Perception & Psychophysics
2000, 62 (1), 137-151

Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias

ROBERT L. WEST
University of Hong Kong, Hong Kong

and

LAWRENCE M. WARD and RAHUL KHOSLA
University of British Columbia, Vancouver, British Columbia, Canada

We report seven experiments in which subjects were trained to respond with numbers to the loudness of 1000-Hz pure tones according to power functions with exponents of 0.60, 0.30, and 0.90. Subjects were then presented with stimuli from other continua (65-Hz pure tones or 565-nm lights varying in amplitude) and were asked to judge the subjective magnitude of these stimuli on the same numerical scale. Stimuli from the training continuum were presented, with feedback, on every other trial in order to maintain the trained scale. Except for the 0.90 scale, subjects readily learned the predetermined scales and were able to use them to judge the non-training stimuli with group results consistent with those usually reported. Also, in contrast to the usual magnitude estimation results, these results produced extremely low levels of intersubject variability. We argue that such learned scales can be used as "rulers" for measuring perceived magnitudes, according to a common unit.

S. S. Stevens established direct scaling as one of the primary tools of psychophysics by amassing an internally consistent, nomothetic network of scaling results, so consistent and elegant that many of the scales he created are now broadly accepted worldwide. For example, Stevens' use of a power function with an exponent of 0.60 to describe the relationship between sound pressure and loudness for 1000-Hz tones is the standard for the System Internationale d'Unites (International Organization for Standardization, 1959). Central to this interlocking set of results was S. S. Stevens's demonstration that magnitude estimation (ME) results for many modalities could be described using the power law (sometimes called *Stevens's law*):

$$R = a S^m, \tag{1}$$

where R is the subjective magnitude of a sensory experience measured by a direct scaling technique (e.g., median or geometric mean magnitude estimation), S is the stimulus magnitude, a is a constant representing the unit of the scale, and m is a constant that varies across sensory continua. Moreover, exponents of power functions fitted to cross-modality matching (CMM) results can be successfully predicted from the ME exponents for the continua involved (S. S. Stevens, 1975), resulting in an internally coherent mathematical structure.

However, although a considerable amount of evidence indicates that, as a first order approximation, subjects do obey the power law (see Bolanowski & Gescheider, 1991, and S. S. Stevens, 1975, for reviews), and exponent values vary considerably across individuals within the same experiment (e.g., Algom & Marks, 1984; Logue, 1976; Luce & Mo, 1965; Marks & J. C. Stevens, 1965; Rule & Markley, 1971; Wanschura & Dawson, 1974) and across time within individuals (Logue, 1976; Marks, 1991; M. Teghtsoonian & R. Teghtsoonian, 1983). The fact that this occurs despite averaging across multiple responses for each stimulus level indicates that subjects systematically differ in their responses to the same stimuli—a fact that has long been known in psychophysics (see Poulton, 1989, for a review). Table 1 provides a sample of intersubject variability from various well-known laboratories. We employed two statistics to describe the variability of exponent values across individual subjects: the ratio of highest to lowest exponent values and the standard deviation of exponent values divided by the mean of the exponent values (i.e., the proportion of variation). According to Marks (1974a), individual ME experiments produce highest to lowest exponent ratios of at least 2:1. From the studies summarized in Table 1, this would seem to be accurate: Only one study produced a level of variability lower than this, and several were much higher. Note, though, that studies with more subjects will be more likely, by chance, to produce greater highest-to-lowest exponent ra-

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to L.M.W. The results of Experiments 1A and 4 were first reported at the annual meeting of the Psychonomic Society, November 1995, Los Angeles. The authors thank the reviewers for useful critiques. Correspondence should be addressed to R. L. West, Department of Psychology, Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6 Canada (e-mail: robert_west@carleton.ca) or to L. M. Ward, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, V6T 1Z4 Canada (e-mail: lward@cortex.psych.ubc.ca).

—Accepted by previous editor, Myron L. Braunstein

Table 1
Variability in a Convenient Sample of ME and CCM Experiments

	SD/M	High/Low	Method	*N*	Stimulus	Study
1	0.286	2.75	ME	11	loudness	Stevens & Guirao (1964)
2	0.436	n.a.	ME	32	loudness	M. Teghtsoonian & R. Teghtsoonian (1983)
3	0.388	n.a.	ME	35	loudness	M. Teghtsoonian & R.Teghtsoonian (1983)
4	0.290	3.951	ME	8	loudness	Algom & Marks (1990)
5	0.293	2.296	ME	11	loudness	Algom & Marks (1990)
6	0.444	3.32	ME	8	loudness	Ward (1982)
7	0.446	6	ME	8	brightness	Ward (1982)
8	0.186	1.6	ME	10	loudness	Hellman & Meiselman (1988)
9	0.274	2.267	ME	6	heaviness	Luce & Mo (1965)
10	0.231	1.746	ME	6	loudness	Luce & Mo (1965)
11	0.328	3.368	CMM	20	duration to loudness	Lilienthal & Dawson (1976)
12	0.347	3.808	CMM	20	loudness to duration	Lilienthal & Dawson (1976)
13	0.392	2.435	CMM	5	loudness to line length	Zwislocki (1983)
14	0.326	2.4	CMM	10	duration to loudness	Ward (1975)

Note—*SD/M*, the standard deviation of exponent values divided by the mean of the exponent values; High/Low, the ratio of the highest to lowest individual exponent; ME, magnitude estimation; CMM, cross-modality matching; n.a., not applicable.

tios. Therefore, this measure should be interpreted accordingly (Table 1 also lists the number of subjects in each study).

While it is possible to construe such individual differences as differences in the way that subjects perceive the same stimuli, this requires inferring larger individual differences than seem reasonable on the basis of sensory system differences. To understand this discrepancy, it is important to distinguish between matching processes that occur in a magnitude scaling experiment and matching processes that occur naturally. As Zwislocki (1991) points out, magnitude matching is a process that occurs frequently in everyday life. For example, to throw a rock at a target, the weight of the rock and the distance to be thrown must be combined and matched to the appropriate combination of force and launch angle. This type of matching is generally an unconscious process. In contrast, it can be argued that magnitude scaling techniques involve a conscious effort to produce an internally consistent set of matches over an entire set of stimuli. In our opinion, a primary cause of intersubject variability is that, instead of relying solely on a naturally occurring matching process, such as that mediating the throwing of a rock, subjects have a tendency to impose higher level cognitive constraints on the matching process. An obvious example is the imposition of subjects' personal experience with the magnitude of numbers on their responses in an ME task (e.g., engineers sometimes use unusually large or unusually small numbers). Yet, a quick glance at Table 1 reveals that CMM experiments fare little better than ME experiments, indicating that the problem is not limited to issues surrounding the use of numbers. We suggest that this is because subjects also impose more general ad hoc constraints early in the scaling process. For example, in terms of the response range, subjects may develop a fixed set of convenient responses early in the experiment, thus defining a fixed response range. To the subject, relying on a particular response range might provide the illusion of having a more well-defined scale, whereas in reality, it can arbitrarily influence the power function exponent value (R. Teghtsoonian, 1971). Anecdotally, when S. S. Stevens was developing ME, he found that, to his surprise, subjects provided less variable psychophysical functions when they were given less information. For example, the inclusion of a reference tone increased variability rather than lowering it. Subjects performed best when they were encouraged not to think about the scale but to just respond with whatever response *felt* natural (S. S. Stevens, 1975, p. 28). It seems as if the added information only encouraged subjects to construct idiosyncratic cognitive constraints, rather than relaxing and accessing the naturally occurring process.

The process of magnitude matching can be represented in the following way (Marks, 1991),

$$M(S) = R, \tag{2}$$

where S is the stimulus amplitude, R is the response magnitude, and M is the function relating them. The M function can then be decomposed into an initial perceptually based function, P, that is the same (or highly similar) across healthy, normal individuals, followed by a function, C, representing cognitively imposed constraints (cf. Curtis, Attneave, & Harrington, 1968):

$$M(S) = C[P(S)] = R. \tag{3}$$

The idea that the "true" psychophysical function (P) is obscured during the scaling process by the introduction of a second function (C) has a long history within psychophysics and is usually referred to as the problem of response bias. In general, the approach to this problem has been to attempt to minimize the effect of C. For example, in free magnitude estimation, C is minimized by providing a situation with as few constraints as possible and by encouraging subjects to respond in a natural way (e.g., see

Zwislocki & Goodman, 1980). As noted above, better results are often achieved by providing an environment that does not prompt subjects to cognitively augment the scaling process.

Another approach has been to measure the effect of C in individual subjects and then mathematically compensate for it (e.g., Berglund, 1991; Marks et al., 1988). This most commonly involves modeling both P and C as power functions (see Curtis et al., 1968, for a justification of this assumption). Typically, a standardizing procedure is employed in which subjects use ME to scale a common stimulus dimension, such as line length. Assuming that P is the same across subjects for this task, the differences between subjects in terms of C can be derived. By further assuming that C remains constant across tasks, these results can be used to "undo" the effects of C in subsequent ME tasks for different stimulus modalities. A third approach to minimizing the effect of C on scaling results is magnitude matching (J. C. Stevens & Marks, 1980). In magnitude matching, stimuli from two different modalities are presented (usually in alternation) during a single ME session, and subjects are told to respond using the same response scale for both modalities. For individual subjects, the responses for the two modalities can then be plotted against each other, effectively canceling out the effect of C (again, assuming that C was the same for both modalities). Finally, it is possible to avoid the use of direct responses and to infer the psychophysical function using methods such as nonmetric scaling (e.g., Schneider, 1980) or additive conjoint scaling (e.g., Schneider, 1988).

In this study, we took a different approach to dealing with the unwanted C function. Instead of trying to minimize the effect of C, we attempted to control and manipulate C so that it was the same in all of our subjects. Although providing subjects with a little information (e.g., a modulus) has proven detrimental (S. S. Stevens, 1975), we propose that providing subjects with a lot of information is actually the best way to control C. More specifically, we propose that C is cognitively penetrable and can be controlled to a high degree under the right conditions. Many studies have been done on the factors that influence C (see Poulton, 1989, for a review), but very little work has been done on subjects' own ability to control and shape C. The purpose of this study was to examine the extent to which the C function could be calibrated across subjects and the extent to which such calibration could reduce intersubject variability.

Our approach was based on Ward's (1992) argument that subjects should be taught to map their subjective perceptions of stimulus magnitudes to a response continuum according to a standardized scale (selected for its mathematical desirability and learnability) before participating in a direct scaling task. Applied to the power law, this involves subjects first learning to respond to a standard set of stimuli according to a power function with a predetermined exponent value (West & Ward, 1994). In theory, such training would provide subjects with the same, or equivalent, sets of cognitive constraints, give or take some error. If subjects could then generalize this set of constraints to other stimuli, on which they had not been trained, it would eliminate or greatly reduce the use of idiosyncratic and/or ad hoc constraints (see Baird, Kreindler, & Jones, 1971, for similar reasoning). West and Ward (1994) termed this general approach *constrained scaling*, and we will adhere to this nomenclature in what follows.

Several studies have already been performed indicating the feasibility of training subjects to judge sensation magnitude on a standard scale. King and Lockhead (1981), Koh and Meyer (1991), Koh (1993), West and Ward (1994), and Marks, Galanter, and Baird (1995) have all provided evidence that, given feedback, subjects can learn to respond to sensory stimuli according to power functions with a given exponent, quickly and with a high degree of accuracy. In addition, West and Ward (1994) and Marks et al. (1995) addressed the issue of using learned scales to investigate basic psychophysical phenomena. West and Ward trained subjects to respond to 1000-Hz tones according to a power function with an exponent of 0.60, then removed the feedback and randomly presented 65-, 100-, 1000-, and 8000-Hz tones. Subjects, responding according to the learned scale, produced results consistent with the research on equal loudness contours (e.g., see Marks, 1974b; Ward, 1990). Specifically, subjects' responses indicated that they found the 65-, 100-, and 8000-Hz tones to be significantly less loud than the 1000-Hz tones at equal sound pressures. Marks et al., using a similar methodology, trained subjects to respond to 500-Hz binaural tones according to power functions with exponents of 0.3, 0.6, and 1.2. Following this, the feedback was removed, and subjects were presented with binaural 500-Hz tones alternated with monaural 500-Hz tones. For the data averaged across subjects, they found that loudness summation was unaffected by the training. Thus, both West and Ward (1994) and Marks et al. (1995) provided evidence that using a learned scale does not alter basic psychophysical phenomena.

However, it is obvious that having subjects respond according to a learned scale will alter the interpretation and meaning of the resulting exponent values. We will defer a discussion of the validity of exponents obtained using constrained scaling until after the presentation of our data. However, at this point, we would note that we view the two approaches of minimizing response bias and manipulating response bias as complementary. For example, it can be argued that a scale that is difficult to learn is an unnatural scale—a topic that we explore in this paper. The paper by Marks et al. (1995) also provides an excellent example of how these two approaches can be combined to deal with specific theoretical issues.

EXPERIMENT 1A

Similar to West and Ward (1994), we used feedback to train subjects to respond to 1000-Hz tones according to

a variant of S. S. Stevens's sone scale (i.e., a power function with an exponent of 0.60). After the learning phase, subjects entered the test phase, in which the same task was repeated but with feedback absent on every second trial. All subjects completed two test phase conditions: one in which the tones on the no-feedback trials were 65-Hz tones, and a control condition in which the tones on the no-feedback trials were 1000-Hz tones. For the trials with feedback, the same set of 1000-Hz tones was used. The purpose of this was to provide subjects with an environment that promoted to the maximum extent their use of the learned scale. We expected, on the basis of well-established psychoacoustical results, that power function exponents for the 65-Hz tones would be larger than those for the 1000-Hz tones (see Marks, 1974b, and Ward, 1990, for reviews). This would reflect the fact that the rate of increase in loudness, relative to sound intensity, is greater for 65-Hz tones than for 1000-Hz tones. In terms of relative loudness, because the absolute threshold for 65-Hz tones is higher than that for 1000-Hz tones, 1000-Hz tones are already reasonably audible at the threshold for 65-Hz tones. Following from the power law, the difference in loudness narrows as the sound pressure increases, due to the higher exponent for the 65-Hz tones (see R. Teghtsoonian, 1971). An analysis of predicted exponents for the 65-Hz tones is made in the Results and Discussion section.

Method

Subjects. Six volunteers were paid to participate. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task. None had participated in a scaling experiment before.

Apparatus. The subjects were seated in a dimly lit sound-attenuation chamber, and the tones were monaurally presented to them through high-quality headphones. A program written in Visual Basic for DOS allowed the subjects to enter their responses by using a mouse to manipulate a specially designed scroll bar that appeared on a computer monitor facing the subjects. Using the mouse, the subjects could move the scroll bar cursor by increments of 1, 10, or 0.1 and access all numbers from 0 to 99.9 with a precision of one decimal place. A text box displayed the value at which the scroll bar was set. To receive a tone, the subjects used the mouse to click on a button marked "PLAY TONE." After the subjects had indicated their response by using the scroll bar, they used the mouse to click on a button marked "OK," at which point their response displayed in the text box was replaced with feedback or the message "NO FEEDBACK" for the no-feedback trials. The feedback was accurate to five digits, as many as would fit in the box, to encourage the subjects to use the full three digits of precision available to them.

Stimuli. The tones were produced by a custom-built sound generator controlled by the computer. They were 1000-Hz or 65-Hz sinusoids of 1-sec duration, with rise/fall times of 6 msec, presented at levels ranging from 33 to 99 dB SPL.

Procedure. In the training phase, the subjects were trained to respond to different amplitudes of a 1000-Hz pure tone according to the power function

$$R = 16.6\,P^{0.60}, \tag{4}$$

where R is the correct response, and P is the sound pressure in dynes/cm^2. The exponent was set to 0.60 to make the scale consistent with S. S. Stevens's (1975) sone scale (which is an excellent candidate for a standard scale of loudness). The multiplicative constant was set to 16.6 so that $R = 100$ would result from a P equivalent to approximately 100 dB SPL. Before the training, the subjects were instructed that they would be learning to use a particular number scale to judge the loudnesses of pure tones. The level of the training tones was varied by randomly selecting a decibel value between 33 and 99 for presentation on each trial, resulting in a response range of approximately 1 to 94. Because individual stimuli were selected at random, there were seldom repeats of particular sound levels. This procedure contrasts with the more common procedure in which relatively few stimulus levels are selected for repeated presentation. However, it was not possible to use only a few selected intensities in the training phase since their identity would have been revealed through the feedback. To maintain consistency, this procedure was also used in the no-feedback condition during the test phase. The subjects were given 100 learning trials prior to the test phase.

In the test phase, the subjects were instructed to use the scale they had just learned to judge the loudnesses of either 1000-Hz or 65-Hz test tones. The test tones were presented, in the manner described above, on no-feedback trials that were alternated with feedback trials on which 1000-Hz tones were presented. It was emphasized that, on all of these trials, the frequency of the tones was to be ignored and that responses should be based solely on perceived magnitude. Also, because of the method of presentation, some of the lower level 65-Hz tones in the test phase would have been perceived as requiring a response below $R = 1$. The subjects were informed that this would happen. The subjects were also told to respond with $R = 0$ for any tones they could not hear. In addition, the subjects were told that the loudest no-feedback tones might or might not be as loud as the loudest feedback tones. For both the 1000-Hz and the 65-Hz test trial sessions, the subjects completed 400 trials (with one rest break when desired). The order in which the subjects completed the two test trial sessions was counterbalanced, with each subject performing an additional 100 training trials between sessions.

Results and Discussion

For all of the 1000-Hz results, the logarithm of the response (R) was linearly regressed against the logarithm of the sound pressure (P) in dynes/cm^2 according to the following equation:

$$\log R = m \log (P) + \log (a) + e, \tag{5}$$

where m and a are constants, and e is error. Consistent with past research, all subjects were able to learn the scale reasonably well. For the combined 1000-Hz learning sessions, the subjects' individual exponent values indicated that they had learned the scale (0.56, 0.60, 0.59, 0.57, 0.59, and 0.60; $M = 0.59$, $SD = 0.02$; note that all means are rounded to two decimal places). Although there was a slight tendency for the subjects' exponent values to be below the prescribed value of 0.60, this was to be expected, since any errors in responding would, by statistical regression, bias the fit by making m lower (S. J. Rule, personal communication, April 1998). To estimate the effect of this, let Y be the logarithm of the correct responses according to Equation 4, and let X be the logarithm of the stimulus amplitudes. The correct value for m can then be expressed as SD_Y/SD_X. Now let Y' be the logarithm of the subjects' responses. Assuming the range of Y' is approximately the same as the range of Y, it should be the case that $SD_{Y'}$ is approximately equal to SD_Y.

Therefore, the correct value for m can be approximated by $SD_{Y'}/SD_X$. Following from least squares theory, the least squares estimate of the subjects' m values, m', is equal to $(SD_{Y'}/SD_X)r_{XY'}$. Substituting m for $(SD_{Y'}/SD_X)$, we can see that $m' = m\ r_{XY'}$, indicating that m' will always be smaller than m, unless $r_{XY'} = 1$ (S. J. Rule, personal communication, April 1998). In all of the following experiments, there was a tendency for the subjects' exponent values to be slightly lower than the values on which they were trained. In all cases, this can be reasonably accounted for by the explanation above.

1000-Hz test phase results. For the feedback trials, the mean exponent value was 0.54, the standard deviation divided by the mean was 0.045, and the highest-to-lowest exponent ratio was 1.11:1. For the no-feedback trials, the mean exponent value was 0.54, the standard deviation divided by the mean was 0.071, and the highest-to-lowest exponent ratio was 1.21:1. These results indicated no unusual effects due to the test phase procedure. Also, note that both feedback and no-feedback trials produced the same mean exponent value, indicating that the subjects were using the same scale to respond to all stimuli. This was also reflected in the consistency of results for the individual subjects (see Table 2 for the individual results).

65-Hz test phase results. For the 1000-Hz feedback trials, the mean exponent value was 0.55, the standard deviation divided by the mean was 0.066, and the highest-to-lowest exponent ratio was 1.09:1. Figure 1 displays the raw data of the individual subjects for the 65-Hz no-feedback trials. Recall that these data represent judgments of a stimulus continuum for which no training whatsoever had been provided. Because these results exhibited the well-documented effect of an increase in slope for stimuli near threshold (e.g., see S. S. Stevens, 1975), the function

$$\log R = m \log (S - T) + \log (a) + e, \qquad (6)$$

recommended by S. S. Stevens to correct for this effect, was fitted to these data. One interpretation of this function is that subjects use their functional threshold (T) as zero, which causes a distortion since the power law implicitly assumes that subjects will use the point associated with the physical absence of a stimulus as zero. Subtracting T from S corrects for this problem and provides an estimate of T. The results of this analysis are displayed in Table 3 and in Figure 1. The mean of the individual exponent values for the 65-Hz tones was 0.77, which, according to a t test, was significantly higher than the mean of the exponents for the 1000-Hz tones with which they were alternated ($p < .001$). Also, the estimates of T were remarkably stable, with 4 of 6 subjects having the same value of $T = 0.008$ dynes/cm^2 (Subjects 2 and 6 gave T values of 0.011 and 0.003 dynes/cm^2, respectively). In terms of intersubject variability, the standard deviation divided by the mean was 0.132, and the highest-to-lowest exponent ratio was 1.33:1. These results were substantially less variable than is typical in direct scaling (see Table 1).

Table 2
Test Trial Results Using 1000-Hz Tones for the Feedback and No-Feedback Trials in Experiment 1A

	Feedback Trials		No-Feedback Trials	
Subject	Exponent	R^2	Exponent	R^2
1	0.52	.82	0.51	.83
2	0.58	.83	0.61	.87
3	0.51	.80	0.52	.84
4	0.55	.90	0.54	.89
5	0.53	.88	0.50	.83
6	0.56	.88	0.55	.88
M	0.54		0.54	
SD	0.02		0.04	

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.60.

We also analyzed the 65-Hz data excluding the subjects' responses below $R = 1$, this time using Equation 5, reasoning that, since the subjects were not trained to respond below 1, they might have exhibited idiosyncratic tendencies for stimulus amplitudes in this range. Consistent with this idea, the results of this analysis displayed less variability than the Equation 6 analysis. The mean 65-Hz exponent was 0.70, the standard deviation divided by the mean was 0.115, and the highest-to-lowest exponent ratio was 1.42:1 (see Table 3). However, an F test for variability indicated that the difference in intersubject variability between the two analyses was not significant, and a t test indicated that the difference between the 65-Hz exponents from the two analyses was not significant. Also, a t test revealed that the Equation 5 mean exponent was still significantly higher than the mean exponent obtained from the alternating 1000-Hz trials ($p < .001$).

Finally, we compared the 1000-Hz and 65-Hz results in a way roughly equivalent to plotting matching functions from a magnitude matching experiment, in which subjects judge stimuli from two different continua (usually presented in alternation) within the same scaling task (J. C. Stevens & Marks, 1980). The only difference in our task was that the subjects first received training on one of the stimulus continua and then received feedback on this continuum during testing. As discussed in the introduction, assuming that response bias remains the same across stimulus continua, taking a ratio of the exponents from two such continua should cancel the effect of response bias and reveal the true ratio of the exponents. Taking the ratio of the 1000-Hz exponents to the 65-Hz exponents for individual subjects, we found even lower levels of intersubject variability. For the Equation 6 results, we obtained a mean exponent value of 0.72, a standard deviation divided by the mean equal to 0.116, and a highest-to-lowest exponent ratio of 1.38:1. For the Equation 5 results, the mean of the subjects' ratios was 0.79, the standard deviation divided by the mean was 0.086, and the highest-to-lowest exponent ratio was 1.28:1. In terms of the specific ratio values, it is important to note that the arithmetic mean of the exponent ratios is not uniquely defined. For example, the mean of subjects' exponent ratios is not the same as the ratio of

Table 3
Test Trial Results Using 1000-Hz Tones for the Feedback Trials and 65-Hz Tones for the No-Feedback Trials in Experiment 1A With Equations 5 and 6 Analyses

	Equation 5					Equation 6		
	1000-Hz Tones		65-Hz Tones		1000:65	65-Hz Tones		1000:65
Subject	Exponent	R^2	Exponent	R^2	Ratio	Exponent	R^2	Ratio
1	0.54	.83	0.70	.87	0.77	0.78	.90	0.697
2	0.59	.87	0.73	.75	0.81	0.87	.87	0.679
3	0.49	.84	0.56	.84	0.87	0.64	.89	0.765
4	0.57	.89	0.67	.90	0.85	0.66	.92	0.866
5	0.58	.83	0.75	.87	0.77	0.85	.90	0.682
6	0.54	.88	0.80	.85	0.68	0.85	.93	0.633
M	0.55		0.70		0.79	0.77		0.720
SD	0.04		0.08		0.07	0.10		0.083

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.60.

the means of subjects' exponent values (S. J. Rule, personal communication, April 1998). Therefore, interpretations of the mean ratio values should be made with this caveat in mind (we used the arithmetic mean in order to employ *t* tests and *F* tests on the results). In terms of intersubject variability, the lower variability for the exponent ratios suggests that the subjects maintained some idiosyncratic biases during the test phase, which were subsequently removed by taking ratios of the exponents.

Following Marks (1974b) and Ward (1990), the exponents for different sound frequencies can be approximated by the following equations:

$$F \leq 400 \text{ Hz: } m = 2\,[H + G\,(400 - F)] \quad (7)$$

$$F > 400 \text{ Hz: } m = 2\,(H), \quad (8)$$

where F is frequency, m is the exponent, and H and G are constants. Equation 7 describes a linear approximation to the relationship between frequency and exponent values for frequencies less than or equal to 400 Hz and can be rearranged into the more familiar $Y = AX + B$ form,

$$F \leq 400 \text{ Hz: } m = (-2G)F + [2H + 2G(400)]. \quad (9)$$

As Equation 9 illustrates, $-2G$ describes the increase in exponents for frequencies ≤400 Hz. Ward (1990) found a value for G equal to 0.0004. Using the average exponent found for the no-feedback 1000-Hz trials (0.54), it was determined that G was equal to 0.0003 when the mean 65-Hz exponent was 0.77 (the Equation 6 analysis) and 0.0002 when the mean 65-Hz exponent was 0.70 (the Equation 5 analysis). These estimates were close to Ward's

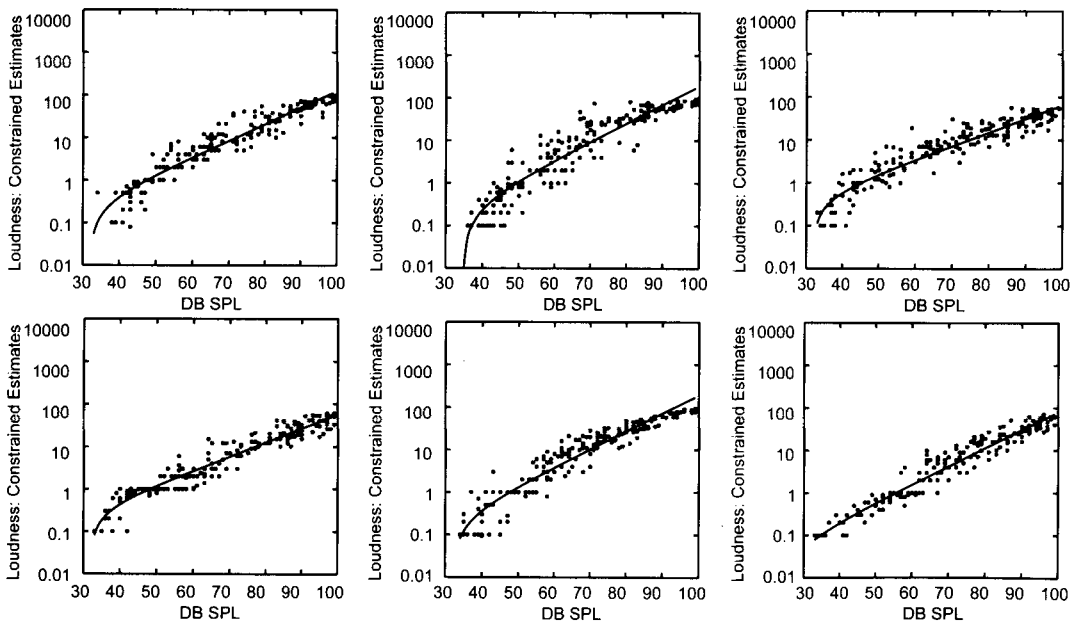

Figure 1. Experiment 1A (no-feedback). Fits of Equation 6 to the constrained estimates of the loudness of 65-Hz tones for each of 6 subjects.

(1990) estimates, illustrating that constrained scaling produces results consistent with established psychophysical methods. However, Marks (1974b), using equal loudness contours and the results of standard ME experiments from various laboratories (see Marks, 1974b), estimated G to be approximately 0.0009, which predicts a 65-Hz exponent of 1.14 (based on a 1000-Hz exponent of 0.54). The problem of obtaining a stable estimate of G is related to the problem of interlaboratory variability, which is discussed in the General Discussion section.

EXPERIMENT 1B

Experiment 1A produced unusually low levels of intersubject variability. Experiment 1B simply replicated Experiment 1A, with a few minor differences, to demonstrate that the Experiment 1A results were not an anomalous occurrence.

Method

Subjects. Seven volunteers, different from those in Experiment 1A, participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task.

Apparatus. The apparatus was the same as that in Experiment 1A.

Procedure. The learning phase procedure was the same as that in Experiment 1A, except that the subjects were given only 50 learning trials instead of 100, and the most intense stimulus level was 89 dB rather than 99 dB. This reduced stimulus range was correspondingly maintained during the test phase. To reflect this decrease in range, a marker was placed at $R = 50$, and the subjects were told that no tones would be louder than this. Also, during the test phase, the lowest stimulus level for the 65-Hz tones was made to be 40 dB in order to subjectively equalize the loudness ranges of the 65-Hz and 1000-Hz tones (based on the exponent values from the Equation 5 analysis in Experiment 1A). The subjects were instructed that the lowest levels of the 1000-Hz tones and the 65-Hz tones would be roughly the same, but to use responses of less than 1 if warranted. Also, the total number of test trials was reduced to 100 (i.e. fifty 1000-Hz trials and fifty 65-Hz trials).

Results and Discussion

The results were analyzed by fitting Equation 5 to the raw data. For the 1000-Hz learning phase trials, the mean of the individual exponent values was 0.56, the standard deviation divided by the mean was 0.112, and the highest-to-lowest exponent ratio was 1.35:1. For the 1000-Hz test phase trials, the mean of the individual exponent values was 0.56, the standard deviation divided by the mean was 0.075, and the highest-to-lowest exponent ratio was 1.23:1 (see Table 4). For the 65-Hz test trials, the mean of the individual exponent values was 0.67, the standard deviation divided by the mean was 0.096, and the highest-to-lowest exponent ratio was 1.37:1. Taking the ratio of the test phase 1000-Hz exponents to the 65-Hz exponents, the mean exponent ratio was 0.84, the standard deviation divided by the mean was 0.082, and the highest-to-lowest exponent ratio was 1.19:1 (see Table 4). As can be seen, both in terms of exponent values and intersubject variability, the results of Experiment 1B were highly consistent with the Equation 5 results of Experiment 1A.

Table 4
Test Trial Results Using 1000-Hz Tones for Feedback Trials and 65-Hz Tones for No-Feedback Trials in Experiment 1B With Equation 5 Analysis

	1000-Hz Tones		65-Hz Tones		1000:65
Subject	Exponent	R^2	Exponent	R^2	Ratio
1	0.57	.93	0.68	.91	0.84
2	0.55	.87	0.70	.84	0.78
3	0.60	.86	0.69	.71	0.88
4	0.57	.91	0.72	.77	0.79
5	0.59	.93	0.67	.76	0.89
6	0.51	.73	0.68	.86	0.74
7	0.49	.82	0.52	.69	0.94
M	0.56		0.67		0.84
SD	0.04		0.06		0.07

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.60.

EXPERIMENT 2A

Because the subjects in Experiments 1A and 1B were trained on the generally accepted exponent value of 0.60 for 1000-Hz tones, it was possible, although unlikely given the unusually low level of intersubject variability, that the subjects were merely responding naturally during the test phase. Another possibility was that something other than the use of a learned scale could account for the low level of intersubject variability, such as the restricted response range or the scroll bar that provided the subjects with a visual analog for their numerical judgments. If this was the case, then the learning trials and the feedback on the 1000-Hz tones during the test phase were actually not necessary. Experiment 2A was done as a control to assess the role of the learning trials and the feedback trials. In this experiment, the subjects made the same judgments as in Experiment 1A but without prior training and without feedback on the 1000-Hz tones. The result of this was a procedure very similar to a magnitude matching task (J. C. Stevens & Marks, 1980).

Method

Subjects. Six volunteers, different from those in the previous two experiments, participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task.

Stimuli and Apparatus. The stimuli and the apparatus were the same as those in Experiment 1A.

Procedure. The procedure was exactly the same as that in the 65-Hz test phase of Experiment 1A, except that no training phase preceded the test phase and no feedback was given on the 1000-Hz trials. Instead, the subjects were shown how to operate the scroll bar and were given several trials of practice, in the absence of stimuli, in producing responses over the entire range available. They were also given standard free absolute ME instructions emphasizing a natural match between sensation and number intensities (Zwislocki & Goodman, 1980). The instructions were adapted for the limited response range in that the subjects were told they were free to use any responses between 0 and 99.9. This limitation was attributed to a property of the computer interface. Also, the subjects were given the standard magnitude matching instructions to use the same scale to judge both frequencies of tone.

Table 5
Magnitude Matching Results Using 1000-Hz Tones and 65-Hz Tones in Experiment 2A With Equation 5 Analysis

	1000-Hz Tones		65-Hz Tones		1000:65
Subject	Exponent	R^2	Exponent	R^2	Ratio
1	0.50	.70	0.55	.58	0.90
2	0.23	.52	0.56	.71	0.41
3	0.35	.80	0.63	.65	0.56
4	0.32	.83	0.88	.75	0.37
5	0.51	.77	0.76	.56	0.67
6	0.66	.75	0.98	.72	0.68
M	0.43		0.73		0.60
SD	0.16		0.18		0.20

Results

As in Experiments 1A and 1B, the 1000-Hz data were analyzed according to Equation 5. The 65-Hz results were also analyzed according to Equation 5, excluding responses less than 1, in order to make the analysis comparable to the 65-Hz, Equation 5 analysis in Experiment 1A (excluding responses less than 1 slightly reduced intersubject variability relative to the same analysis including responses less than 1). Since the near-threshold deviation from the power law was not strongly evident, as it was in Experiment 1A, Equation 6 was not employed. The results are displayed in Table 5. The mean exponent value for the 1000-Hz tones was 0.43, and the mean exponent value for the 65-Hz tones was 0.73, replicating the standard result of a larger exponent value for 65-Hz tones (this difference was significant by *t* test, $p = .006$). In terms of intersubject variability, both the 1000-Hz exponents ($SD/M = 0.367$, highest/lowest exponent = 2.88) and the 65-Hz exponents ($SD/M = 0.244$, highest/lowest exponent = 1.77) were within the expected range for ME results (see Table 1). Also, an *F* test for variance indicated that the 65-Hz exponents were significantly more variable than the Experiment 1A, Equation 5, 65-Hz exponents ($p = .054$).

We next examined the relationship between the 1000-Hz exponent values and the 65-Hz exponent values. As in Experiment 1A, we did this by looking at the ratio of 1000-Hz exponents to 65-Hz exponents for each subject. The mean exponent ratio was 0.60, which a *t* test revealed to be significantly lower than the mean exponent ratio of 0.79 found in Experiment 1A, using Equation 5 ($p = .022$). In terms of intersubject variability, the standard deviation divided by the mean was 0.329, and the highest-to-lowest exponent ratio was 2.46:1. This was actually worse than the intersubject variability for the 65-Hz exponents alone. An *F* test for variance showed that this level of intersubject variability was significantly higher than that for the Equation 5 exponent ratios in Experiment 1A ($p = .018$).

EXPERIMENT 2B

As noted above, Experiment 2A had the same design as a magnitude matching experiment. Experiment 2B was also a control study but instead had the same design as a standard free absolute ME experiment.

Method

Subjects. Six volunteers, different from those in the preceding experiments, participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task.

Apparatus. The apparatus was the same as that in Experiment 1A, except that the scroll bar was removed and the subjects were able to type their responses into a text box using the computer keyboard.

Procedure. Experiment 2B followed the same procedure as that in Experiment 2A, except that the subjects typed a number in a text box rather than using the scroll bar. This allowed the subjects to use numbers as high or as low as they pleased. Also, in keeping with standard ME procedure, the subjects were presented with tones of only one frequency. In order to better compare the intersubject variability to that found in the ME literature, 1000-Hz tones were used (1000-Hz tones are far more common than 65-Hz tones in the ME literature). As in Experiment 2A, the subjects were given standard free absolute ME instructions, emphasizing a natural match between sensation and number intensities (Zwislocki & Goodman, 1980). Also, in keeping with the standard free absolute ME procedure, the subjects were told to use any responses that seemed natural. The subjects completed 100 trials.

Results and Discussion

The results were analyzed according to Equation 5 and are displayed in Table 6. The mean of the subjects' individual exponents was 0.64, which was higher than the value of 0.43 found for the 1000-Hz tones in Experiment 2A. A *t* test revealed that this difference was marginally significant ($p = .06$). In terms of intersubject variability, the standard deviation divided by the mean was 0.409, and the highest-to-lowest exponent ratio was 3.05:1. These results agree well with the values in Table 1 and can be considered typical of ME results. The psychophysical functions are displayed in Figure 2 at the same resolution as the constrained scaling results in Figure 1.

Overall, the results of Experiments 2A and 2B indicate that the use of a learned scale and feedback played an important role in producing the low levels of intersubject variability found in Experiments 1A and 1B. Using the same equipment and stimuli as those in Experiments 1A and 1B, neither magnitude matching nor free magnitude estimation could produce levels of intersubject variability as low as constrained scaling. Also, magnitude matching and free magnitude estimation produced dif-

Table 6
Free Magnitude Estimation Results Using 1000-Hz Tones in Experiment 2B With Equation 5 Analysis

Subject	Exponent	R^2
1	0.71	.88
2	0.33	.83
3	0.75	.62
4	1.00	.80
5	0.34	.66
6	0.71	.75
M	0.64	
SD	0.26	

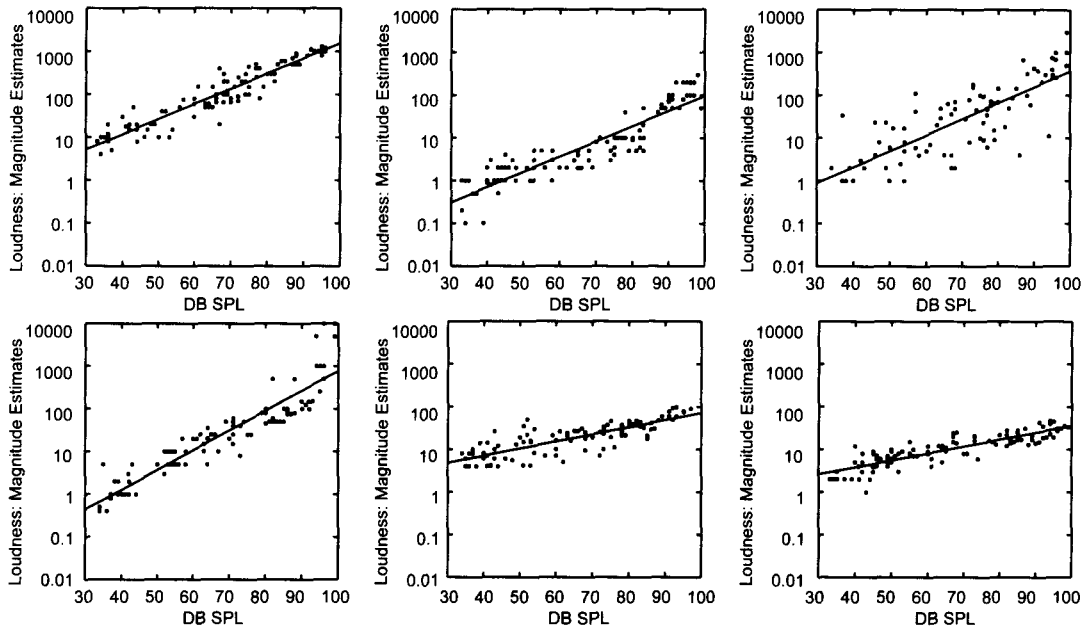

Figure 2. Magnitude estimates of the loudness of 1000-Hz tones by 6 subjects. The equation of the line fitted to the data is given by Equation 5.

ferent estimates of the exponent value for the 1000-Hz tones, and magnitude matching and constrained scaling produced different estimates of the ratio of 1000-Hz exponents to 65-Hz exponents. These different results reflect the more general finding that seemingly innocuous differences between methods can produce significantly different results (Marks, 1974a; S. S. Stevens, 1955), a topic we will return to in the General Discussion section.

EXPERIMENT 3A

Experiments 1A, 1B, 2A, and 2B showed that training and feedback on a learned scale considerably reduced intersubject variability. However, it could be argued that training and feedback will only help when they guide subjects toward the true psychophysical function (i.e., for 1000-Hz tones, a power function with an exponent of approximately 0.60). In this sense, the benefits of training and feedback would be related to reducing bias rather than controlling it (i.e., the training would help subjects to achieve a more natural matching function). Thus, when subjects are trained on an "unnatural" exponent value, the effect could be to increase confusion and raise intersubject variability.

In Experiment 3A, we investigated this possibility by training subjects to respond to 1000-Hz tones according to an exponent of 0.30. Similar to Experiments 1A, 1B, and 2A, the novel stimuli were 65-Hz tones. If constrained scaling truly does provide subjects with a generalizable learned scale, then training subjects on a different scale should alter the results for the no-feedback continuum in a predictable way. In this case, the most straightforward prediction would be that, since 0.30 is half of 0.60, the mean 65-Hz exponent should be approximately half of that found in Experiments 1A and 1B.

In terms of learnability, Marks et al. (1995) used feedback to train subjects to respond according to an exponent of 0.30 to 500-Hz tones (like 1000-Hz tones, 500-Hz tones are characterized by an exponent of approximately 0.60; see Equation 8 above). Taking pooled geometric means within and across subjects, they found a group exponent of 0.31. For individual subjects' exponent values, the standard deviation divided by the mean exponent was 0.068, and the highest-to-lowest exponent ratio was 1.27:1. This pattern of results was very close to the results obtained when Marks et al. trained subjects to respond to 500-Hz tones according to an exponent of 0.60. In this case, the group exponent was 0.58, the standard deviation divided by the mean was 0.060, and the highest-to-lowest exponent ratio was 1.38:1.

Method

Subjects. Seven volunteers participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task.

Apparatus. The apparatus was the same as that in Experiment 1A, except that the response scale went from 0 to 9.99 and the subjects could respond with up to two decimal places of accuracy.

Procedure. The procedure was identical to that in Experiment 1A, except that the subjects received only 50 learning trials and performed only 100 test phase trials, as in Experiment 1B. Also, instead of learning the 0.60 scale, the subjects learned a scale based on an exponent of 0.30. For power law scales of a given exponent, once one stimulus value is matched to a response value, the response values for all other stimuli are fixed. In this case, the response scale was determined by assigning the least intense 1000-Hz tone (33 dB) a value of 1, as was the case in Experiments 1A and 1B. By doing this, the upper bound of the response continuum was

Table 7
Test Trial Results Using 1000-Hz Tones for Feedback Trials and 65-Hz Tones for No-Feedback Trials in Experiment 1A With Equations 5 and 6 Analyses

	Equation 5					Equation 6		
	1000-Hz Tones		65-Hz Tones		1000:65	65-Hz Tones		1000:65
Subject	Exponent	R^2	Exponent	R^2	Ratio	Exponent	R^2	Ratio
1	0.27	.74	0.41	.77	0.64	0.38	.91	0.70
2	0.28	.86	0.47	.81	0.59	0.34	.80	0.83
3	0.30	.86	0.56	.76	0.53	0.41	.92	0.73
4	0.30	.83	0.42	.76	0.72	0.52	.93	0.59
5	0.26	.87	0.36	.87	0.71	0.39	.89	0.66
6	0.27	.70	0.37	.69	0.71	0.35	.88	0.75
7	0.26	.74	0.36	.78	0.71	0.44	.89	0.57
M	0.28		0.42		0.66	0.40		0.69
SD	0.02		0.07		0.07	0.06		0.09

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.30.

fixed at approximately 10. This produced a response scale compatible with the base 10 number system that subjects are more familiar with. The information content, or fidelity, of the 0.30 response scale was equalized to the previous 0.60 response scales by restricting the subjects to three-digit responses. For example, the maximum response value was 9.99 for the 0.30 response scale and 99.9 for the 0.60 response scale.

Results

All 1000-Hz results were analyzed by fitting the best line through the raw data according to Equation 5. For the learning trials, the mean exponent was 0.31, the standard deviation divided by the mean was 0.112, and the highest-to-lowest exponent ratio was 1.39:1. In terms of intersubject variability, this result was very similar to that found in Experiment 1B, which used the same number of training trials. For the 1000-Hz test trials, the mean of the individual exponent values was 0.28, the standard deviation divided by the mean was 0.071, and the highest-to-lowest exponent ratio was 1.192:1 (see Table 7).

The results from the 65-Hz test trails were first analyzed according to Equation 6, to account for the near-threshold deviation from the power law (see Experiment 1A). Individual subjects' raw data were analyzed separately using Statistica's Hook Jeeves nonlinear estimation procedure. Except for 1 anomalous subject (Subject 7), the estimates of the subjects' functional thresholds (T in Equation 4) were very stable and highly consistent with those found in Experiment 1A. Respectively, they were 0.008, 0.007, 0.008, 0.006, 0.007, 0.003, and −0.045 dynes/cm^2. The very close similarity between these estimates of T and the estimates of T derived in Experiment 1A indicates that measurement of the functional threshold in this way is unaffected by training subjects on an a lower exponent value. The mean of the individual exponent values for the 65-Hz tones was 0.40, which, according to a t test, was significantly higher than the mean of the 1000-Hz exponents with which they were paired ($p < .001$). As predicted, this value was also very close to one half of the mean exponent value found in Experiment 1A using Equation 6 (i.e., 0.77/2 = 0.39). In terms of intersubject variability, the standard deviation divided by the mean was 0.153, and the highest-to-lowest exponent ratio was 1.55:1 (see Table 7).

As in Experiment 1A, the 65-Hz results were also analyzed according to Equation 5, excluding responses below $R = 1$, where the subjects had not been trained and where the near-threshold deviation from the power law occurred. For this analysis, the mean of the individual exponent values was 0.42, which again was significantly higher than the mean of the 1000-Hz exponents (by t test, $p < .001$). In terms of intersubject variability, the standard deviation divided by the mean was 0.173, and the highest-to-lowest exponent ratio was 1.57:1 (see Table 7).

Finally, we analyzed the data by taking the ratio of the 1000-Hz exponents to the 65-Hz exponents for each subject. Using the Equation 6 results for the 65-Hz tones, the mean ratio was 0.69, which was quite close to the value of 0.72 found in Experiment 1A using Equation 6 for the 65-Hz tones. The standard deviation divided by the mean was 0.133, and the highest-to-lowest ratio was 1.46:1 (see Table 7). Using the Equation 5 results for the 65-Hz tones, the mean ratio was 0.67, which t tests revealed was significantly lower than the mean ratio of 0.79 found in Experiment 1A using Equation 5 ($p < .001$) and the mean ratio of 0.84 found in Experiment 1B using Equation 5 ($p < .001$). In terms of intersubject variability, the standard deviation divided by the mean was 0.113, and the highest-to-lowest ratio was 1.36:1 (see Table 7). As in Experiments 1A and 1B, the fact that taking the ratio of the exponents lowered intersubject variability suggests that the subjects maintained some idiosyncratic biases that were subsequently removed by the exponent ratio analysis.

When the near-threshold deviation from the power law was taken into consideration (i.e., the Equation 6 analysis), the result of training the subjects to respond to 1000-Hz tones according to an exponent of 0.30 was essentially to produce 65-Hz exponent values half the size of those produced by the subjects trained to respond accord-

ing to an exponent of 0.60. The results excluding responses below $R = 1$ (i.e., the Equation 5 analysis) were not as clear-cut. Although the results were roughly in line with what was predicted, the mean ratio of 1000-Hz exponents to 65-Hz exponents was lower than that found in the Equation 5 results in Experiments 1A and 1B.

EXPERIMENT 3B

Having demonstrated that subjects can learn, and extend to novel stimuli, a power function scale based on an exponent value less than the canonical value of 0.60, we turned to exponents greater than 0.60. Using feedback, Marks et al. (1995) trained subjects to respond to 500-Hz tones according to an exponent of 1.20 but found that subjects produced a group exponent that was significantly different from 1.20 (the exponent was 1.11). In addition, relative to when the same subjects were trained on exponents of 0.30 and 0.60, the standard deviation divided by the mean was higher (0.110 relative to 0.068 and 0.060, respectively), and the highest-to-lowest exponent ratio was higher (1.60:1 relative to 1.27:1 and 1.38:1, respectively). Since these findings indicated that a power function based on an unnaturally high exponent value is more difficult to learn, we predicted a similar result in the present context.

Method

Subjects. Seven volunteers participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task.

Apparatus. The apparatus was the same as that in Experiment 1A, except that the response scale went from 1 to 1,000 and the subjects were not permitted to respond with decimals.

Procedure. The subjects were first trained to respond to 1000-Hz tones according to an exponent of 0.90. An exponent of 0.90 was chosen because it resulted in a response range of 1–1,000 when the response for lowest tone (33 dB) was set to 1. Thus, similar to the response ranges in Experiments 1A, 1B, and 3A, this response range was compatible with the base 10 number system with which subjects are familiar. The information content, or fidelity, of the 1–1,000 response scale was made equal to the scales used in Experiments 1A, 1B, and 3A by restricting the subjects to three-digit responses. However, in this case, the limitation to three-digit precision meant that the subjects could respond only to tones below $R = 1$, with a 1 or a 0. This situation was remedied by subjectively equalizing the bottom of the 65-Hz range to the bottom of the 1000-Hz range in the same way as in Experiment 2B. Also, the subjects were told that the least intense 65-Hz tone would be somewhere "around" $R = 1$ and to use a response of 1 for any audible 65-Hz tones requiring a response of 1 or less. Other than this, the procedure for Experiment 3B was identical to that for Experiment 3A.

Table 8
Test Trial Results Using 1000-Hz Tones for Feedback Trials and 65-Hz Tones for No-Feedback Trials in Experiment 3B With Equation 5 Analysis

	1000-Hz Tones		65-Hz Tones		1000:65
Subject	Exponent	R^2	Exponent	R^2	Ratio
1	0.79	.85	0.68	.55	1.17
2	0.76	.80	0.77	.67	0.98
3	0.74	.81	0.97	.86	0.76
4	0.74	.80	0.71	.56	1.04
5	0.91	.95	1.01	.83	0.90
6	0.48	.45	0.60	.55	0.80
7	0.82	.75	1.01	.74	0.81
M	0.75		0.82		0.92
SD	0.13		0.17		0.15

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.90.

Results and Discussion

As before, the 1000-Hz trials were analyzed by fitting Equation 5 to the raw data. For the learning trials, the mean exponent was 0.700, the standard deviation divided by the mean was 0.178, and the highest to lowest ratio was 1.71:1. For the 1000 Hz test trials, the mean of the individual exponent values was 0.75, the standard deviation divided by the mean was 0.177, and the highest-to-lowest exponent ratio was 1.90:1 (see Table 8). Although these results displayed considerably more intersubject variability than the results above, this appears to have been due primarily to intersubject differences in error, rather than systematic deviations from the feedback provided. As noted in the Results section of Experiment 1A, the effect of error on the estimate of subjects' exponent values can be predicted using the formula $m' = m\, r_{XY'}$, where m' is the least squares estimate of subjects' exponent values, m is the exponent value on which the feedback is based, and $r_{XY'}$ is the correlation between the logarithm of stimulus amplitude and the logarithm of subjects' responses (S. J. Rule, personal communication, April 1998). For the 1000-Hz test phase results, the Pearson correlation coefficient between the exponent values predicted by this formula and the subjects' actual exponent values was $r = .95$, indicating that differences in error rates could account for most of the intersubject variability. Also, the mean of the exponents predicted from the formula for m' was 0.79, which was quite close to the actual mean. We speculate that the difficulty some subjects experienced may have arisen from the fact that they were required to use large numbers.

The 65-Hz test trial results were analyzed in the same way as the 1000-Hz trials. The mean of the individual exponent values was 0.82, which, according a t test, was not significantly higher than the 1000-Hz mean exponent value of 0.75. The standard deviation divided by the mean was 0.209, and the highest-to-lowest exponent ratio was 1.68:1. Taking the ratio of the 1000-Hz exponents to the 65-Hz exponents produced a mean exponent ratio of 0.92, a standard deviation divided by the mean equal to 0.162, and a highest-to-lowest exponent ratio of 1.54:1 (see Table 8). However, t tests revealed that this ratio was significantly higher than the Equation 5 ratio found in Experiment 1A ($p < .05$) and possibly higher than the Equation 5 ratio found in Experiment 1B ($p = .097$). This was very interesting, since the Equation 5 ratio found in Experiment 3A was significantly lower than these two ratios. These results indicate a roughly linear trend for the Equation 5 ratio estimates to increase as the exponent

value for the learned scale is increased. Currently, we do not have an explanation for this. However, the fact that the Equation 6 results seem stable across different learned scales (see Experiment 3A) suggests that this effect may have been due to an interaction between the near-threshold deviation from the power law and the exponent values that the subjects were trained on.

Overall, the low level of intersubject variability found in Experiments 3A and 3B indicated that the effect of the training and feedback in Experiments 1A and 1B was not due to reducing unnatural response bias (i.e., by guiding the subjects toward a more natural matching function) but rather to controlling higher level nonperceptual constraints on the matching process. However, it is also important to note that the intersubject variability found in Experiments 3A and 3B was not as low as in Experiments 1A and 1B, which suggests that subjects may find an exponent of 0.60 for 1000-Hz tones to be more natural than an exponent of 0.30 or 0.90. In particular, these results and the results of Marks et al. (1995) indicate that subjects find unusually large exponents difficult to learn.

EXPERIMENT 4

Experiment 4 was done to investigate the use of constrained scaling for measuring sensation magnitudes from sensory modalities other than the training modality. The procedure was the same as that in the experiments above (i.e., subjects were trained using 1000-Hz tones), but the test modality was brightness instead of loudness. Since we trained subjects on the exponent that S. S. Stevens recommended for the loudness of 1000-Hz tones (approximately 0.60 using dynes/cm^2), we predicted that subjects would also produce the exponent that S. S. Stevens (1975) attained for brightness (approximately 0.33 using footlamberts to measure light intensity).

Method

Subjects. Eight volunteers participated for pay. All claimed to have normal hearing, and there was no evidence of hearing abnormalities during the task. None of the subjects reported any visual system abnormalities other than those corrected for by wearing glasses. The subjects who wore glasses wore them during the experiment.

Apparatus. The apparatus and the 1000-Hz tones were the same as those in Experiment 1A, except that the colors of the computer monitor screen were altered (primarily to dark red on a black background) and the luminance of the monitor was reduced to make the interior of the sound-attenuation chamber as dark as possible without making it too difficult for the subjects to see the screen. The light stimuli were produced in the form of dots, 6.5 mm in diameter and of approximately uniform luminance. They were positioned at eye level, directly in front of the subject. The light stimuli were six levels of luminance, equally spaced on a logarithmic scale: 0.013, 0.76, 0.430, 2.400, 13.800, and 79.400 fL. They were produced by a 565-nm wavelength LED embedded in diffusing plastic and controlled by varying the voltage across the LED. Light stimuli were presented for 1 sec, with rise/fall times of less than 1 msec.

Procedure. The learning procedure was identical to the procedure in Experiment 1A, except that the subjects were given only 50 learning trials, as in Experiment 1B. The testing procedure was also the same as that in Experiment 1A, except that the 65-Hz tones were replaced with the six luminance levels, which were presented randomly. The subjects were unaware that they were receiving only six luminance levels. The subjects performed 100 test trials, alternating between judgments of the loudness of 1000-Hz tones (with feedback) and the brightness of the dots of light (without feedback). There was thus an average of 8–9 presentations of each of the six luminance levels during the test trials—much fewer than in a typical ME experiment.

Table 9
Test Trial Results Using 1000-Hz Tones for Feedback Trials and Light (Brightness) for No-Feedback Trials in Experiment 4 With Equation 5 Analysis

	1000-Hz Tones		Light (Brightness)		
Subject	Exponent	R^2	Exponent	R^2	Ratio
1	0.56	.87	0.36	.89	1.57
2	0.61	.91	0.34	.99	1.83
3	0.55	.87	0.40	.89	1.38
4	0.48	.87	0.27	.92	1.79
5	0.60	.90	0.25	.91	2.44
6	0.50	.84	0.36	.94	1.39
7	0.46	.74	0.31	.93	1.48
8	0.52	.83	0.35	.97	1.47
M	0.53		0.33		1.67
SD	0.06		0.05		0.35

Note—The subjects were trained to respond to the 1000-Hz tones according to a power law function with an exponent of 0.60.

Results and Discussion

As in the previous experiments, psychophysical functions were fitted to the 1000-Hz data according to Equation 5. For the 1000-Hz learning trials, the mean of the individual exponent values was 0.56, the standard deviation divided by the mean was 0.094, and the highest-to-lowest exponent ratio was 1.29:1. For the 1000-Hz trials alternated with the light trials, the mean of the individual exponent values was 0.53, the standard deviation divided by the mean was 0.106, and the highest-to-lowest exponent ratio was 1:1.34 (see Table 9).

Equation 5 was also used for analysis of the light trials, since all luminance levels were well above threshold. Figure 3 shows the psychophysical functions for the mean responses of individual subjects for the light trials, as well as the best-fitting functions according to Equation 5. The mean of the individual exponent values was 0.33, exactly on the predicted value. Also, all of our subjects' individual exponents were close to the predicted value: The exponent values were 0.36, 0.34, 0.39, 0.27, 0.25, 0.36, 0.31, and 0.35 (see Table 9).

In terms of intersubject variability, the standard deviation divided by the mean was 0.152, and the highest-to-lowest ratio was 1.59:1. However, taking the ratios of the brightness exponents to loudness exponents failed to further reduce the intersubject variability, as in the experiments above. The mean of the exponent ratios was 1.67, the standard deviation divided by the mean was 0.212, and the highest-to-lowest exponent ratio was 1.77:1. This, however, was due almost entirely to the results of 1 subject (Subject 5), whose exponent ratio was more than 2 standard deviations from the mean of the exponent ra-

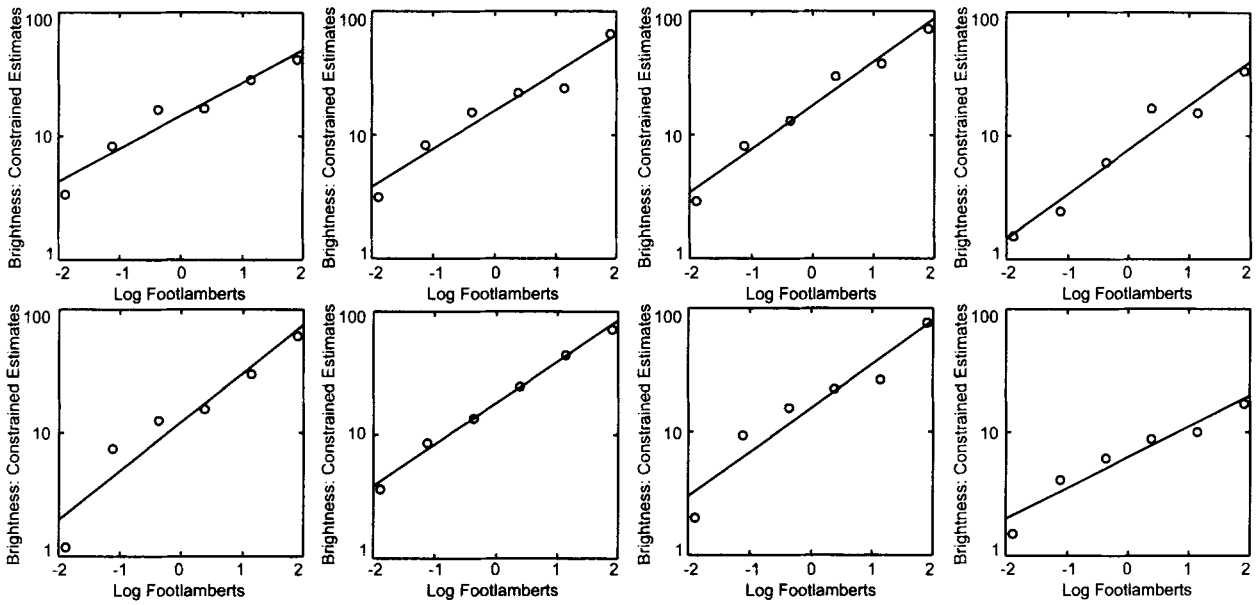

Figure 3. Fits of Equation 5 to the constrained estimates (no-feedback condition) of brightness for each of 8 subjects.

tios, whereas all other subjects' exponent ratios were closer than 1 standard deviation to the mean. Also, the exponent ratio of Subject 5 was well outside a confidence interval of 99.9%, set around the mean of the exponent ratios excluding Subject 5. Excluding this subject, the standard deviation divided by the mean of the ratios was 0.117, and the highest-to-lowest exponent ratio was 1.32:1.

We felt justified in removing Subject 5 because the very low intersubject variability produced by constrained scaling made it obvious that Subject 5's matching process was somehow different from the other subjects. This was unlike the results of the control experiments (Experiment 2A and 2B), in which there were no obvious outliers but rather an overall higher degree of intersubject variability. However, it is difficult to say why Subject 5 was different, since his loudness exponent and brightness exponent were not obvious outliers in the sense that the ratio of the two exponents was (his loudness exponent was at the high end of the loudness exponent range, and his brightness exponent was at the low end of the brightness exponent range). There are three possible reasons why the relationship between reported loudness and reported brightness appeared different for Subject 5 than for the other subjects: (1) an auditory problem, (2) a visual problem, and (3) a failure to extend the learned scale to the visual stimuli. We will deal with the potential for constrained scaling to detect sensory differences between subjects in future research. At this point, we would simply like to note that a low level of intersubject variability, such as provided by constrained scaling, is critical for this enterprise.

GENERAL DISCUSSION

The primary goal of this paper was to demonstrate that constrained scaling can be used to calibrate subjects' response scales and thus reduce intersubject variability. Experiments 1A and 1B demonstrated that constrained scaling produces a level of intersubject variability considerably lower than that of other direct scaling techniques. Experiments 2A and 2B further showed that the training and feedback that the subjects received during the constrained scaling process were necessary to achieve this low variability. Experiments 3A and 3B showed that low levels of intersubject variability could be achieved even when the subjects were trained to respond according to a power function with an unnaturally low exponent (0.30 for 1000-Hz tones) or an unnaturally high exponent value (0.90 for 1000-Hz tones), although there was also evidence that training subjects on a more natural exponent value (e.g., 0.60 for 1000-Hz tones) was more efficacious. Finally, Experiment 4 demonstrated that constrained scaling can also be successfully applied cross-modally.

Using constrained scaling, subjects who experience what should be the same level of subjective magnitude (assuming healthy normal individuals and identical conditions) report approximately the same level of subjective magnitude on the learned scale. In other words, when successfully applied, constrained scaling causes subjects to respond using the same unit of psychological magnitude. Therefore, just as many different physical objects can be measured for length using a single unit (e.g., centimeters, feet, cubits), so too, in theory, psychological magnitude can be measured using a common unit (in this case, the sone).

In terms of the extent to which constrained scaling reduces idiosyncratic response bias, it is interesting to compare the results of this study to results from methods that avoid the use of a response continuum. Specifically, it has been argued that avoiding the use of a response continuum avoids response biases altogether and that any remaining individual differences must be due to real individual sensory system differences (Schneider, 1980, 1988). In order to eliminate the response continuum, Schneider (1980) employed the nonmetric approach and, in a separate study (Schneider, 1988), the conjoint measurement approach. In both cases, subjects were required to make binary judgments of "greater than" or "less than" for paired stimuli. For example, in the Schneider (1980) study, subjects judged which of two pairs of tones displayed the greatest loudness difference. Similarly, Schneider (1988) presented subjects with pairs of two-tone complexes (with each complex in the pair made up of two simultaneous tones of different frequencies and intensities) and asked them to judge which was louder. The resulting scales were consistent with the power law, and exponent values were calculated. In the data of Schneider (1980), for 1200-Hz tones, the *SD/M* was 0.308, and the ratio of highest to lowest exponents was 2.55:1. In the data of Schneider (1988): for 2-kHz tones, the *SD/M* was 0.133, and the ratio of highest to lowest exponents was 1.36:1; for 5-kHz tones, the *SD/M* was 0.178, and the ratio of highest to lowest exponents was 1.57:1.

The use of the nonmetric approach (Schneider, 1980) produced a level of intersubject variability no better than standard ME and CMM (see Table 1). We suggest that this was because of the sequential nature of the concatenating procedure, in which at least one tone (the first) had to be held in memory for a later comparison. In contrast, the conjoint measurement approach (Schneider, 1988), in which the tones were presented simultaneously for concatenation, reduced intersubject variability considerably over standard ME and CMM (see Table 1), to a level comparable with constrained scaling. As in the present study, Schneider (1988) presented two different tone frequencies within the same experiment. Examining the ratios of the exponents of the 2-kHz tones to those of the 5-kHz tones for individual subjects, we found that the mean ratio was 0.91, the standard deviation was 0.19, and the *SD/M* was 0.203. Compare this with the results of Experiments 1A and 1B of the present study (using 65- and 1000-Hz tones), which produced exponent ratio *SD/M* levels of 0.086 and 0.082, respectively.

Constrained scaling seems to produce considerably less intersubject variability than other scaling techniques, which suggests that the technique effectively ameliorates the problem of response bias. But why should this be so? As S. S. Stevens (1975) noted, reporting the magnitude of a sensation involves abstracting that magnitude from a more complex multidimensional sensation experience. One possibility is that abstracting magnitudes for the purpose of conscious manipulation involves a higher level cognitive component that is variable across individuals, but also cognitively penetrable. Any matching or concatenating process occurring after this process would consequently reflect this variability (note, as discussed in the introduction, this would not necessarily apply to naturally occurring unconscious matching and concatenating processes). In this case, the training involved in constrained scaling may act to calibrate this component across subjects. A second possibility is that holding two representations of sensory magnitude in memory and then processing them through a matching or concatenating function is not an easy task. In this case, the training involved in constrained scaling may augment subjects' cognitive resources for performing this task by providing them with a well-learned scale. Furthermore, the feedback provided to subjects on the learned scale during the test phase may help to keep the scale fresh in memory. Finally, the process of trying to "guess" the correct answer for the learned scale is engaging (according to subjects' self-reports), which may help focus subjects on the task.

However, although it is clear that constrained scaling reduces variability and thereby increases the reliability of the scaling process, there remains the issue of the validity of the scaling results obtained from the technique. Experiments 3A and 3B demonstrated that, by changing the exponent of the power function on which subjects are trained, the exponent that subjects produce for the novel stimuli in the test phase can be systematically altered. Thus, the exponent values produced through constrained scaling reflect the choice of the training scale more than anything else. However, this is not true of the ratio between the training continuum exponent and the test continuum exponent, which reflects the relative rates of growth in perceived magnitude. Therefore, in theory, constrained scaling should be able to produce valid measures of the relative psychophysical relations between stimulus continua (although this may not be the case for the Equation 5 analysis method used in this study; see Experiment 3B, Results and Discussion). Furthermore, our results, as well as those of Marks et al. (1995) and West and Ward (1994), indicate that the use of a learned scale does not distort basic psychophysical phenomena.

Finally, in addition to addressing the problem of excessive intersubject variability, constrained scaling may also be able to help resolve the problem of interlaboratory variability. In a review of loudness scaling results, Marks (1974a) found that averaged group exponent values from different laboratories also show considerable variation. For example, for 1000-Hz tones, magnitude estimation produced averaged group exponents ranging from 0.43 to 0.70, and ratio production produced averaged group exponents ranging from 0.44 to 0.85 (Marks, 1974a). S. S. Stevens (1955) also surveyed the literature and found that the change in loudness level needed to produce a 2:1 ratio of apparent loudness for 1000-Hz tones varied over a range of 20 dB across experiments. Although this cannot be demonstrated in a single study, we

expect that the significant advantages that constrained scaling offers in terms of its ability to reduce intersubject variability will extend to interlaboratory results, particularly for results reflecting the relationship between stimulus continua.

REFERENCES

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

ALGOM, D., & MARKS, L. E. (1990). Range and regression, loudness scales, and loudness processing: Toward a context-bound psychophysics. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 706-727.

BAIRD, J. C., KREINDLER, M., & JONES, K. (1971). Generation of multiple ratio scales with a fixed stimulus attribute. *Perception & Psychophysics*, **9**, 399-403.

BERGLUND, M. B. (1991). Quality assurance in environmental psychophysics. In S. J. Bolanowski, Jr., & G. A. Gescheider (Eds.), *Ratio scaling of psychological magnitude: In honor of the memory of S. S. Stevens*. Hillsdale, NJ: Erlbaum.

BOLANOWSKI, S. J., JR., & GESCHEIDER, G. A. (EDS.) (1991). *Ratio scaling of psychological magnitude: In honor of the memory of S. S. Stevens* (pp. 140-162). Hillsdale, NJ: Erlbaum.

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

HELLMAN, R. P., & MEISELMAN, C. H. (1988). Prediction of individual loudness exponents from cross modality matching. *Journal of Speech & Hearing Research*, **31**, 605-615.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (1959). *Expression of physical and subjective magnitudes of sound* [ISO/R-131-1959(E)]. Geneva: International Organization for Standardization.

KING, M. C., & LOCKHEAD, G. R. (1981). Response scales and sequential effects in judgment. *Perception & Psychophysics*, **30**, 599-603.

KOH, K. (1993). Induction of combination rules in two-dimensional function learning. *Memory & Cognition*, **21(5)**, 573-590.

KOH, K., & MEYER, D. E. (1991). Function learning: Induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 811-836.

LILIENTHAL, M. G., & DAWSON, W. E. (1976). Inverse cross-modality matching: A test of ratio judgment consistency for group and individual data. *Perception & Psychophysics*, **19**, 252-260.

LOGUE, A. W. (1976). Individual differences in magnitude estimation of loudness. *Perception & Psychophysics*, **19**, 279-280.

LUCE, D. R., & MO, S. S. (1965). Magnitude estimation of heaviness and loudness by individual observers: A test of a probabilistic response theory. *British Journal of Mathematical & Statistical Psychology*, **18**, 159-174.

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

MARKS L. E. (1991). Reliability of magnitude matching. *Perception & Psychophysics*, **49**, 31-37.

MARKS, L. E., GALANTER, E., & BAIRD, J. C. (1995). Binaural summation after learning psychophysical functions for loudness. *Perception & Psychophysics*, **57**, 1209-1216.

MARKS, L. E., & STEVENS, J. C. (1965). Individual brightness functions. *Perception & Psychophysics*, **1**, 17-24.

MARKS, L. E., STEVENS, J. C., BARTOSHUK, L. M., GENT, J. G., RIFKIN, B., & STONE, V. K. (1988). Magnitude matching: The measurement of taste and smell. *Chemical Senses*, **13**, 63-87.

POULTON, E. C. (1989). *Bias in quantifying judgements*. London: Erlbaum.

RULE, S. J., & MARKLEY, R. P. (1971). Subject differences in cross-modality matching. *Perception & Psychophysics*, **9**, 115-117.

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

SCHNEIDER, B. (1988). The additivity of loudness across critical bands: A conjoint measurement approach. *Perception & Psychophysics*, **43**, 211-222.

STEVENS, J. C., & MARKS, L. E. (1980). Cross-modality matching functions generated by magnitude estimation. *Perception & Psychophysics*, **27**, 379-389.

STEVENS, S. S. (1955). The measurement of loudness. *Journal of the Acoustical Society of America*, **27**, 815-829.

STEVENS, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New York: Wiley.

STEVENS, S. S., & GUIRAO, M. (1964). Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology*, **66**, 177-186.

TEGHTSOONIAN, M., & TEGHTSOONIAN, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception & Psychophysics*, **33**, 203-214.

TEGHTSOONIAN, R. (1971). On the exponents in Stevens' law and the constants in Ekman's law. *Psychological Review*, **78**, 71-80.

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

WARD, L. M. (1975). Sequential dependencies and response range in cross-modality matches of duration to loudness. *Perception & Psychophysics*, **18**, 217-223.

WARD, L. M. (1982). Mixed-modality psychophysical scaling: Sequential dependencies and other properties. *Perception & Psychophysics*, **31**, 53-62.

WARD, L. M. (1990). Critical bands and mixed-frequency scaling: Sequential dependencies, equal-loudness contours, and power function exponents. *Perception & Psychophysics*, **47**, 551-562.

WARD, L. M. (1992). Who knows? In G. Borg & N. Neely (Eds.) *Fechner Day 92* (pp. 79-100). Stockholm: International Society for Psychophysics.

WEST, R. L., & WARD, L. M. (1994). *Constrained scaling*. In L. M. Ward (Ed.), Fechner Day 94 (pp. 225-230). Vancouver: International Society for Psychophysics.

ZWISLOCKI, J. J. (1983). Group and individual relations between sensation magnitudes and their numerical estimates. *Perception & Psychophysics*, **33**, 460-468.

ZWISLOCKI, J. J. (1991). Natural measurement. In S. J. Bolanowski & G. A. Gescheider (Eds.), *Ratio scaling of psychological magnitude: In honor of the memory of S. S. Stevens* (pp. 18-26). Hillsdale, NJ: Erlbaum.

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

(Manuscript received September 26, 1996;
revision accepted for publication September 17, 1998.)