Perception & Psychophysics
1988, 43, 211-222

The additivity of loudness across critical bands: A conjoint measurement approach

BRUCE SCHNEIDER
University of Toronto, Mississauga, Ontario, Canada

Five subjects were required in each trial to compare directly two sounds and to indicate which sound was louder. Each of the 64 sounds employed consisted of a combination of one of eight intensity levels of a 2-kHz tone and one of eight intensities of a 5-kHz tone. If, as Fletcher and Munson (1933) argued, loudness is additive for tone combinations in which the frequencies are widely separated, then subjects' judgments should reflect the summed loudnesses of the 2- and 5-kHz tones in a two-tone combination. Judgments of individual subjects were shown to satisfy the conditions for an additive structure, and individual loudness scales were constructed. These loudness scales varied from subject to subject. Since this paired comparison procedure minimized response biases, the results suggest substantial individual differences in the sensory representation of sound intensity. The relations among sensory scales derived from other structured sensory judgments, such as binaural loudness, are discussed.

When two or more pure tones differing in frequency are presented simultaneously to the same ear, the listener is able to parcel out perceptually the individual tones in the complex, providing that the components are not too numerous and differ sufficiently in frequency (Ohm's acoustical law). This does not mean, however, that the perception of the complex is without global attributes. Despite the decomposability of the percept into its components, the complex, as a whole, possesses a definite timbre and total loudness. The problem then is to relate the perceptual attributes of the components (pitch and loudness) to the global attributes of the complex (timbre and total loudness). In this paper, a conjoint measurement approach is used to determine the relationship between the loudness of the components of a complex and the global loudness of the complex.

Total Loudness and the Critical Band

The relationship of the loudness of a complex sound to its individual components depends on the frequency separation of the components. If the individual components all fall within a narrow range of frequencies (i.e., within a critical band), the overall loudness depends solely on total acoustic energy (see, e.g., Scharf, 1970; Zwicker, Flottorp, & S. S. Stevens, 1957), that is, $L_t = f(\Sigma E_i)$, where L_t is global loudness and E_i is the energy in tone i. When the component frequencies are widely separated (fall in separate critical bands), it is generally believed that the loudness of the complex is the sum of the loudnesses within each critical band, that is, that $L_t = \Sigma L_j$, where L_j is the loudness in critical band j (see Fletcher & Munson, 1933; Marks, 1978b). Only Marks (1978b), however, has carefully looked at this relationship, and his investigation focused on group data rather than individual subjects.

Testing Loudness Additivity with a Paired Comparison Design

If loudness additivity holds for two tones separated by more than a critical band, then total loudness $L_t = L_1 + L_2$. If we have numerical measures of L_1, L_2, and L_t, it is relatively easy to test for additivity among these measures (see, e.g., Anderson, 1970, 1974). It is, however, difficult to obtain unbiased numerical measures of loudness. Numerous experiments involving judgments of sensory sums, sensory differences, and sensory ratios have demonstrated the existence of nonlinear response biases in numerical estimates (Algom & Marks, 1984; Anderson, 1974; Curtis, Attneave, & Harrington, 1968; Curtis & Rule, 1972; Fagot, Stewart, & Kleinknecht, 1975; Marks, 1978b; Rule, Curtis & Markley, 1970; Rule, Laye, & Curtis, 1974; Schneider, Parker, Valenti, Farrell, & Kanow, 1978). If numerical estimates are nonlinearly biased, then testing for additivity among these measures is not a valid test for additivity of loudness. Loudness might indeed be additive while our measures of it are not.

The dependency of standard tests of additivity on numerical estimates of loudness can be avoided by using a paired comparison procedure. Consider a comparison between two complex sounds, both of which consist of the same two pure tones, but at different intensities. Representing the intensities of the first frequency with the letters $a,b,c,\ldots$ and the intensities of the second frequency with the letters $p,q,r,\ldots$, a comparison between two complex sounds can then be represented as (a,p) versus (b,q).

Preparation of this article was supported in part by the Natural Sciences and Engineering Research Council of Canada. I would like to thank Randall Bissett for writing the program and conducting the simulations found in the Appendix. I would also like to thank Scott Parker and Annabel Cohen for their helpful comments on an earlier version of the manuscript.
Requests for reprints should be sent to Bruce Schneider, Department of Psychology, Erindale College, University of Toronto, Mississauga Road, Mississauga, ON L5L 1C6, Canada.

The listener's task is to declare whether (a,p) is louder than (b,q) or vice versa—that is, to decide which of the complex sounds has the greater total loudness. If additivity holds, then this judgment should depend on whether or not $L_1(a) + L_2(p)$ is greater than $L_1(b) + L_2(q)$, where the subscripts 1 and 2 stand for frequencies 1 and 2, respectively. Luce and Tukey (1964) have shown that binary judgments such as these not only permit a test of additivity, but also can be used to construct loudness scales for the individual tones and for the tonal complex. Because numerical judgments are not included, we avoid the problem of response bias.

In the present experiment, eight intensity values of a 2-kHz tone were paired with eight intensity values of a 5-kHz tone for a total of 64 tone-pair combinations. If additivity holds in this 8×8 matrix, then the total loudness for pair (a,p) is

$$L_t = L_2(a) + L_5(p), \qquad \text{(Additivity)}$$

where the subscripts 2 and 5 stand for 2- and 5-kHz tones. Furthermore, if observers are actually comparing the loudnesses of combination tones, they should respond that combination (a,p) is louder than some other combination (b,r) if and only if L_t for pair (a,p) is greater than L_t for pair (b,r). Given 64 combination tones, there are $64 \times 63/2 = 2{,}016$ distinct paired comparisons. Not all of these paired comparisons need be tested. For example, consider the comparison (60,45) versus (55,40). Since the intensities of both the 2- and the 5-kHz stimuli in the first pair are greater than their counterparts in the second pair, it is clear that the first pair should be judged louder than the second pair. In general, if we assume that the loudness of a two-tone complex is monotonic with the intensities of the pure tones, we can eliminate 1,236 comparisons in which monotonicity specifies which two-tone complex is louder. This leaves 784 comparisons that require testing.

As Krantz, Luce, Suppes, and Tversky (1971) have shown, additivity requires that the paired comparisons satisfy certain constraints. One of these constraints is transitivity, that is,

$$(a,p) \geq (b,r) \text{ and } (b,r) \geq (c,s)$$
$$\text{implies } (a,p) \geq (c,s). \qquad \text{(Transitivity)}$$

A second constraint, which is not so immediately obvious, is referred to as double cancellation. If loudnesses are additive, then

$$(a,p) \geq (b,r) \text{ and } (b,s) \geq (c,p)$$
$$\text{implies } (a,s) \geq (c,r). \qquad \text{(Cancellation)}$$

We can identify the reason for this condition if we substitute in for each of the pairs in the equation above, their corresponding loudnesses. If we do that, we have

$$L(a)+L(p) \geq L(b)+L(r)$$

and

$$L(b)+L(s) \geq L(c)+L(p).$$

If we now add these two inequalities, we have

$$L(a)+L(p)+L(b)+L(s) \geq L(b)+L(r)+L(c)+L(p).$$

Eliminating common terms on both sides of the inequality leaves

$$L(a)+L(s) \geq L(c)+L(r).$$

Thus, if loudness is additive, double cancellation must hold. Krantz et al. (1971) have shown that for all effective purposes, if double cancellation is not violated, the system is additive. We can, therefore, test whether or not we are justified in assuming that loudnesses are additive across critical bands.

Deriving Loudness Scales from Comparisons of Combination Tones

Provided that the system is additive, the paired comparison judgments can be used to derive loudness scales for the 2- and the 5-kHz tones that are, for all practical purposes, unique up to addition and multiplication by a constant. Specifically, if P_2 and P_5 are loudness scale values derived from the paired comparison judgments for 2- and 5-kHz tones, respectively, then P_2 and P_5 are related to the "true" loudness values by the following transformations: $P_2 = aL_2+b_2$ and $P_5 = aL_5+b_5$, where a, b_2, and b_5 are constants.

In the present experiment, 6,272 comparisons of two-tone complexes were obtained from each of 5 subjects. These comparisons were used to determine whether the judgments of individual subjects satisfied transitivity and double cancellation, that is, to determine whether loudness is additive. Loudness scales were also constructed from these judgments to look for individual differences, if any, among subjects.

METHOD

Subjects

Four undergraduate students and a research assistant served as subjects. All 5 subjects had previously participated in psychophysical experiments.

Apparatus

The two pure tones (2 and 5 kHz) that appeared in the first tone pair were generated by two Hewlett-Packard oscillators (Model 204C). The output of each oscillator, after attenuation, was sent to the same operational amplifier to add the two signals. The output of this amplifier was gated by an electronic switch with a rise and decay time of 10 msec to dampen transients. Pressing button 1 gated this circuit to the earphone and produced a 750-msec presentation of this two-tone combination.

An identical circuit was used to generate the second tone pair. Pressing button 2 gated the second circuit to the earphone and produced a 750-msec presentation of this tone pair. The subjects had unlimited opportunity to switch back and forth between tone pairs before making a judgment as to which tone pair was louder. Successive buttonpresses were not effective until 750 msec following the end of the last tone-pair presentation.

Listening was monaural (right ear) throughout the entire experiment. The subject sat inside an Industrial Acoustics single-walled sound-attenuating chamber and communicated with the experimenter via an intercom.

Procedure

The subjects were requested to judge which two-tone combination was louder. For both the 2- and 5-kHz tones, there were eight different levels of sound pressure (-48, 40, 48, 54, 60, 68, 74, and 80 dB SPL). Note that the first one of these levels is clearly below threshold and is equivalent to a null presentation. These eight levels for each tone result in 64 distinct two-tone combinations, and $64\times63/2 = 2{,}016$ distinct comparisons. By assuming that the loudness of a two-tone combination is monotonic with the intensities of the individual tones, the number of combinations actually tested was reduced to 784. Each subject listened to each of these 784 comparisons a total of eight times over the course of 72 sessions.

During the first 9 sessions, a subject was presented once with each of the 784 comparisons in a random order. The first 8 sessions contained 90 comparisons, and the 9th contained 64 comparisons. The second set of 784 comparisons was presented in a different random order during the next 9 sessions. For each comparison, (a,p) versus (b,r) in the first set, the assignment of tone pair to button was reversed in the second set of 784 comparisons. This same procedure was followed for Replications 3 and 4, 5 and 6, and 7 and 8. Thus, order effects were balanced across replications within an individual. Sessions generally lasted 45-60 min with a 10-min break halfway through the session.

RESULTS

Transitivity and Double Cancellation

The data from each subject consisted of the number of times he or she judged one two-tone combination to be louder than another for each of the 784 comparisons. If combination (a,p) was judged to be louder than combination (b,r) on more than one-half of the trials, we write $(a,p) > (b,r)$. Occasionally, ties occurred; that is, (a,p) was judged louder than (b,r) on exactly one-half of the trials. In that case $(a,p) = (b,r)$. The number of such ties is shown in Table 1 for each subject. To check for transitivity, ordinal numbers were assigned to the tone pairs such that whenever

$$(a,p) \geq (b,r) \text{ and } (b,r) \geq (c,s) \qquad \text{(Order)}$$

$$\text{then } N(a,p) \geq N(b,r) \geq N(c,s),$$

where N stands for a numerical assignment. If there are no violations of transitivity, then the pairs can be ordered. If, on the other hand, there are violations of transitivity, there will be instances in which it is not possible to maintain this assignment procedure. The number of such instances indicates the extent to which transitivity is violated. An ordering was sought that resulted in the minimal number of violations. Table 1 (column 2) presents the number of violations for the 5 subjects. The number of violations ranged from 7 for Subject B.E. to 21 for Subject S.E. Given the relatively low number of violations of transitivity for each subject, we can conclude that the transitivity requirement is essentially satisfied.[1]

To test the double-cancellation requirement, all possible cases in which violations of double cancellation could occur were examined. It should be noted that, because ties were possible, some of the antecedent conditions for double cancellation involved ties. Antecedent conditions containing ties are denoted as weak tests; those that do not involve ties are denoted as strong tests. The reason for this distinction is that ties in the antecedent condition are unlikely to reflect equality in the sensory domain but, rather, lack of precision in our measurements. Therefore, an antecedent condition containing an equality may turn into either a strong test or a nontest, depending on the resolution of the inequality. For each of the subjects, the number of cases that met the antecedent conditions for double cancellation was computed. This number is presented in column 3 of Table 1 for weak tests and in column 4 for strong tests. Columns 5 and 6 give the percentage of failures for each of these two tests, respectively. Table 1 shows that double cancellation fails in only about 2% of the strong tests and 4% of the weak tests. This suggests that judgments of loudness satisfy to a reasonable degree both transitivity and double cancellation, and that the data matrix is additive.

Loudness Scale Values

Schneider (1980b) developed a nonmetric program designed for comparisons of differences. This nonmetric program was modified (see Appendix) to determine scale values for each set of eight stimuli. The program assigns projection values to each of the stimuli so that whenever $(a,p) \geq (b,r)$, then $P(a)+P(p) \geq P(b)+P(r)$. Of course, it is not possible to do this perfectly for data containing an error component; hence, the program minimizes the number of disagreements between the predicted paired comparisons and the obtained comparisons. Formally, the program minimizes the quantity, $G = N_d/(N_r-N_t)$, where N_d is the number of disagreements, N_r is the number of comparisons, and N_t is the number of cases in which the predicted comparisons are indeterminate; that is, $P(a)+P(p) = P(b)+P(r)$. Values of G are shown in column 7 of Table 1.

These values of G can be used to estimate the degree to which the projection values obtained from this program actually represent interval scale measurement. To accom-

Table 1

			Double Cancellation							
			Tests		Failure				p	
Subject	Ties	Transitivity Violations	Weak	Strong	Weak	Strong	G	Estimated CM	2 kHz	5 kHz
B.E.	39	7	1814	1460	4.2%	2.7%	.024	.998	.25	.28
N.E.	51	15	1976	1454	5.5%	2.3%	.036	.996	.34	.34
S.E.	54	21	1824	1258	5.8%	2.1%	.036	.996	.25	.36
E.D.	47	11	1878	1470	4.1%	2.5%	.020	.998	.28	.35
S.U.	77	21	2279	1478	4.0%	1.2%	.027	.997	.27	.23

Note—G is the goodness of fit index, CM is coordinate metric recovery, and p is the exponent as a function of sound intensity.

plish this, the index of coordinate metric recovery (*CM*) was estimated (see Appendix). *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 using these techniques are the true values known, but the Appendix shows how this can be estimated given the number of stimuli and the value of *G*. Hence, if the estimated value of *CM* is sufficiently high, the point coordinates can be properly regarded as representing interval scale measurement. Table 1 presents estimates of *CM* for all 5 subjects. Note that *CM* is estimated to be above .995 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.

Figure 1 plots the projection values as a function of decibels of sound pressure for each subject. Since the projection values for each frequency are unique only up to translation and a common expansion, they have been normalized so that the 2-kHz stimuli have the same range (1.0), and so that the lowest stimulus for both frequencies (-48 dB) has a coordinate value of 0. (Recall that the intensity value of this stimulus is so low that it is effectively a null stimulus. A break in the abscissa is used to call attention to this fact.) Note that the points are positively accelerated in these coordinates, with the degree of acceleration varying from listener to listener. Note also that the range of 5-kHz projection values and the position of those values relative to the 2-kHz stimuli vary from person to person.

The positive acceleration found in these coordinates is consistent with a power function representation for loudness. If loudness is a power function of intensity, then the projection values should be linearly related to stimulus intensity raised to a power, that is, $P_2 = a \cdot k_2 \cdot I^{p(2)} + b_2$ and $P_5 = a \cdot k_5 \cdot I^{p(5)} + b_5$, where $k_2 \cdot I^{p(2)}$ represents the loudness of the 2-kHz tones and $k_5 \cdot I^{p(5)}$ is the corresponding loudness for the 5-kHz tones. An iterative least squares procedure was used to find the best values of a, k, p, and b for both the 2- and 5-kHz functions. The curves in Figure 1 represent the best-fitting power functions to the coordinate values. Figure 1 shows that a power function representation for loudness provides a good fit to the data. The last two columns in Table 1 show the exponents for these power functions. Exponents relating loudness to sound intensity for the 2-kHz stimuli range from .25 to .34; those for the 5-kHz stimuli range from .23 to .35.

Intrasubject Reliability

For each subject, every 9 sessions constituted a replication of the basic experiment; that is, over the course of 9 sessions, each subject judged each of the 784 comparisons exactly once. In order to look for possible changes over time, projection values were obtained from the nonmetric program for each of the eight replications. With the exception of the lowest stimulus value (represented by a square), straight lines connect the projection values for a given replication in Figure 2. Again, because the projection values are unique only up to translation and to expansion by a common factor, the projection values in Figure 2 were adjusted so that the projection values for the 2-kHz stimuli had an average range of 1.0, and so that the average value for the lowest stimulus in both cases was 0.0. Figure 2 shows that the individual replications overlap to a considerable extent. The squared correlation coefficients between the average projection values in Figure 2 and the projection values obtained in Figure 1 exceed .991 in all cases.

Figure 2 suggests that the variability observed across replications is random. If it is random, then the r^2 values between projection values obtained for consecutive replications should be the same as the r^2 values between projection values obtained from replications more widely separated in time; that is, the average r^2 for Replications 1 and 2, 2 and 3, 3 and 4, ..., 7 and 8 should be the same as the average for Replications 1 and 4, 2 and 5, ..., 5 and 8. If, on the other hand, the underlying loudness scale was changing in a systematic way, the r^2 for consecutive replications should be greater than the r^2 for replications more widely separated in time. Figure 3 plots the mean r^2 between replications as a function of the number of intervening replications. (Mean r^2s were computed by first averaging rs using Fisher's transformation [see Hays, 1973] and then squaring the mean r.) Hence, the number 0 on the abscissa represents the consecutive replications (1 and 2, 2 and 3, etc.), and number 3 represents the replications separated by exactly three replications (1 and 5, 2 and 6, 3 and 7, 4 and 8). (The functions are not plotted for replications separated by more than three replications, since the number of r^2 values entering into the average becomes too small.) Figure 3 shows no indication of any trend.

Equal Loudness Contours and Intersubject Variability

To illustrate the extent of intersubject variability and to relate the results of this study to earlier ones on additivity of loudnesses across critical bands, equal loudness contours were constructed from the functions shown in Figure 1. For each subject, the loudness value of a 48-dB 2-kHz tone was determined from the function in Figure 1 and added to the loudness value of a 48-dB 5-kHz tone, also determined from the functions in Figure 1. This sum represents the loudness of a (48,48) combination of 2- and 5-kHz tones. From the functions in Figure 1, we also determined other combinations of intensities of 2- and 5-kHz tones, which also produced the same total loudness as that produced by the (48,48) combination. These paired intensities, then, determine an equal-loudness contour. Equal-loudness contours were determined for each of the 5 subjects for a (48,48) standard and also for standards at (54,54), (60,60), and (68,68).

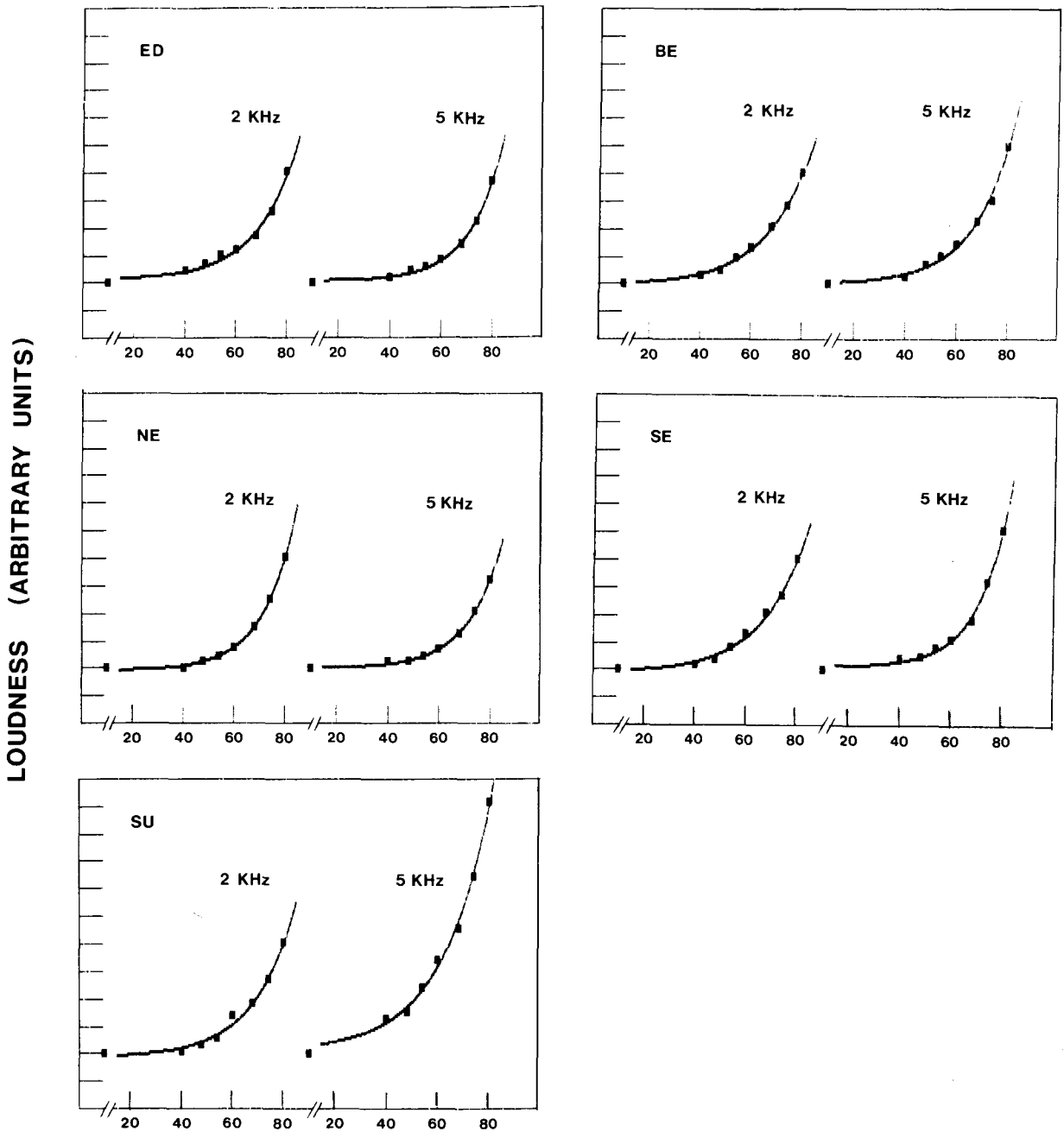

Figure 1. Individual subjects' loudness values as a function of decibels of sound pressure for 2- and 5-kHz tones derived from comparisons of the loudnesses of two-tone combinations. The curves represent the best-fitting power functions.

The calculated equal-loudness curves are depicted in Figure 4. Each contour specifies the tradeoff between 2- and 5-kHz tones that produce the same total loudness. Also shown in Figure 4 are the equal-loudness contours for Subjects S.U. and N.E. plotted on the same coordinates. As Figure 4 shows, there are striking individual differences in equal-loudness contours, sometimes exceeding 10 dB. These differences reflect differences in the loudness functions for the 2- and 5-kHz stimuli across subjects (see Table 1 and Figure 1).

DISCUSSION

Additivity and Scale Validity

Fletcher and Munson (1933) proposed that the loudness of a tonal complex, in which the tones were widely spaced

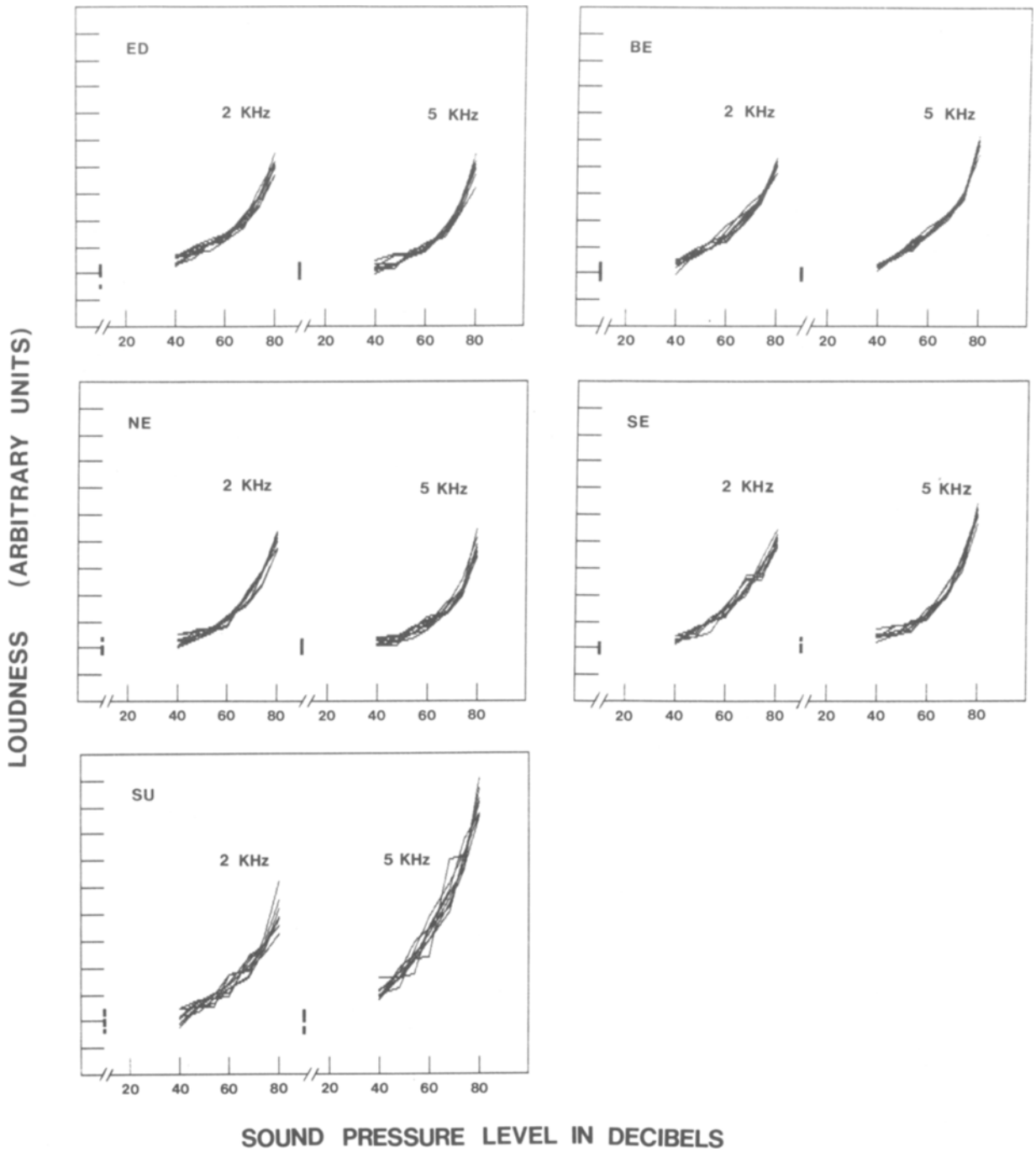

Figure 2. Loudness values as a function of decibels of sound pressure for 5 subjects. Eight replications are shown for each subject. The loudness values for the seven highest intensities within a replication are connected by straight lines. The loudness values for the below-threshold intensity are denoted by filled squares.

in frequency, was the sum of the loudnesses of the individual components. This simple hypothesis, however, is difficult to test if we suspect that our measures of total loudness and/or of the loudnesses of the individual tones are biased in a nonlinear fashion. For example, Marks (1978b) had subjects give magnitude estimates of the total loudness of a two-tone combination consisting of 2- and 5-kHz tones. These magnitude estimates were shown to be inconsistent with the notion that total loudness was equal to the sum of the loudnesses of the two tones. Thus, if these magnitude estimates accurately reflect total loudness, we must conclude that the loudness of widely separate tones are not additive, a conclusion directly opposite to the hypothesis of Fletcher and Munson (1933).

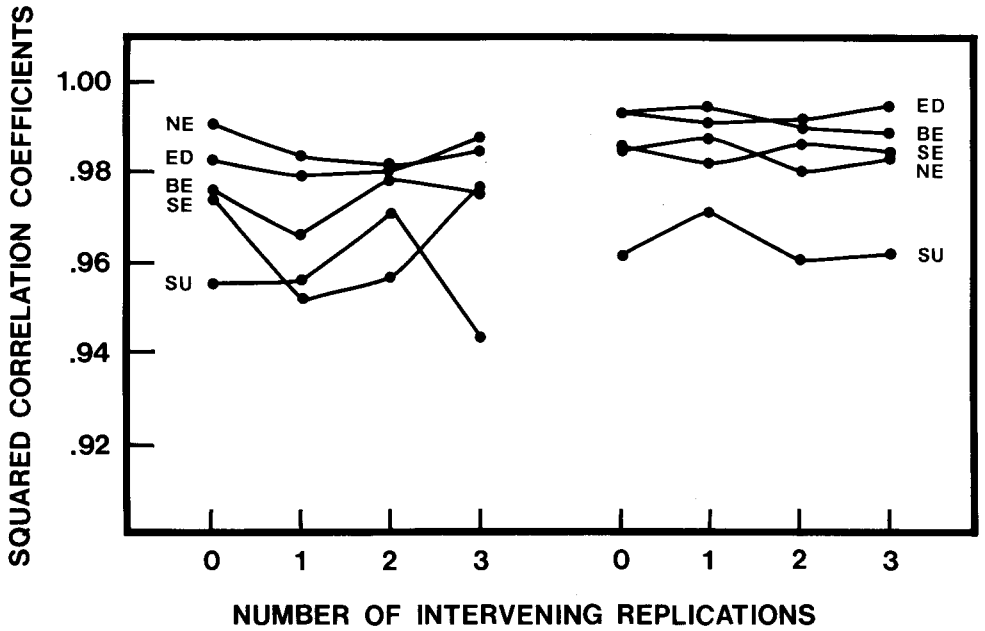

Figure 3. Averaged squared correlation coefficients between replications for each of 5 subjects at two frequencies as a function of the number of intervening replications.

Rather than conclude that the additivity assumption was false, Marks (1978b) proposed that the magnitude estimates of total loudness were biased. Marks transformed the magnitude estimates by raising them to the 1.33 power, which resulted in transformed estimates that were consistent with additivity. Such a transformation is consistent with Rule and Curtis's (1982) two-stage model of magnitude estimation, in which sensation is a power function of intensity and numerical estimates are a power function of sensations. Hence, either magnitude estimates are biased or additivity does not hold. In general, it is difficult to test for additivity if we believe that our response measures might be biased. This point is perhaps best illustrated in a paper by Algom and Marks (1984), in which they considered five different ways in which the judgments of individual subjects could be transformed to be consistent with different indicators of binaural additivity.

The present experimental paradigm avoids the problem of numerical biases by using a paired comparison paradigm. If we assume that subjects can accurately report which of two sounds is louder, we can use these paired comparisons to test the assumption of additivity by determining the extent to which the comparisons satisfy the double-cancellation condition. If loudness is additive across critical bands, then violations of double cancellation should be infrequent. This is, indeed, the case in this experiment.

Unfortunately, a statistical test of double cancellation is not possible in the present experiment because the error theory necessary for such a test is lacking. Without an adequate error theory, we cannot determine the extent to which the observed violations of double cancellation reflect (1) variability in the judgments, or (2) true violations of additivity. It should be noted that other designs do permit statistical tests of additivity. For example, Anderson (1970, 1974) devised methods to test for additivity when quantitative response measures of the joint effect of two stimuli are obtained in a factorial design. Carterette and Anderson (1979) have shown that it is possible, under the appropriate circumstances, to apply these tests even when the response measure is monotonically distorted. Further research, using these or other methods (see Falmagne, 1976), is needed in order to provide a statistical basis for evaluating additivity across critical bands.

A failure to detect departures from additivity does not, however, preclude the possibility that a nonadditive process is involved. Until we know how sensitive our tests are to reasonable departures from additivity (that is, until we can estimate the power of these tests), we should be cautious in concluding that an additive process characterizes the data. Nonadditive processes can, under some circumstances, generate data that appear to be additive. More powerful tests, or more sensitive experimental paradigms, may be required to uncover such nonadditive processes.

Although the lack of an appropriate error theory precludes a statistical test of double cancellation, we can show that the failure rate for double cancellation in this experiment is smaller than that found in some factorial experiments in which statistical tests failed to detect any departures from additivity. Marks (1979a), in a factorial design, had subjects give graphic rating estimates of the fused loudness of 1-kHz tones simultaneously presented

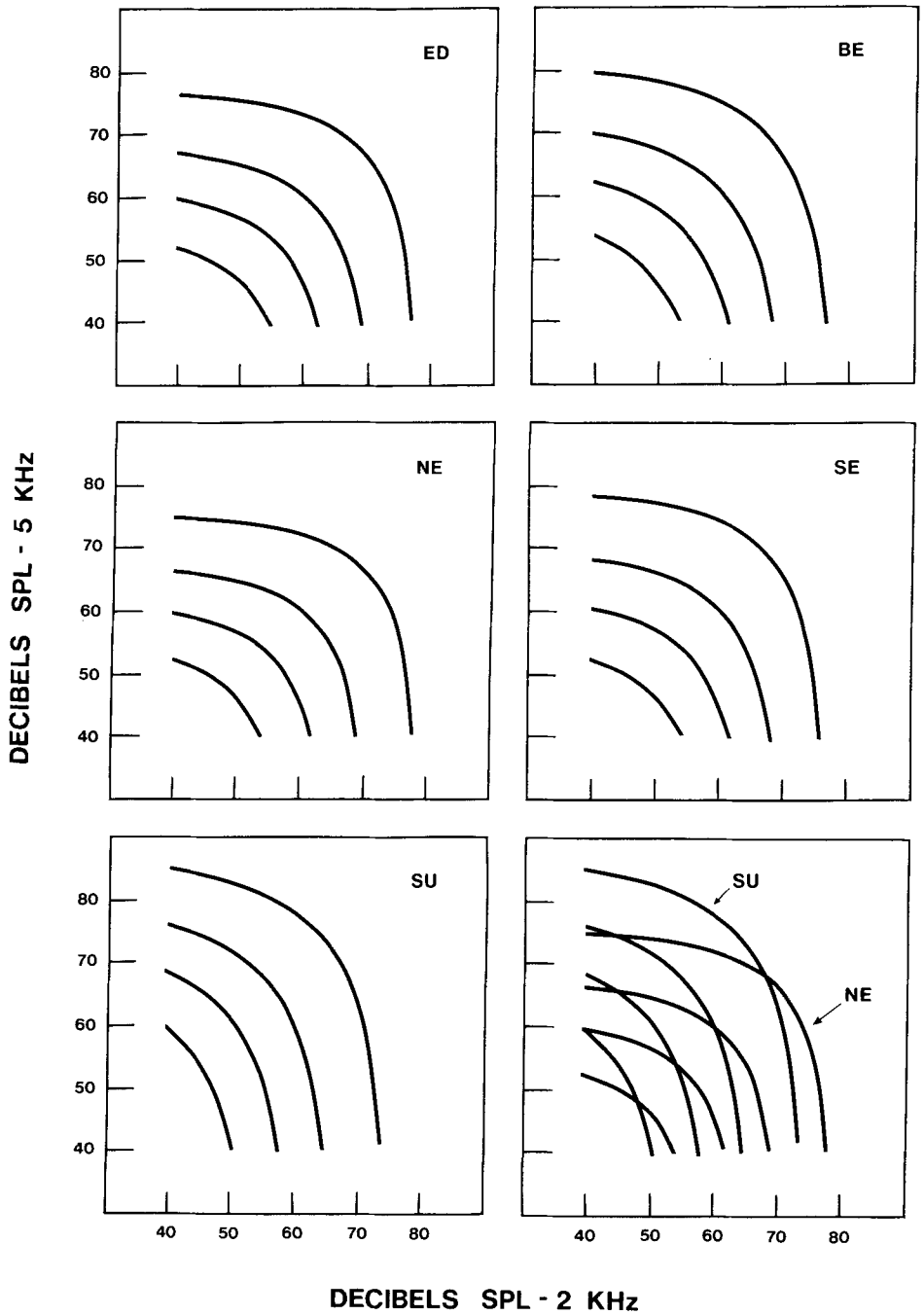

Figure 4. Equal-loudness curves constructed from the functions in Figure 1 for 5 subjects. Each curve specifies the combination of SPLs of the 2- and 5-Khz tones that sound equally loud (see text). The lower right-hand panel shows the contours for 2 subjects in the same coordinates.

to the two ears. He also had subjects give both magnitude and graphic rating estimates to the total loudness of these same tone pairs presented sequentially to the two ears. Statistical analyses of the judgments in the latter two experiments revealed no departures from additivity. A statistical analysis of the monotonically transformed judgments of the first experiment also failed to detect departures from linearity. Marks concluded, as a result, that loudness was additive for both simultaneous and seque tial presentations. The number of violations of double ca cellation in any one of these experiments can be dete mined from the rank order of the tone pairs. For ea experiment, the rank order of the tone pairs was dete mined from the data presented in Figures 1, 6, and 8 the Marks (1979a) paper. (Tied ranks were employ when two or more tone pairs appeared to have the san

value along the ordinate.) Hypothetical paired comparisons were generated from this rank order and tested for double cancellation. The failure rates for weak tests of double cancellation were 4.5%, 5.4%, and 5.2% for the three experiments, respectively. The corresponding failure rates for strong tests of double cancellation were 3.9%, 4.8%, and 4.6%. These failure rates are higher than those found in the present experiment. This suggests that departures from additivity in the present experiment are likely to be small for tones widely separated in frequency.

Loudness Scales and Sensory Processes

In the present experiment, loudness values were sought for 2- and 5-kHz tones such that whenever $(a,q) > (b,r)$, then $L(a)+L(q) > L(b)+L(r)$. In other words, loudness values were derived with respect to a hypothesized sensory process, namely, the summation of loudnesses across critical bands. Because the comparisons themselves supported an additive model (in the sense that they satisfied the testable axioms of additive conjoint measurement), these loudness values define the sensory representation of the stimuli within this *additive* process, and the relationship between these values and stimulus intensities represents the sensory transformation used in this additive process. In other words, the additive process reveals the scale values and the sensory transformation. Thus, in Mark's (1974) terms, discovery of a psychosensory relation constrains a psychophysical one.

Loudness values derived from magnitude estimates of single stimuli, however, are not as rigorously constrained. The only constraint on the judgments is that they should be monotonic with intensity. This provides no simple way to "correct for" any nonlinear biases that may exist in magnitude estimation experiments and no strong way in which to restrict the permissible transformation allowed on these estimates. Indeed, Krantz (1972) has argued that magnitude estimation and cross-modality matching yields loudness values that are unique only up to power transformations. Hence, for Krantz, sensory exponents are relative rather than absolute. The constraints imposed by loudness additivity are far narrower.

Since an additive process provides such strong restrictions on permissible transformations, it is interesting to compare loudness scales derived from this process with loudness scales derived from other, equally strict sensory processes. Several investigators have attempted to determine loudness scales from judgments of binaural loudness. Levelt, Riemersma, and Bunt (1972) had subjects judge which of two binaural tones was louder (the right- and left-ear intensities of the tones were varied). They showed that the judgments satisfied the requirements for binaural loudness additivity, and constructed loudness scales for the left and right ears of 2 subjects (but see Gigerenzer & Strube, 1983). These loudness scales were found to be power functions of stimulus intensity with exponents in the range of 0.20 to .31. Algom and Marks (1984) and Marks (1978a, 1979a, 1980), using magnitude estimation of binaural loudness, have found power functions for loudness with exponents ranging from .25 to .40. Note that these values are reasonably close to those found here for summation across critical bands. Hence, it appears that the same loudness scale may underlie two different sensory processes involving summation.

Scales derived from other kinds of sensory comparisons are not in agreement with those found employing sensory sums. For example, scales involving sensory differences typically result in exponents that are approximately half those found for sensory sums (see Marks, 1979b). It should be noted, however, that these scales are also based on a well-defined judgmental process, namely, loudness difference, since comparisons of loudness differences have been shown to satisfy the conditions for a positive difference structure (see Krantz et al., 1971) for both group data (Schneider, Parker, & Stein, 1974) and individual subjects (Schneider, 1980a). Thus, different judgmental processes require different loudness scales, as Marks (1979b) has argued.

Marks (1979b) has noted that magnitude estimation experiments result in exponents close to those found in sensory sum-experiments. I would be inclined to regard this congruence as fortuitous rather than as an indication of the greater validity for one scale than another (see also Popper, Parker, & Galanter, 1986). Typically, departures from the "expected" 0.3 for magnitude estimation value are often attributed to other factors. For example, Algom and Marks (1984) speculated that the low values of the loudness exponents found when intensity and duration are varied in a single experiment were due to the difficulty of the task. Other factors, such as the number of different tones being judged (Schneider, Wright, Edelheit, Hock, & Humphrey, 1972) or the stimulus range (Teghtsoonian, 1973), will affect the value of the exponent. Given that the value of the exponent can be affected by all of these factors, it becomes more and more difficult to declare one value as being most valid when magnitude estimation of single stimuli is involved. Furthermore, when a process such as summing or differencing is involved, the derived loudness scales represent, uniquely up to affine transformations, how sound intensities entering into the process are actually transformed during neural processing. Hence, it becomes more meaningful to speak of these scales as representing a stage in sensory processing, since it is the structure of the judgments and not their numerical values which require a power function transformation of stimulus intensity. In straightforward magnitude estimation experiments, the structure of the judgments requires only a loudness function that is monotonic with intensity. Therefore, it is difficult to decide exactly what stage of processing such judgments represent.

Intersubject Differences

Figure 2 shows that the underlying sensory representations for the 2- and 5-kHz tones were reasonably stable across replications within subjects. Figures 1, 2, and 4, however, suggest that there is considerable variation

across subjects. Exponents for both 2- and 5-kHz tones range from about .24 to .35. Note also, that the relative loudnesses of 2- and 5-kHz tones vary considerably across subjects. Recall that the lowest stimulus intensity for both tones was effectively below threshold. In Figure 1, this intensity has been assigned the same loudness value (0). Hence, horizontal cuts in Figure 1 can be used to determine the intensity values of 2- and 5-kHz tones that are equally loud. Using this procedure, we can see, for example, that a 74-dB 2-kHz tone is equal in loudness to a 72-dB 5-kHz tone for Subject S.E. while a 74-dB 2-kHz tone is equal in loudness to a 57-dB 5-kHz tone for Subject S.U. It is important to note that comparably large individual differences in equal-loudness contours for 2- and 5-kHz stimuli are found using traditional methods. For example, Ross (1967), using a matching procedure found individual differences on the order of 20 to 25 dB among subjects when 2- and 5-kHz stimuli were being matched. One of his 3 subjects (F.P., Table IV) matched an 80-dB 2-kHz tone to a 67.5-dB 5-kHz tone, while another (L.L., Table III) matched a 79-dB 2-kHz tone to a 95.5-dB 5-kHz tone. Thus, the large intersubject differences in equivalently loud stimuli found in this study are not peculiar to the paired comparison task but, rather, reflect true individual differences among subjects. Individuals vary not only in their growth rates for loudness, but also in the relative loudnesses of 2- and 5-kHz tones. The combination of these two factors can lead to a wide variation in contours for equally loud two-tone combinations (see Figure 4).

It is interesting to note that the range found for the individual exponents for both 2- and 5-kHz stimuli is comparable to that observed by Algom and Marks (1984) and Levelt et al. (1972) for binaural summation, once the response biases were factored out. (Algom & Marks actually employed five different ways to factor out response biases. The one that produced the minimum variability, but employed the strongest assumptions is the one reported here.) Algom and Marks used magnitude estimates of monaural and binaural loudness to determine, for individual subjects, the average decibel difference, N, between equally loud monaural and binaural stimuli. They showed that (1) if loudness is the same power function in both ears, and (2) if binaural summation is perfect, then the loudness exponent, p, equals $\log 2/(N/10)$. Using this equation to estimate individual exponents, they found exponents that ranged from .27 to .40, a range not greatly different from the one found here. It is interesting to note, however, that to obtain this small range it was necessary to assume additivity. If, instead, the magnitude estimates were nonlinearly transformed to produce the best fit to additivity (the procedure followed by Marks, 1978b), the individual exponents ranged from .30 to .595, a variation substantially larger than found in the present experiment.

Both the results of the present experiment and those of Algom and Marks (1984) and Schneider (1980a) point to sizable individual differences in loudness where loudness is derived from either summation or difference judgments. Furthermore, in two of these experiments (Schneider, 1980a, and the present experiment), use of a paired comparison design eliminated any numerical bias. Thus, with every effort made to reduce sources of intersubject variability due to non-sensory sources, we are still left with significant intersubject differences. This suggests that not all of the intersubject variability found in magnitude estimation experiments (Green & Luce, 1974; Luce & Green, 1978; J. C. Stevens & Guirao, 1964) is due to response biases, but that some portion is due to individual differences in sensory coding.

Intrasubject Variability

In the present study and in Schneider's (1980a), although there were substantial intersubject differences, the sensory representations appeared to be reasonably stable within subjects. This is to be contrasted with the data from magnitude estimation, where loudness functions have been known to drift over days (Schneider & Lane, 1963; Wanschura & Dawson, 1974). Such within-subject variation in magnitude estimation experiments could be due to two sources: (1) changes in sensory encoding, or (2) changes in response bias. Since the paired comparison design minimizes response bias, we would expect little intrasubject variation using this design if the sensory encoding for a particular sensory process (summation or difference) remains stable over time. This, indeed, appears to be the case for the two studies in which this design has been applied to individual subjects.

If, in addition, it could be shown that stimulus range does not affect the loudness function in a paired comparison design, although it does have an effect in other designs, we could be reasonably confident that the sensory encoding for a particular sensory process is a fixed component of the sensory system, and therefore is unaffected by non-sensory variables. If, on the other hand, range effects occurred using the paired comparison design, we would have to conclude that the encoding process was highly flexible and sensitive to a number of factors, not all of which are sensory. The resolution of this issue awaits further experimentation.

REFERENCES

ALGOM, D., & MARKS, L. E. (1984). Individual differences in loudness processing and loudness scales. *Journal of Experimental Psychology: General*, **113**, 571-593.

ANDERSON, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, **77**, 153-170.

ANDERSON, N. H. (1974). Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Vol. II. Psychophysical judgment and measurement* (pp. 215-298). New York: Academic Press.

CARTERETTE, E. C., & ANDERSON, N. H. (1979). Bisection of loudness. *Perception & Psychophysics*, **26**, 265-280.

CURTIS, D. W., ATTNEAVE, F., & HARRINGTON, T. L. (1968). A test of a two-stage model of magnitude judgment. *Perception & Psychophysics*, **3**, 25-31.

CURTIS, D. W., & RULE, S. J. (1972). Magnitude judgments of brightness and brightness difference as a function of background reflectance. *Journal of Experimental Psychology*, **95**, 215-222.

FAGOT, R. F., STEWART, M. R., & KLEINKNECHT, R. E. (1975).

Representations for biased numerical judgments. *Perception & Psychophysics*, **17**, 309-319.

FALMAGNE, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, **83**, 65-79.

FLETCHER, H., & MUNSON, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, **5**, 82-108.

GIGERENZER, G., & STRUBE, G. (1983). Are there limits to binaural additivity of loudness? *Journal of Experimental Psychology: Human Perception & Performance*, **9**, 126-136.

GREEN, D. M., & LUCE, R. D. (1974). Variability of magnitude estimates: A timing theory analysis. *Perception & Psychophysics*, **15**, 291-300.

HAYS, W. L. (1973). *Statistics for the social sciences*. New York: Holt, Rinehart & Winston.

HUBERT, L. (1976). Seriation using asymmetric proximity measures. *British Journal of Mathematical & Statistical Psychology*, **29**, 32-52.

KENDALL, M. G., & BABINGTON-SMITH, B. (1939). On the method of paired comparisons. *Biometrika*, **31**, 324-345.

KRANTZ, D. H. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology*, **9**, 168-199.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., & TVERSKY, A. (1971). *Foundations of measurement: Vol. 1. Additive and polynomial representations*. New York: Academic Press.

LEVELT, W. J. M., RIEMERSMA, J. B., & BUNT, A. A. (1972). Binaural additivity of loudness. *British Journal of Mathematical & Statistical Psychology*, **25**, 51-68.

LUCE, R. D., & GREEN, D. M. (1978). Two tests of a neural attention hypothesis for auditory psychophysics. *Perception & Psychophysics*, **23**, 363-371.

LUCE, R. D., & TUKEY, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, **1**, 1-27.

MARKS, L. E. (1974). *Sensory processes: The new psychophysics*. New York: Academic Press.

MARKS, L. E. (1978a). Binaural summation of the loudness of pure tones. *Journal of the Acoustical Society of America*, **64**, 107-113.

MARKS, L. E. (1978b). PHONION: Translation and annotations concerning loudness scales and the processing of auditory intensity. In J. J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 7-31). Hillsdale, NJ: Erlbaum.

MARKS, L. E. (1979a). Sensory and cognitive factors in judgments of loudness. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 426-443.

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

MARKS, L. E. (1980). Binaural summation of loudness: Noise and two-tone complexes. *Perception & Psychophysics*, **27**, 489-498.

POPPER, R., PARKER, S., & GALANTER, E. (1986). Dual loudness scales in individual subjects. *Journal of Experimental Psychology: Human Perception & Performance*, **12**, 61-69.

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.

RULE, S. J., CURTIS, D. W., & MARKLEY, R. P. (1970). Input and output transformations from magnitude estimation. *Journal of Experimental Psychology*, **86**, 343-349.

RULE, S. J., LAYE, R. C., CURTIS, D. W. (1974). Magnitude judgments and difference judgments of lightness and darkness: A two-stage analysis. *Journal of Experimental Psychology*, **103**, 1108-1114.

SCHARF, B. (1970). Critical bands. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. 1, pp. 157-202). New York: Academic Press.

SCHNEIDER, B. (1980a). Individual loudness functions determined from direct comparisons of loudness intervals. *Perception & Psychophysics*, **27**, 493-503.

SCHNEIDER, B. (1980b). A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika*, **45**, 357-372.

SCHNEIDER, B., & LANE, H. (1963). Ratio scales, category scales, and variability in the production of loudness and softness. *Journal of the Acoustical Society of America*, **35**, 1953-1961.

SCHNEIDER, B., PARKER, S., & STEIN, D. (1974). The 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.

SLATER, P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika*, **48**, 303-312.

STEVENS, J. C., & GUIRAO, M. (1964). Individual loudness functions. *Journal of the Acoustical Society of America*, **36**, 2210-2213.

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

WANSCHURA, R., & DAWSON, W. (1974). Regression effect and individual power functions over sessions. *Journal of Experimental Psychology*, **102**, 806-812.

YOUNG, F. W. (1970). Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika*, **35**, 455-473.

ZWICKER, E., FLOTTORP, G., & STEVENS, S. S. (1957). Critical bandwidth in loudness summation. *Journal of the Acoustical Society of America*, **29**, 548-557.

NOTE

1. Finding the minimum number of violations of transitivity is the same as finding the minimum number of reversals of the comparisons that would be necessary to satisfy transitivity. Slater (1961) refers to the minimum number of reversals necessary to achieve transitivity as the minimum number of inconsistencies in a set of paired comparisons. Thus, we are not referring to the number of circular triads, which is another common way of evaluating transitivity (Kendall & Babington-Smith, 1939). As Hubert (1976) points out, determining the set of rank orders that minimizes inconsistencies is, in general, not a trivial task. However, if the number of inconsistencies is small, as it was in this case, it can be done by hand.

APPENDIX

Schneider (1980b) developed an algorithm to search for the best one-dimensional representation for a set of stimuli whose psychological "distances" have been compared. This is equivalent to determining the location of each stimulus (its coordinate value) along a straight line segment such that the absolute difference between the coordinate values of any two stimuli represent their psychological distance. Coordinate values are sought such that whenever $(a,b) > (x,y)$, then $P(b)-P(a) > P(y)-P(x)$. Initially, the stimuli are arbitrarily assigned locations (coordinate values) along the line segment. The program begins by taking the first stimulus and moving it alternately to the left and right along the line segment until either the number of discrepancies between predicted and obtained comparisons (G) is reduced or the distance of the point from its original position on the line segment exceeds a criterial value. If G is reduced, the stimulus is moved to the location that produced the reduction. If G is not reduced before the boundaries are exceeded, then the stimulus remains at its current position. Thus, the program attempts to move the stimulus to a location that reduces the number of discrepancies, G. It then moves on to another stimulus along the

line segment and repeats this process until there is no further improvement in G.

The additive conjoint version of the program attempts to locate the 2- and 5-kHz stimuli along two orthogonal dimensions such that whenever $(a,p) > (b,q)$, then $P(a)+P(p) > P(b)+P(q)$. Again it starts with arbitrary starting locations for both 2- and 5-kHz stimuli and adjusts the locations of each stimuli serially until G is minimized.

The ability of the program to recover a configuration was tested using the Monte-Carlo procedures described by Young (1970) and used by Schneider (1980b). Briefly, comparisons of sums were determined from random starting configurations for the 2- and 5-kHz stimuli. With no error term added to the coordinate values, the average value of G for eight point configurations was .000, with an average index of coordinate metric recovery of .999. Thus, the program was able to recover the original configurations with an excellent degree of accuracy. To obtain estimates of metric recovery from values of G, progressively larger error terms were added to the coordinate values before the sums were computed, following the procedure developed by Young (1970) and used by Schneider (1980b). Thus, the relation between G and CM was determined. The estimates of CM found in this paper were based on these simulations.

(Manuscript received September 4, 1986;
revision accepted for publication August 28, 1987.)

Announcement

Third Symposium on Progress in Research on Brain and Behavior
"Advances in Research on Cerebral Laterality Effects"
Toledo, Ohio
April 8-9, 1988

The Third Symposium on Progress in Research on Brain and Behavior, sponsored by the University of Toledo Department of Psychology and the University of Toledo Alumni Foundation, will be held in the Driscoll Center for Continuing Education on the University of Toledo campus on April 8 and 9, 1988.

The theme for this symposium is Advances in Research on Cerebral Laterality Effects. Invited papers will be presented on advances in our knowledge of laterality effects with topics including research in the areas of anatomy, electrophysiology, neuropsychology, and behavior.

The participants for this symposium are M. P. Bryden, A. S. Gevins, C. R. Hamilton, C. Hardyck, J. B. Hellige, M. Kinsbourne, S. M. Kosslyn, J. Levy, W. F. McKeever, D. Molfese, J. Sergent, J. Ward, S. Witleson, and F. Wood.

For registration and other information, contact Fred Kitterle, Department of Psychology, University of Toledo, Toledo, Ohio 43606 (Phone: (419)537-2722).