. Consider a case in which all of the 500-Hz tones were loud and all of the 2500-Hz tones were soft. If subjects were to assign numbers to accurately represent the loudnesses of these tones, the lowest number assigned to the 500-Hz tone would have to be greater than the largest number assigned to the 2500-Hz tone. Teghtsoonian (1973) found that subjects in magnitude estimation experiments tend to use the same numerical range, independently of the range of physical intensity. Because 500- and 2500-Hz tones are easily discriminable and, in fact, are separated by several critical bands, these subjects may have considered them as two separate sets of stimuli and adjusted their numerical ranges accordingly. Thus, a 45-dB 2500-Hz tone in a set of soft 2500-Hz tones would, when presented in conjunction with a set of loud 500-Hz tones, be given a higher numerical estimate than when it is included within a set of loud 2500-Hz tones paired with a set of soft 500-Hz tones. Such a response bias might overcome the instruction to judge the loudnesses of all stimuli on the same scale.

Other response biases, having comparable effects on stimulus matching, may be operative as well. For example, Marks, Szczesiul, and Ohlott (1986) showed how sequential dependencies in numerical estimation could account for a similar shift in the stimulus-matching function when their subjects compared tonal loudness to vibratory intensity and to brightness. Thus, the influence of context on loudness might be wholly attributable to the presence of response biases.

If the Marks phenomenon could be reproduced in a situation in which numerical biases could not operate, we would be left with the conclusion that context influences loudness. In the present experiments, we employed a nonmetric scaling technique that eliminates numerical response biases, to see if the Marks phenomenon would still appear. In this technique, subjects are presented with pairs of tones that can differ in both frequency (500 or 2500 Hz) and intensity. A subject hears two such pairs in a trial and is asked to choose the pair that has the greater loudness difference. Notice that the tones within a pair can differ in frequency, which forces the subject to evaluate all loudnesses on a single scale. Schneider and Bissett (1987) have shown that subjects in such a task do indeed act as if they were evaluating the tones along a unitary loudness dimension. Thus, unlike in the magnitude estimation technique, in which it is possible for subjects to differentially rate the loudnesses of 500- and 2500-Hz stimuli, here tones do not appear in isolation, subjects make no numerical ratings, and explicit loudness comparisons occur on every trial. Because loudness is evaluated on a single scale, we can change the range of the 500- and 2500-Hz tones and see whether or not the range changes affect the relative loudnesses of particular 500- and 2500-Hz tones.

METHOD

Subjects

Nine students and research assistants associated with the psychology department of the University of Toronto served as subjects in these experiments. Seven of the subjects served in Experiments 1 and 2; Subject M.S. served in Experiment 1 only; Subject A.K. served in Experiment 2 only. All were paid for their participation, and none had any known auditory pathology.

Apparatus

The subjects presented themselves with one of the two pairs of tones to be compared by pressing Button 1; pressing Button 2 resulted in the presentation of the second pair. The tones in a pair were 750 msec in duration and were separated by an intertone interval of 900 msec.

The pure tones were generated by a Hewlett-Packard programmable function generator (Model 3325A) under microcomputer control (Commodore C-64). The amplitude of the signal was controlled by a programmable attenuator and switched on and off with 10-msec rise and decay times. When Button 1 was pressed, the computer, during an initial delay of 50 msec, programmed the first fre-

Table 1
Intensity Values (dB SPL) of the Stimuli Employed in Experiments 1 and 2

Experiment 1				Experiment 2			
Condition A		Condition B		Condition A		Condition B	
.5 kHz	2.5 kHz	.5 kHz	2.5 kHz	.5 kHz	2.5 kHz	.5 kHz	2.5 kHz
35	32		32	35	32	35	
47	48		48	47	48	47	
55	60	55	60	55	60	55	60
68	72	68	72	68	72	68	72
	84	82	84	82		82	84
	95	90	95	90		90	95

quency to appear in Pair 1 and set the attenuator to the appropriate level. During the intertone interval, the frequency and intensity of the second tone were set. Pressing Button 2 delivered the second pair of tones in an identical manner. The subjects were allowed to listen to each pair as many times as they wished before indicating their judgment by pressing one of the two response buttons.

The earphones (TDH-49) were calibrated with a Brüel and Kjær 2209 sound-level meter using a 1-in. microphone in a 6-cm^3 coupler. The subjects were seated in a sound-attenuating chamber. Listening was monaural (right ear) in all experiments.

Procedure

The 10 stimuli in both experiments were tones of 500 and 2500 Hz. In each experiment, 4 tones were used at one frequency and 6 tones at the other. A list of stimuli used in Experiments 1 and 2 is shown in Table 1. Notice that the two experiments differed only in that the roles of the 500-Hz tones and the 2500-Hz tones were reversed. Each experiment included two conditions. In Condition A, there was a set of four low-intensity tones and a set of six tones with a broad intensity range; in Condition B, there was a set of four high-intensity tones and a set of six tones with a broad intensity range. Thus of the 45 ($10\times9/2$) pairs of tones that can be constructed from the set of 10, 24 pairs involve tones of different frequencies. From the 45 pairs, 990 ($45\times44/2$) pairs of pairs can be constructed. These pairs of tone-pairs were the stimuli presented to the subjects for comparison.

In each experimental condition (Experiment 1A-1B, Experiment 2A-2B), each subject was required to make 55 comparisons in each of 18 sessions. The 990 comparisons were presented in a different random sequence for each of the 8 subjects in each experiment. For any pair of pairs, there were eight possible arrangements of the component tones. Within a single pair, there were two possible orders in which the tones could occur; since there were two pairs, this means there were four possible orders of tones within pairs. Since, in addition, each pair could be assigned to Button 1 or Button 2, there were eight possible arrangements. The eight arrangements for each comparison of pairs were randomly assigned across the 8 subjects.

At the end of Experiments 1 and 2, the subjects were tested in two magnitude estimation sessions, which constituted Experiments 3 and 4. In Experiments 3A and 4A, they estimated the loudnesses of the 10 tones used in Condition A. In Experiments 3B and 4B, they estimated the loudnesses of the 10 tones used in Condition B. Within each session, they were presented with 10 blocks of the 10 tones, with a different random sequence of stimuli within each block. The order of conditions (A or B) was randomized across subjects. A free magnitude estimation procedure with no modulus was employed.

RESULTS

Comparisons of Loudness Intervals

In each of the experiments, 8 subjects judged which of two pairs of tones had the greater loudness difference for each of the 990 comparisons. If 5 or more of the 8 subjects judged pair (i,j) to differ more in loudness than pair (k,l), then we would consider the loudness interval in the first pair to be greater than the loudness interval in the second pair [i.e., $(i,j) > (k,l)$]. If only 4 subjects judged the first larger than the second, then we would consider the loudness difference comparison to be indeterminate [i.e., $(i,j) \approx (k,l)$]. The number of such indeterminate pairs or ties is given in column 1 of Table 2.

To determine loudness scales for each of the experiments, a nonmetric scaling procedure specifically designed for comparisons of intervals (Bissett & Schneider, 1990) was employed. This procedure attempts to assign loudness values, L, to each of the tones, so that whenever $(i,j) > (k,l)$ then $|L(i)-L(j)| > |L(k)-L(l)|$. Of course, in errorful data it is not possible to do this perfectly; hence, the program minimizes θ^2, a goodness of fit measure devised by Johnson (1973), which indexes the extent to which the above condition is satisfied. If $\theta^2 = 0$, then the loudness difference comparisons are perfectly predicted by the difference in loudness values. Values of θ^2 for Experiments 1 and 2 are shown in column 2 of Table 2.

Table 2
Number of Indeterminate Comparisons, θ^2, Coordinate Metric Recovery, Number of Tests of Monotonicity, Failure Rate for Monotonicity, Number of Conservative Tests of Monotonicity, Failure Rate for Conservative Tests

Experiment	Indeterminate Comparisons	θ^2	Coordinate Metric Recovery	Tests of Monotonicity	Monotonicity Failure Rate	Conservative Tests	Conservative Failure Rate
1A	97	.0008	.997	1643	11%	410	0.3%
1B	91	.0020	.996	1542	13%	283	2.1%
2A	111	.0012	.997	1777	10%	377	1.3%
2B	108	.0023	.996	1691	11%	296	0.3%

These values of θ^2 can be used to estimate the degree to which the projection values obtained from this program actually represent interval scale measurement. To accomplish this the index of coordinate metric recovery (CM) was estimated. CM is the squared Pearson correlation coefficient between the true values of the stimuli (which presumably generated the obtained comparisons) and the projection values produced by the program. Hence, CM varies between 0 and 1, and CM = 1 means that the true coordinate values have been perfectly recovered. In no empirical investigation done with these techniques are the true values known, but Bissett and Schneider (1990) have shown how CM can be estimated given the number of stimuli and the value of θ^2. Hence, if the estimated value of CM is sufficiently high, the point coordinates can be properly regarded as representing interval scale measurement. Column 3 of Table 2 presents the estimated values of CM in these experiments. Note that CM is estimated to be above .996 in all cases. Given these high values of CM, the projection values for each of the tones can be taken as representing interval scale measurement.

The loudness scale values determined in this fashion are unique only up to affine transformation—that is, up to addition of a constant and multiplication by a constant. If, for example, the loudness values determined from the nonmetric program were all doubled and 10 was then added to each, their differences would still predict the subjects' judgments equally well. So in order to compare loudness values from Conditions A and B, we first normalized them so that the loudness values for the 2500-Hz tones in Experiment 1A had a range of 1.0 and a mean of 1.0. A similar normalization procedure was followed for the tones in Experiment 1B. Each normalization process involved only affine transformations. Figure 1 plots the normalized loudness values of the 2500-Hz tones for Experiments 1A and 1B. Figure 1 shows that the interstimulus spacing in loudness is virtually identical in both conditions for the 2500-Hz tones. Recall that Experiments

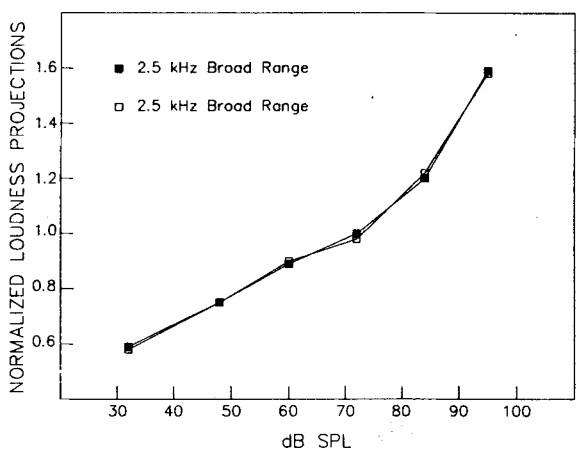

Figure 1. Loudness projections for the 2500-Hz tones in Experiments 1A (filled squares) and 1B (unfilled squares). The loudness projections in each experiment were normalized so that the 2500-Hz tones had a mean loudness of 1 and a loudness range of 1.

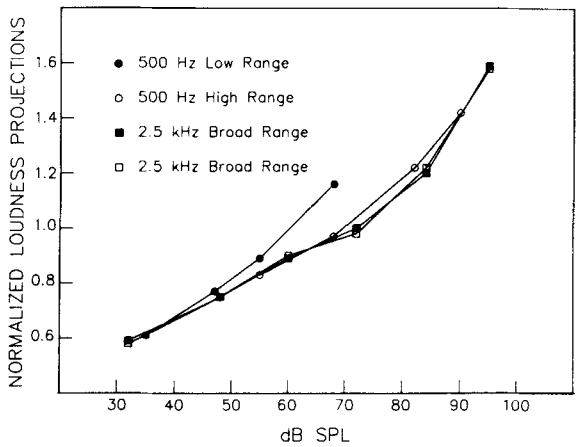

Figure 2. Loudness projections for the 500-Hz tones (circles) and 2500-Hz tones (squares) in Experiments 1A (filled symbols) and 1B (unfilled symbols). The loudness projections in each experiment were normalized so that the 2500-Hz tones had a mean loudness of 1 and a loudness range of 1.

1A and 1B differed in that different ranges of the 500-Hz tones were used. If the range of the 500-Hz tones were to have affected the loudnesses of the 2500-Hz tones, then we might have seen a difference in the two contours in Figure 1. The fact that the two contours are coincident means that the change in range of the 500-Hz tones from Condition A to Condition B changes the loudnesses of the 2500-Hz tones by at most an affine transformation (addition and multiplication by constants). Thus a change in the range of the 500-Hz tones leaves the interval scale properties of the loudness of the 2500-Hz tones unaffected.

In order not to violate the interval scale properties of the loudness of tones in Experiment 1A, the same affine transformation that normalized the loudness values for the 2500-Hz tones in Experiment 1A was also applied to the loudness values of the four 500-Hz tones in Experiment 1A. Similarly, to maintain interval scale properties of the loudnesses of tones in Experiment 1B, the affine transformation that normalized the 2500-Hz tones in that experiment was also applied to the 500-Hz tones. Note that we cannot apply different affine transformations to 500- and 2500-Hz tones within an experiment without altering the predicted directions of the loudness interval comparisons when the comparisons involve tones of both frequencies.

Figure 2 plots the normalized loudness scale values for both 500- and 2500-Hz tones. Note that the 68-dB, 500-Hz tone receives very different loudness values in Conditions A and B. In Experiment 1A, its loudness value is about equal to that of the 82-dB, 2500-Hz tone; in Experiment 1B, its loudness value is about equal to that of the 70-dB, 2500-Hz tone. Thus, a change in range produced a 12-dB shift in the equivalently loud, 2500-Hz tone.

Figure 3 shows the equivalent plot for the broad-range 500-Hz stimuli in Experiment 2. Note that in Figure 3 the loudnesses of the tones were normalized so that the

500-Hz tones had a mean loudness of 1 and a range of 1. Figure 3 shows that loudness intervals among the 500-Hz tones were unaffected by the change in range of the 2500-Hz tones between Conditions A and B.

Figure 4 plots the 2500-Hz tones along with the 500-Hz tones in Experiment 2. Note that the loudness of the 72-dB, 2500-Hz tone changes from Condition A to Condition B. In Experiment 2A (low range), its loudness is approximately equivalent to that of a 74-dB, 500-Hz tone; in Experiment 2B, its loudness is approximately equal to that of a 69.5-dB 500-Hz tone.

Thus, in both experiments the change from Condition A to Condition B does affect loudness. In both cases, the loudness intervals of the broad-range set of tones (2500 Hz in Experiment 1, 500 Hz in Experiment 2) are unaffected

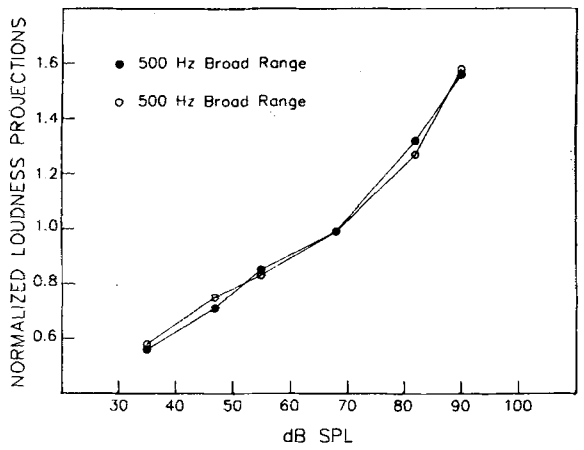

Figure 3. Loudness projections for the 500-Hz tones in Experiments 2A (filled circles) and 2B (unfilled circles). The loudness projections in each experiment were normalized so that the 500-Hz tones had a mean loudness of 1 and a loudness range of 1.

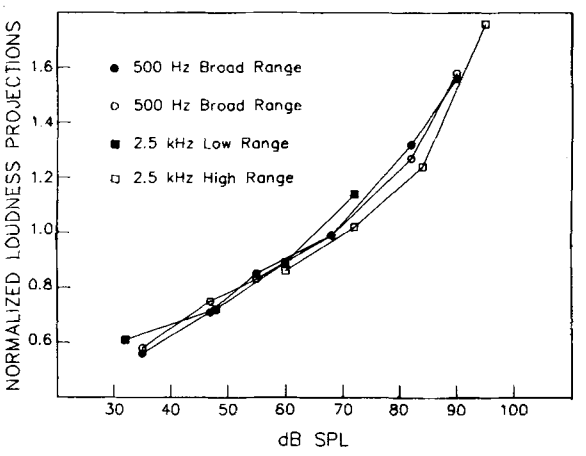

Figure 4. Loudness projections for the 500-Hz tones (circles) and 2500-Hz tones (squares) in Experiments 2A (filled symbols) and 2B (unfilled symbols). The loudness projections in each experiment were normalized so that the 500-Hz tones had a mean loudness of 1 and a loudness range of 1.

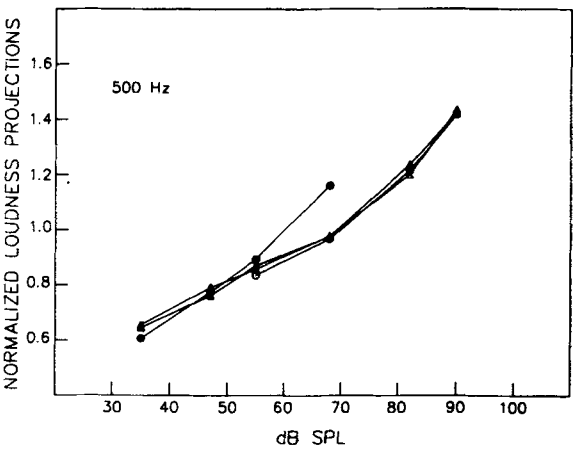

Figure 5. Normalized loudness projections (see text) for the 500-Hz tones in Experiments 1A (filled circles), 1B (unfilled circles), 2A (filled triangles), and 2B (unfilled triangles).

by changes in the range of the tones at the other frequency. However, when the loudness range for a particular frequency is changed, at least some of its loudness intervals change in size. An examination of equal-loudness relationships suggests that the loudness changes occur only when low intensities at one frequency are combined with a broad range of intensities at the other frequency. First, if we examine, for example, the 68-dB 500-Hz tone that was common to Conditions A and B in Experiments 1 and 2, we find that its equivalently loud 2500-Hz tone is 70 dB in Experiments 1B and 2B, and 82 and 67 dB in Experiments 1A and 2A, respectively. Note that Experiments 1B and 2B include only high intensities at one frequency, combined with a broad range of intensities of the other frequency. In both of these experiments (1B and 2B), we obtained the same loudness match for the 68-dB, 500-Hz tone. Moreover this loudness match is approximately what would be expected on the basis of equal-loudness matching procedures (Churcher & King, 1937; Robinson & Dadson, 1956; Ross, 1967). Experiments 1A and 2A combine low intensities at one frequency with a broad range of intensities at the other frequency. Depending on which frequency constitutes the broad-range set, the loudness match is considerably different (a change from 67 to 82 dB). This suggests that subjects do not experience any difficulty in maintaining equivalent loudness scales at the two frequencies when high intensities at one frequency are combined with a broad range of intensities at the other frequency, independent of which frequency constitutes the broad range. On the other hand, subjects apparently do not maintain the same loudness scale when low intensities at one frequency are combined with a broad range of intensities at the other frequency. The extent of this effect for the 500-Hz tones is shown in Figure 5.[2] The comparable plot for the 2500-Hz tones is shown in Figure 6.

Because the nonmetric scaling procedures do not easily permit statistical tests of differences between obtained

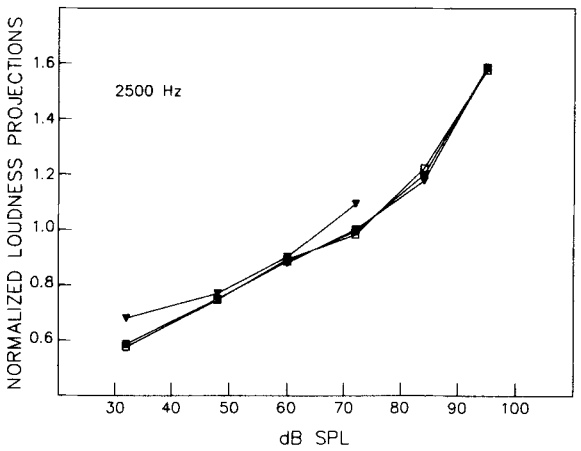

Figure 6. Normalized loudness projections (see text) for the 2500-Hz tones in Experiments 1A (filled squares), 1B (unfilled squares), 2A (filled triangles), and 2B (unfilled triangles).

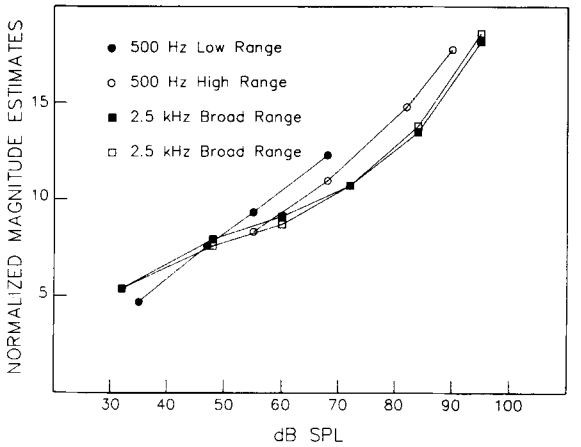

Figure 7. Magnitude estimates for the 500-Hz tones (circles) and 2500-Hz tones (squares) in Experiments 3A (filled symbols) and 3B (unfilled symbols). Magnitude estimates were normalized so that the geometric mean of the magnitude estimates of the 2500-Hz tones was the same in both Experiment 3A and Experiment 3B.

loudness projections, the data from each individual subject at each condition were submitted to the nonmetric program in order to obtain individual loudness values for each stimulus. For purposes of comparison, these loudness values were normalized in the same fashions as were the values from the group data. The group data from Experiment 1 suggest that the 68-dB, 500-Hz stimulus has a higher loudness value when it is part of a low-intensity rather than a high-intensity group (see Figure 2). This pattern was observed in all 8 subjects ($p < .008$, two-tailed, binomial test). The group data from Experiment 2 suggest that the 72-dB 2500-Hz tone has a higher loudness value when it is part of a low-intensity rather than a high-intensity group (see Figure 4). This pattern was also observed in all 8 subjects ($p < .008$, two-tailed, binomial test). Thus, the range effect found in both Experiments 1 and 2 is clearly statistically significant.

Magnitude Estimates of Loudness

The stimuli in Experiments 3 and 4 were the same as those in Experiments 1 and 2, respectively, but loudness values for the stimuli were obtained with the method of magnitude estimation. Figure 7 plots the magnitude estimates obtained in Experiments 3A and 3B. The 2500-Hz stimuli were multiplied by a constant so that the geometric mean magnitude estimate was the same in both Condition A and Condition B. The multiplicative constant applied to the 2500-Hz stimuli in Condition A was applied to the 500-Hz stimuli in the same condition to maintain loudness matches across frequencies. The corresponding procedure was followed for the 500-Hz tones in Condition B. Figure 7 shows that the normalized magnitude estimates are nearly equivalent for the 2500-Hz tones, and that there is a definite range effect. The pattern of these results differs from those shown in Figure 2 primarily in that the loudness matches across frequency differ from those in Experiment 1. For example, the 68-dB, 500-Hz tone when it is a member of the high intensity set matches a 70-dB, 2500-Hz tone in Experiment 1 and a 73-dB tone in Experiment 3. In general, three of the four high-intensity tones in Experiment 3B appear to be louder (in terms of their matches with 2500-Hz tones) than they are in Experiment 1. Figure 8 shows the corresponding results for Experiment 4. Again, the normalized 500-Hz tones are nearly identical in the two conditions, with the magnitude estimates for the 2500-Hz tones demonstrating a strong range effect. There is no evidence, however, that the loudness matches for the high-intensity 2500-Hz tones are substantially different from those in Experiment 2 (see Figure 4). Thus the results of the magnitude estimation experiment are consistent with those from the

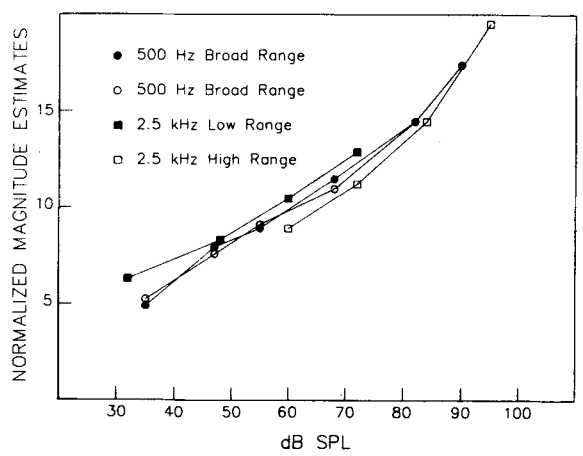

Figure 8. Magnitude estimates for the 500-Hz tones (circles) and 2500-Hz tones (squares) in Experiments 4A (filled symbols) and 4B (unfilled symbols). Magnitude estimates were normalized so that the geometric mean of the magnitude estimates of the 500- Hz tones was the same in Experiments 4A and 4B.

loudness comparison experiment, with the exception that the loudness matches for the 500-Hz tones in Experiment 1B, in which paired comparisons of loudness intervals were used, differ from those obtained in Experiment 3B, in which a magnitude estimation technique was employed.

DISCUSSION

Nonindependence from Irrelevant Alternatives

Experiments 1-4 confirm Marks' (1988) observation that when the stimulus set includes two or more frequencies, the apparent loudnesses of tones are altered by changing the range of stimuli at one or both of the frequencies, and that this effect occurs in magnitude estimation and in the paired comparison of loudness intervals. This phenomenon has disturbing consequences for the notion of a unitary loudness scale or even a finite number of loudness scales (Marks, 1979). The data from these experiments show, for example, that the loudness of a 68-dB 500-Hz tone depends, in part, on the other 500-Hz tones in the stimulus set. What is even more disturbing is that it is difficult to attribute this effect to numerical response biases that could be imposed at an output-response stage—that is, at some cognitive level of processing that is remote from the experience of loudness itself (Rule & Curtis, 1982). In Experiments 1 and 2, the possibility of numerical response biases was removed by asking subjects to indicate which of two loudness intervals was larger. Nevertheless, we still observed a range effect. What this means is that the subjects' judgments as to which of two loudness intervals was larger were influenced by tone intensities that appeared in other comparisons but were not a part of the comparison being made; that is, the subjects' comparisons of loudness intervals were influenced by irrelevant alternatives (Luce, 1959). We can see this in the direct inspection of the comparisons. If we compare the tone sets in Experiments 1A and 1B, we will see that they have 8 stimuli in common. We would expect that the loudness interval comparisons that involve only these 8 stimuli should be identical in Experiments 1A and 1B, apart from experimental error, if the comparisons were independent of irrelevant loudness intervals. Figure 2, however, suggests that the loudness value of the 68-dB, 500-Hz tone is larger in Experiment 1A than in Experiment 1B. Therefore, we would expect that at least some of the comparisons involving that stimulus would change from 1A to 1B. After excluding indeterminate comparisons [cases in which $(i,j) \approx (k,l)$ in either or both of the experiments], we found the percentage of times in which $(i,j) > (k,l)$ in Experiment 1A but $(i,j) < (k,l)$ in Experiment 1B for the 133 comparisons involving this stimulus and for the 186 comparisons that did not include this stimulus. The rate of disagreement for the comparisons involving the 68-dB, 500-Hz stimulus was 0.15, whereas it was only 0.07 for the others. This difference is significant ($z = 2.2$, $p < .05$). Similarly, in Experiments 2A and 2B, the rate of disagreement for the 125 comparisons involving the 72-dB, 2500-Hz stimulus was .16, whereas it was .05 for the 176 remaining comparisons ($z = 2.9$, $p < .01$). These data clearly show that the loudness interval comparisons in these experiments are not independent of irrelevant alternatives. Rather, the choice of which of two loudness intervals is the larger depends, in part, on what other loudness intervals are in the experiment.

It is not altogether clear what this effect is due to. The simplest interpretation is that it represents a contextual effect on tonal loudness. This interpretation rests on the assumption that there is some single dimension called tonal loudness within which context has its effect. Another interpretation is that in a complicated stimulus environment subjects cannot maintain one unidimensional loudness scale within which all comparisons are performed. The magnitude estimation paradigm does not allow one to distinguish between these two alternatives. The present paradigm permits us, in theory, to determine whether or not these comparisons are being made within a single unidimensional loudness scale.

The Unidimensionality of Loudness Judgments

The presumption that underlies this set of experiments is that subjects can successfully locate all the stimuli along a single dimension, assess stimulus differences within it, and compare those loudness differences. That subjects can successfully do this for a set of 10 1200-Hz stimuli was shown in Schneider, Parker, and Stein (1974), and confirmed for individual subjects by Schneider (1980). Moreover, Schneider and Bissett (1987) showed that subjects could assess the loudness differences among a set of stimuli comprising two different frequencies with comparable loudness ranges along a single loudness dimension. Because the stimuli in the present experiments are more widely separated in frequency than were the stimuli in Schneider and Bissett and also subtend different loudness ranges, it is important to check to see whether or not the present subjects also reached their judgments in that way. As task difficulty and complexity increase, subjects may become unable to evaluate all loudness intervals along a single dimension.

Krantz, Luce, Suppes, and Tversky (1971) have shown that if difference judgments are to be represented as distances along a unidimensional psychological scale such as loudness, then difference comparisons—in the present case, of loudness—must satisfy the axioms of a positive difference structure. The critical axiom for unidimensionality is the monotonicity condition. A set of loudness difference comparisons satisfies monotonicity if $(a,b) \geq (x,y)$ and $(b,c) \geq (y,z)$ implies $(a,c) \geq (x,z)$, for $a \leq b \leq c$ and $x \leq y \leq z$. Here, point b divides interval (a,c) into two interior sections, and point y divides interval (x,z) into two interior sections. Satisfaction of monotonicity requires that whenever the two interior sections of one interval exceed the corresponding interior segments of another interval, then the interval with the larger interior sections is the larger interval. The condition is illustrated

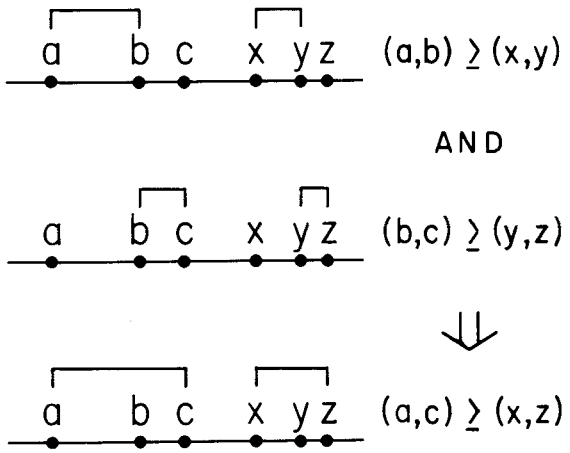

Figure 9. Illustration of the monotonicity condition for points on a line segment.

in Figure 9 and must be true of any points a,b,c and x,y,z taken from the same line. If the loudnesses of the 500- and 2500-Hz tones are to be represented as points on a line segment, then loudness comparisons must satisfy the monotonicity condition. If the loudness comparisons do not satisfy monotonicity, then one or more of the points must be off the line, thereby necessitating that more than one dimension represent the loudnesses of the stimuli.

We tested for monotonicity in Experiments 1A, 1B, 2A, and 2B by looking for all of the comparisons of pairs for which we also had comparisons of interior sections. To do this, we needed to know the loudness order of the 10 tones in each experiment. We used for this purpose the L values assigned to the tones in each experiment by the nonmetric scaling program. In each experiment, we found the percentage of instances in which triplets of comparisons violated the monotonicity condition. The number of tests of monotonicity and the percentage of such tests that violated monotonicity are seen in columns 4 and 5 of Table 2. Notice that the rate of violation is nearly constant—approximately 11% in all four experiments.

This value, 11%, is considerably higher than the 1.6% found in Schneider, Parker, and Stein (1974) for a group of 8 subjects who compared loudness intervals defined by pairs of 1200-Hz tones. It is also larger than the range of values (0.3%-2.2%) found for the individual subjects in Schneider (1980), who were tested at the same frequency. Rates of violation of monotonicity ranged from 3.4%-5.6% in the three experiments of Schneider and Bissett (1987), whose subjects compared intervals involving tones of two frequencies with similar loudness ranges. Those results provide a context for the present ones.

The most natural interpretation of the higher failure rates in Schneider and Bissett (1987) is that to increase the complexity of the stimuli in the tone pairs by using tones of two frequencies increases the rate of violation of the monotonicity condition.[3] Our experiment complicated the subject's task further, in that the loudness ranges differed substantially and the frequency separation was larger. This increase in failure along with increases in experimental complexity could reflect increased sources of variability in the subject's judgments, or increasing departures from unidimensionality. The lack of an error theory for monotonicity complicates the task of deciding which of these two alternatives is more likely. Nevertheless, some other features of the data may provide some guidance with respect to this issue.

In testing for monotonicity in all of these experiments, indeterminate comparisons $[(i,j) \approx (k,l)]$ were included. In theory, indeterminate comparisons are most likely to occur for intervals that are nearly equal psychologically. Given judgmental variability, the inclusion of these indeterminate comparisons may lead to an increase in failure rate. To see this, consider an example. Suppose that (a,b) is slightly smaller than (x,y), but that because of judgmental variability, 4 subjects judge it larger $[(a,b) \approx (x,y)]$. Suppose further that (b,c) is slightly larger than (y,z) and 6 of the 8 subjects judge it so $[(b,c) > (y,z)]$. Finally, suppose that (a,c) is slightly smaller than (x,z), so that 5 subjects judge it as smaller $[(a,c) < (x,z)]$. This triplet of comparisons violates monotonicity, because subject variability can produce errors in comparisons of nearly equal loudness intervals. To reduce the effect of subject variability on violations of monotonicity, we reanalyzed the data, regarding not only four-four splits but also five-three splits as indeterminate rather than decisive. This narrows the range of decisive votes—now, at least 6 subjects must agree that (a,b) exceeds (x,y) for us to record $(a,b) > (x,y)$. If 5 subjects report that (a,b) exceeds (x,y), we record $(a,b) \approx (x,y)$. This means that there has to be strong agreement in any triplet of comparisons that is to violate monotonicity. Furthermore, we required all three comparisons in any test of monotonicity to be decisive. This is tantamount to replacing "$\geq$" in the monotonicity test with "$>$." This more conservative test of monotonicity should substantially reduce the failure rate for randomly perturbed judgments from a unidimensional loudness scale. Table 2 shows that the failure rates for this conservative test ranged from 0.3% to 2.1%. Thus, broadening our definition of indeterminacy and using only decisive comparisons produced relatively few violations of monotonicity. This is indeed what we would expect if the violations of monotonicity in our original test reflected the effects of random perturbations on judgments of loudness intervals. If, on the other hand, violations of monotonicity were primarily due to the fact that more than one dimension were required to represent the data, we would not necessarily expect such a large reduction in the observed number of violations. Similar application of the conservative monotonicity tests to the Schneider and Bissett (1987) experiments reduced the rate of failure to zero in one experiment and 0.2% in the other two.

The fact that the rate of failure in the present study for both the regular and the conservative tests of monotonicity is higher than in Schneider and Bissett (1987) may in-

dicate that there is simply a larger random error term in the present experiment. Of course, it is always possible that it reflects deviations from unidimensionality. To check for deviations from unidimensionality, we determined two-dimensional solutions to the data in the present experiments. Two-dimensional solutions produced only marginal reductions in the size of θ^2, indicating that the addition of another dimension did not particularly improve the goodness of fit. In none of the four experiments did an inspection of the solution yield an obvious interpretation. Furthermore, no two of the four solutions looked alike (except for a coarse ordering of the stimuli with respect to loudness), which is what we would expect if the second dimension were simply capturing some of the random error. All things considered, we have no strong evidence that the judgments are not based on a unidimensional loudness scale.

Implications for Loudness Scales

If the judgments are truly based on a unidimensional loudness scale, we are forced to conclude that context influences how the auditory system encodes or represents the loudnesses of sounds. Specifically, Figures 5 and 6 suggest that if soft tones at one frequency are combined with a broad range of tones at another frequency, the loudnesses of at least some of the soft tones are increased. On the other hand, a combination of loud tones at one frequency and a broad range of tones at the other frequency does not seem to affect the loudnesses of the tones at either frequency for the paired comparison judgments. (The fact that it does appear to do so in Experiment 3 suggests the operation of additional biases in magnitude estimation responses.) This raises the interesting questions of what mechanisms are responsible for the loudness change and why they are operative only for the set of low-intensity stimuli.

The change in loudness that occurs for the low-intensity set is consistent with the notion that context influences the "gain" or amplitude in that channel or critical band. During the course of the experiment, the subject is exposed to only low-intensity tones at one frequency. If subjects can "turn up" the gain in this frequency region, they might be able to process these stimuli more easily. In fact, the results from both Experiment 1 and Experiment 2 are consistent with a variable linear gain. If all incoming signals at that frequency are multiplied by a constant, the net effect would be to add a constant number of decibels to each signal. Note, however, that at very low intensities, a 4-dB change in intensity will produce a negligible change in loudness. At higher intensities, the same 4-dB change can produce a considerable change in loudness. The pattern of results in Figures 5 and 6 are roughly consistent with this model.

If the "gain" hypothesis is correct, this context effect should not occur for tonal frequencies separated by less than a critical band. For in that instance, the same gain would apply to all of the tones at both frequencies, thereby eliminating the context effect. Indeed, Marks (in press) reports that the effect does not occur for two sets of tones whose frequencies fall within a critical band.

Note that this kind of model is an example of top-down processing. If the system has a variable gain control available, it can utilize this feature to enhance, for example, information transfer. In the present example, the system, because of past input, expects only low-intensity sounds at one frequency. To compensate for this, it adjusts its gain to a higher level, thereby changing the loudness function. It is interesting to note that gain controls have been proposed as a way of accounting for certain other auditory phenomena such as binaural unmasking (the equalization portion of Durlach's equalization and cancellation model; Durlach, 1972).

The implication of a variable gain control for loudness is that there are at least as many loudness scales as there are settings on the gain control. The loudness scale that is operative at any one time will depend on context. The need for at least two loudness scales has been recognized by Marks (1979), to account for differences between magnitude estimation experiments and interval estimation experiments. A variable gain control would give us many more. It could also account for phenomena such as the Teghtsoonian (1973) range effect, as well as the effects of stimulus spacing on magnitude and category estimation. To determine the extent to which such a mechanism can account for these phenomena, it will be necessary to separate its effects from those of numerical response biases in these procedures.

REFERENCES

BAIRD, J. C., & NOMA, E. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

BISSETT, R. J., & SCHNEIDER, B. (1990). *Spatial and conjoint models based on pairwise comparisons of dissimilarities and conjoint effects: Complete and incomplete designs*. Unpublished manuscript.

CHURCHER, B. G., & KING, A. J. (1937). The performance of noise meters in terms of the primary standard. *Journal of the Institute of Electrical Engineering*, **81**, 57-90.

DURLACH, N. I. (1972). Binaural signal detection: Equalization and cancellation theory. In J. V. Tobias (Ed.), *Foundations of modern auditory theory: Vol. II* (pp. 369-462). New York: Academic Press.

ENGEN, T., & LEVY, N. (1955). The influence of standards on psychophysical judgments. *Perceptual & Motor Skills*, **5**, 193-197.

KRUEGER, L. E. (1989). Reconciling Fechner and Stevens: Towards a unified psychophysical law. *Behavioral & Brain Sciences*, **12**, 251-267.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., & TVERSKY, A. (1971). *Foundations of measurement: Vol. I. Additive and polynomial representations*. New York: Academic Press.

LUCE, R. D. (1959). *Individual choice behavior*. New York: Wiley.

MARKS, L. E. (1974). On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as science. *Perception & Psychophysics*, **16**, 358-376.

MARKS, L. E. (1979). A theory of loudness and loudness judgments. *Psychological Review*, **86**, 256-285.

MARKS, L. E. (1988). Magnitude estimation and sensory matching. *Perception & Psychophysics*, **43**, 511-525.

MARKS, L. E. (in press). The dynamics of ratio scaling. In G. A. Gescheider & S. J. Bolanowski (Eds.), *Ratio scaling of psychological magnitudes*. Hillsdale, NJ: Erlbaum.

MARKS, L. E., SZCZESIUL, R., & OHLOTT, P. (1986). On the cross-

modal perception of intensity. *Journal of Experimental Psychology: Human Perception & Performance*, **12**, 517-534.

MOLINO, J. A. (1973). Pure-tone equal-loudness contours for standard tones of different frequencies. *Perception & Psychophysics*, **14**, 1-4.

ROBINSON, D. W., & DADSON, R. S. (1956). A redetermination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, **7**, 166-181.

ROBINSON, G. H. (1976). Biasing power law exponents by magnitude estimation instructions. *Perception & Psychophysics*, **19**, 80-84.

ROSS, S. (1967). Matching functions and equal-sensation contours for loudness. *Journal of the Acoustical Society of America*, **42**, 778-793.

RULE, S. J., & CURTIS, D. W. (1982). Levels of sensory and judgmental processing: Strategies for evaluation of a model. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 107-122). Hillsdale, NJ: Erlbaum.

SCHARF, B., & STEVENS, J. C. (1961). The form of the loudness function near threshold. In *Proceedings of the Third International Congress of Acoustics* (pp. 80-82). Amsterdam: Elsevier.

SCHNEIDER, B. (1980). Individual loudness functions determined from direct comparisons of loudness intervals. *Perception & Psychophysics*, **28**, 493-503.

SCHNEIDER, B. A., & BISSETT, R. J. (1987). Equal loudness contours derived from comparisons of sensory differences. *Canadian Journal of Psychology*, **41**, 429-441.

SCHNEIDER, B. A., PARKER, S., & STEIN, D. (1974). Measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology*, **11**, 259-273.

SCHNEIDER, B., PARKER, S., VALENTI, M., FARRELL, G., & KANOW, G. (1978). Response bias in category and magnitude estimation of difference and similarity for loudness and pitch. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 483-496.

SCHNEIDER, B., WRIGHT, A. A., EDELHEIT, W., HOCK, P., & HUMPHREY, C. (1972). Equal loudness contours derived from sensory magnitude judgments. *Journal of the Acoustical Society of America*, **51**, 1951-1959.

STEVENS, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153-181.

STEVENS, S. S. (1971). Issues in psychophysical measurement. *Psychological Review*, **78**, 426-450.

STEVENS, S. S., & GALANTER, E. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, **54**, 377-411.

TEGHTSOONIAN, R. (1973). Range effects in psychophysical scaling and a revision of Stevens' law. *American Journal of Psychology*, **86**, 3-27.

ZWISLOCKI, J. J., & GOODMAN, D. A. (1980). Absolute scaling of sensory magnitudes: A validation. *Perception & Psychophysics*, **28**, 28-38.

NOTES

1. For a more detailed discussion of the factors that can affect the form of the loudness function, see Baird and Noma, 1978, chapter 6.

2. In constructing Figure 5, the broad-range 500-Hz tones in Experiments 2A and 2B were first normalized to have a mean loudness of 1.0 and a range of 1.0. Second, the 500-Hz high-intensity tones in Experiment 1B were adjusted by an affine transformation so that they would have the same mean value and loudness range as their counterparts within the broad-range tones of Experiment 2. This same affine transformation was applied, of course, to the 2500-Hz tones in Experiment 1B, to maintain the loudness matches observed in this experiment. The loudnesses of the tones in Experiment 1A were then adjusted so that the mean and range of the 2500-Hz tones were coincident with the mean and range of the 2500-Hz tones in Experiment 1B. Thus, the broad-range 500-Hz tones in Experiment 2 were adjusted by affine transformations to have the same mean and range, and the high-intensity tones from Experiment 1B were adjusted to be as coincident as possible with the loudnesses of the 500-Hz tones from Experiment 2. Finally, in order to preserve the loudness matching relations observed in Experiment 1, the loudness values from Experiments 1A and 1B were adjusted so that the 2500-Hz tones had the same mean and range.

3. One procedural difference between Schneider et al. (1974) and Schneider and Bissett (1987) that might contribute to the higher failure rate is that in the earlier experiments, the ordering of the tones with respect to loudness was known in advance of testing. Therefore, it was not necessary to test certain intervals (see Schneider et al., 1974, pp. 261-262). In Schneider and Bissett (1987), as in the present study, the ordering was not known, and therefore all 990 comparisons of intervals were tested instead of the 540 tested in Schneider et al. (1974) and Schneider (1980). Given that subjects' judgments of intervals are variable, an additional 450 comparisons increases the opportunity for violations of monotonicity. Because the total number of possible tests of monotonicity is the same in both cases, this could lead to a higher failure rate even without an intrinsic increase in the variability of the subject's judgments. However, the increase in failure rate from 4.5% in Schneider and Bissett (1987) to 11% in the present experiment cannot be explained in this fashion, because the full set of 990 comparisons of intervals was employed in both.

(Manuscript received September 11, 1989;
revision accepted for publication June 5, 1990.)