Summer 6-1-2003

Recalibrating the Auditory System: A Speed–Accuracy Analysis of Intensity Perception

Yoav Arieh
Montclair State University, ariehy@montclair.edu

Lawrence E. Marks
Yale University, lawrence.marks@yale.edu

Recalibrating the Auditory System: A Speed-accuracy Analysis of Intensity Perception

Article *in* Journal of Experimental Psychology Human Perception & Performance · July 2003

DOI: 10.1037/0096-1523.29.3.523 · Source: PubMed

CITATIONS
40

READS
84

2 authors:

Yoav Arieh
Montclair State University
13 PUBLICATIONS **321** CITATIONS
SEE PROFILE

Lawrence E Marks
The John B. Pierce Laboratory
241 PUBLICATIONS **7,740** CITATIONS
SEE PROFILE

Journal of Experimental Psychology:
Human Perception and Performance
2003, Vol. 29, No. 3, 523–536

0096-1523/03/$12.00 DOI: 10.1037/0096-1523.29.3.523

Recalibrating the Auditory System: A Speed–Accuracy Analysis of Intensity Perception

Yoav Arieh and Lawrence E. Marks
John B. Pierce Laboratory and Yale University

Recalibration in loudness perception refers to an adaptation-like change in relative responsiveness to auditory signals of different sound frequencies. Listening to relatively weak tones at one frequency and stronger tones at another makes the latter appear softer. The authors showed recalibration not only in magnitude estimates of loudness but also in simple response times (RTs) and choice RTs. RTs depend on sound intensity and may serve as surrogates for loudness. Most important, the speeded classification paradigm also provided measures of errors. RTs and errors can serve jointly to distinguish changes in sensitivity from changes in response criterion. The changes in choice RT under different recalibrating conditions were not accompanied by changes in error rates predicted by the speed–accuracy trade-off. These results lend support to the hypothesis that loudness recalibration does not result from shifting decisional criteria but instead reflects a change in the underlying representation of auditory intensity.

People's momentary perception of loudness is remarkably susceptible to preceding auditory stimulation: In loudness adaptation, the perceived intensity of an ongoing low-level tone diminishes over time, and in auditory fatigue, sensitivity and loudness decline after prolonged exposure to intense tones (Botte, Charron, & Bouayad, 1993; Scharf, 1983; Ward, 1973). In both phenomena, exposure duration is critical; the change in loudness or sensitivity is directly related to the duration of the adapting or fatiguing tone. Marks (1988) reported a new variety of time-dependent loudness modification, in which transient, moderate stimulation can lead to substantial changes in loudness. In the first report of this phenomenon, listeners were asked to judge the loudness of brief, weak 500-Hz tones alternating with stronger 2500-Hz tones in one condition (A) and brief, weak 2500-Hz tones alternating with stronger 500-Hz tones in another condition (B). When loudness matches were derived from the judgments (magnitude estimates), it turned out that they differed substantially across the two conditions. In Condition A, a 500-Hz tone of 65 dB was judged as loud as a 2500-Hz tone of 70 dB, whereas in Condition B, the same 65-dB 500-Hz tone was judged as loud as a 53-dB 2500-Hz tone. Thus, loudness judgments shifted across the two contextual conditions by the equivalent of 17 dB (Marks, 1988). Subsequently, Marks (1994, 1996) dubbed these loudness changes "recalibration" to designate a temporary adjustment in auditory responsiveness in different frequency bands.

Other terms have been used to describe the same phenomenon. Marks (1992, 1993) initially used the more descriptive expression "differential context effect," referring to the fact that the effects were first observed under conditions in which each of two sound frequencies took on different contextual sets of sound pressure levels (SPLs). More recently, B. Scharf (personal communication, January 8, 2002) offered the term *induced loudness reduction,* emphasizing the fact that the effects appear to reflect reductions in loudness at the sound frequency receiving high SPLs. We continue to use the term *recalibration,* however, for three reasons: to preserve continuity with terminology used elsewhere (Mapes-Riordan & Yost, 1999; Marks, 1994, 1996); to reflect our view, discussed at the end of this article, that analogous phenomena also characterize intensity processing in other modalities, such as vision and taste; and to distinguish recalibration from the many contextual effects that are commonly attributed to decisional biases.

From the outset, considerable effort has been directed to answering this question: What exactly is being recalibrated? Is it the perceptual representation of loudness that is altered by presenting the specific stimulus ensemble of loud and weak tones at different frequencies? Or are the loudness judgments altered through a response bias, or a criterion shift, operating differentially at the two frequencies? The current study may be the most direct and explicit effort yet to answer this question. We reshaped the recalibration paradigm to fit a speeded classification task in order to measure response times (RTs) and errors rather than loudness per se. As discussed later, in a variety of circumstances, RT is closely correlated with loudness, with RT decreasing monotonically as loudness increases. So RT may in a sense serve as a surrogate measure of loudness—although, for present purposes, it is sufficient to assume that the processes underlying changes in loudness evident in paradigms of recalibration are largely the same as the processes underlying changes in RT.

Yoav Arieh, John B. Pierce Laboratory, New Haven, Connecticut, and Department of Epidemiology and Public Health, Yale University; Lawrence E. Marks, John B. Pierce Laboratory and Department of Epidemiology and Public Health and Department of Psychology, Yale University.

This research was supported by National Institutes of Health Grant DC 03842-04 from the National Institute on Deafness and Other Communication Disorders to Lawrence E. Marks. We thank Nora Williams for her help in running the experiments and the three reviewers, Scott Parker, Robert Teghtsoonian, and an anonymous reviewer, for cogent and thoughtful comments.

Correspondence concerning this article should be addressed to Yoav Arieh or Lawrence E. Marks, John B. Pierce Laboratory, 290 Congress Avenue, New Haven, Connecticut 06519. E-mail: yarieh@jbpierce.org or lmarks@jbpierce.org

In the present study, we analyzed the results obtained in a speeded classification procedure within a conceptual framework in which RTs and errors were used jointly to distinguish changes in "sensitivity" from changes in response criterion, analogous to the way that the conceptual framework of signal-detection theory uses correct detections and false-positive responses to distinguish sensitivity from criterion in sensory discrimination. Specifically, we asked whether choice RT in different stimulus conditions would show recalibration (RT increases when loudness decreases) and, if so, whether the changes in RT would be accompanied by corresponding changes in error rate, as predicted by a speed–accuracy tradeoff. Measuring the speed–accuracy relation for the stimuli undergoing recalibration makes it possible to ask whether changes in RT and error rate can be predicted by the speed–accuracy relation, implying a decisional basis for recalibration, or, alternatively, whether the changes in RT reflect modifications in the underlying perceptual representations of the auditory signals.

Recalibration of Loudness: Sensory or Decisional Basis?

The earliest interpretation of loudness recalibration (Marks, 1988) considered response bias as an explanation, suggesting that recalibration might originate in the ways that people use numbers in magnitude estimation. When people assign numbers to perceptual intensities they tend to use a more or less constant numerical range, independent of the physical range of the stimuli (Teghtsoonian, 1973). Thus, it is possible that, in Marks's (1988) experiment, listeners applied the same response range to the 500-Hz tones and the 2500-Hz tones, regardless of the actual range of SPLs presented at each frequency. Such a strategy would entail wholly relativistic judgment and would result in different implicit loudness matches in the two experimental conditions. Results of subsequent studies, however, cast doubt on this interpretation. Thus, recalibration has also been shown in paradigms that did not require listeners to make numerical judgments. Schneider and Parker (1990) had listeners compare differences in the loudness of various pairs of tones while the average SPL at both frequencies shifted across conditions. Mapes-Riordan and Yost (1999) had listeners compare loudnesses at two frequencies immediately after presenting a brief recalibrating tone at one of the frequencies. Both the comparisons of loudness difference reported by Schneider and Parker and the comparisons of loudness reported by Mapes-Riordan and Yost revealed recalibration, much like that observed with magnitude estimation. Thus, recalibration is revealed in several psychophysical paradigms.

Though it has successfully eliminated numerical response bias as a plausible explanation, research thus far has not demonstrated unequivocally that recalibration indeed involves changes in the underlying perceptual representations. It remains possible to account for the changes in intensity levels that match, indirectly or directly, for loudness by assuming that listeners adjust their response criteria differently at each frequency—that is, by assuming that listeners change the amount of information required at one signal frequency relative to another in order to judge signals at the two frequencies to be equally loud.

Consider the model sketched in Figure 1. Assume for convenience that there is no difference in sensitivity, at threshold or above, to tones at two frequencies, f1 and f2; that is, at equal SPL, the underlying representations of loudness at f1 and f2 are the same. In Figure 1A, the SPLs presented at f1 and f2 are the same, so there is no shift in criterion for loudness-based responses. When asked to judge or compare the loudness of f1 and f2, listeners set their criteria at equal values of SPL. The result is a "correct" loudness match, so that, for example, 60 dB at f1 is judged equal in loudness to 60 dB at f2. Now consider what happens when the SPLs at f1 are lower than those at f2. Biasing tendencies—for instance, a tendency to categorize, at least implicitly, the stimuli at each frequency as "soft" or "loud," and to use the categories equally often (cf. Parducci, 1965)—could induce the listeners to lower the criterion for loudness-based responses at f1 and to raise it at f2, as shown in Figure 1B. The result is a shift in loudness matches. A 60-dB tone at f1 now matches a 70-dB tone at f2. Finally, Figure 1C shows the complementary case in which SPLs

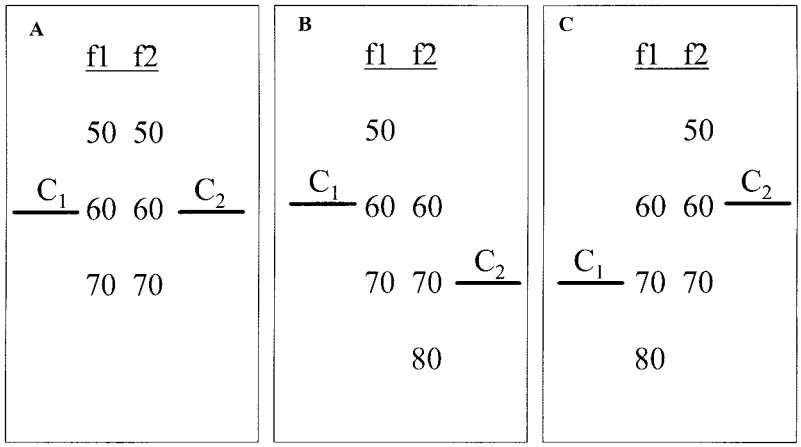

Figure 1. Theoretical model depicting the way in which criteria (C_1 and C_2) for loudness comparisons might be placed when different intensity distributions are allocated to two frequencies (f1 and f2). In Panel A, the intensity distributions at f1 and at f2 are equal; in Panel B, the sound pressure levels (SPLs) at f2 are greater than those at f1; and in Panel C, the SPLs at f1 are greater than those at f2.

at f1 are greater, on average, than SPLs at f2, and the listeners raise and lower, respectively, the criteria at f1 and f2, thereby shifting the matching SPLs in the opposite direction.

The shifting-criterion model, although plausible, has not received much empirical support. In fact, several lines of evidence suggest that the recalibration process is better explained in terms of a change in the representation of auditory intensity (loudness) than as a product of differential response bias. If loudness recalibration reflects nothing more than shifts in decision criteria, independent of the sensory representations of the stimuli, then one would expect the decisional process to be a general one, characterizing judgments at most (if not all) intensity levels, frequencies, and stimulus dimensions. In a series of studies, Marks and colleagues uncovered three important characteristics of loudness recalibration that are hard to reconcile with a general model of shifting criteria (Marks, 1992, 1993; Marks & Warner, 1991).

First, loudness recalibration depends on the intensity levels of the stimuli. One experiment (Marks, 1993, Experiment 15) used a "selective adaptation" method, in which listeners were first presented with a series of exposure tones and then asked to compare test tones at 500 and 2500 Hz. In one condition, the exposure tones had a single frequency and SPL (500 Hz at 53 dB, 500 Hz at 73 dB, 2500 Hz at 48 dB, or 2500 Hz at 68 dB). Only the louder tone at each frequency influenced the subsequent judgments. Exposure to the 500-Hz tone at 73 dB substantially decreased the probability of judging a subsequent 500-Hz tone as louder than a 2500-Hz tone previously equated to it, and exposure to the 2500-Hz tone at 68 dB substantially increased that probability. Exposure to the softer tones had essentially no effect. Mapes-Riordan and Yost (1999) also reported evidence that recalibration arises from stronger, but not from weaker, exposure levels. Thus, when participants are presented with low SPLs at one frequency and high SPLs at another, the changes in relative loudness, such as those reported by Marks (1988), could simply reflect a reduction in loudness at the frequency at which SPLs are high. It follows that loudness recalibration does not require the presentation of stimuli at two sound frequencies. Presenting high SPLs at one frequency suffices (see also Parker & Schneider, 1994), though the introduction of a second frequency provides a convenient yardstick by which to measure recalibration.

Second, the magnitude of the recalibration effect depends on the difference between the two frequencies presented: the smaller the frequency difference, the smaller the shift in matching SPLs (Marks, 1994; Marks & Warner, 1991). Roughly speaking, when the difference between the two frequencies is smaller than a critical band, the amount of recalibration is negligible. In other words, when the two frequencies fall within a critical band, the signals are presumably processed through a common channel and thus show no relative difference in responsiveness.

Third, reversing the roles of the intensity and frequency dimensions—that is, having listeners judge the pitch of soft low-frequency tones and loud high-frequency tones in one condition but loud low-frequency tones and soft high-frequency tones in another—yields no recalibration at all (Marks, 1992). Thus, loudness recalibrates but pitch does not.

Compelling as these findings may be in favoring a sensory over a criterial explanation, they are indirect at best. One can always envision a case in which listeners shift their loudness criteria under a particular set of conditions but not under another. For example, Marks and Warner (1991) noted that the dependence of recalibration on the difference between sound frequencies could be explained in terms of perceived similarity. When the two frequencies fall inside a critical band, they are more similar than when they lie several critical bands apart. It might be easier for listeners to use a different response criterion at each frequency when the frequencies are perceived as dissimilar.

A more direct way to ask whether recalibration reflects changes in sensory representations or decisional criteria is to manipulate and measure shifts in criteria concurrently with measures of loudness—or with surrogates for loudness that, like loudness itself, show recalibration. In essence, the present study aimed to determine whether conditions inducing recalibration do in fact produce shifts in criteria and, if so, whether the shifts in criteria are sufficient to account for the recalibration effect. We extended the recalibration paradigm to a task of speeded classification, in which listeners respond as quickly as possible to each sound by identifying its frequency as low or high, while the intensity levels (irrelevant to the task) vary from condition to condition. Simultaneous measurement of RT and errors affords the opportunity to quantify decisional criteria—in essence, to interpret recalibration effects within the framework of the speed–accuracy trade-off. This approach relies on two main assumptions: first, that RT can serve as a surrogate measure for loudness, depending on comparable mechanisms of auditory intensity processing, and, second, that measures of RT, like measures of loudness per se, show recalibration. Most of the evidence bearing on the first of these assumptions comes from studies of simple response time (SRT).

Loudness and SRT

How closely does SRT reflect loudness perception? In an auditory SRT task, listeners are asked, while timed, to press a key as quickly as possible when they detect a tone. Ample evidence exists in support of an inverse relation between tone intensity (SPL) and SRT. Classic experiments by Cattell (1886), Chocholle (1940), and, more recently, Kohfeld (1971) have convincingly shown that RT decreases monotonically with corresponding increases in SPL (see also Luce & Green, 1972). Further, Chocholle and Greenbaum (1966) showed that when loudness decreases by masking a target tone with a noise, SRT to that tone increases. The inverse relation between SRT and loudness is not perfect, however; as Kohfeld, Santee, and Wallace (1981) reported, at low intensity levels (20 and 40 phons—i.e., levels matching the loudness of a 1000-Hz tone at 20- and 40-dB SPL), equally loud tones at different frequencies yield similar but not identical SRTs. Consequently, it is pertinent to ask whether loudness recalibration, so clearly evident in magnitude estimation and paired comparison tasks, would reveal itself in SRT. We did so by measuring recalibration in loudness (Experiment 1) and SRT (Experiment 2).

If presenting greater SPLs at one frequency, f1, relative to another, f2, acts to depress responsiveness to signals at f1, then SRT should be greater at f1 than at f2. Thus, when the average SPL at 500 Hz is smaller than the average SPL at 2500 Hz, we would expect that responses to 500-Hz tones would be faster than responses to 2500-Hz tones. But when the intensity relations are reversed, we would expect responses to 2500-Hz tones to be faster than responses to 500-Hz tones.

Grice (1968) proposed an evidence-accumulation model to account for the effect of intensity on SRT. The model suggests that in order to initiate a response, information about neural events accumulates over time in the sensory pathways until a certain criterion, presumably one enabling the system to distinguish external stimulation from internal noise, has been satisfied. Because information accumulates more rapidly when intensity is higher, stronger signals reach the response criterion sooner than weaker signals, leading to quicker response.

One way to track changes in criterion is to measure false-alarm responses in SRT tasks by recording anticipatory responses. If a person lowers the criterion for responding—that is, reduces the amount of information required to initiate a response—then noise events alone will more often surpass the criterion. Thus, given constant stimulus intensity, faster response is associated with greater proportions of false-alarm responses (Green & Luce, 1971). The presence of a trade-off between speed and accuracy implies that criterion is shifting. Unfortunately, for present purposes, within a block of trials containing different signal frequencies presented at different SPLs (the recalibration paradigm), it is not possible to attribute the anticipatory responses (false alarms) to individual stimuli within the block. Thus, it is not possible to determine, in an SRT task, whether changes in SRT at one frequency relative to another are accompanied by corresponding changes in anticipatory responses. To use speed–accuracy relations to track recalibration, it is necessary to assign errors to each stimulus in the ensemble. Fortunately, this is possible within a choice response time (CRT) task.

Loudness and CRT

In adapting the recalibration paradigm to a choice procedure, we asked participants to classify tones on the basis of their frequency (one response to high tones, another to low tones) while SPL varied irrelevantly. If CRT follows loudness, then at any fixed SPL, CRT should be greater at whichever frequency the SPLs are greater. This adaptation of the recalibration paradigm assumes, of course, that CRTs in some fashion follow or mirror loudness—for example, CRT (in a task of frequency classification) should decline as SPL increases, much as SRT (in a detection task) declines with increasing SPL. Working in the visual modality, Pins and Bonnet (1996) reported that both SRT and CRT declined as light intensity increased. But not all studies have been so encouraging. Some earlier studies suggested that CRTs sometimes follow signal intensity as a U-shaped function. This seems particularly to be the case when, as is typical in choice tasks, the foreperiod (time between the warning signal and the onset of the test stimulus) is fixed in duration—as reported by van der Molen and Keuss (1979) in hearing. Subsequently, however, these investigators reported a monotonic relation between CRT (in a task requiring responses to signal frequency) and SPL when the foreperiod was both variable and relatively long (Keuss & van der Molen, 1982).

The findings of Keuss and van der Molen (1982) suggest that the recalibration paradigm can be adapted to a choice task, in which CRT may serve as a correlate to or a surrogate for loudness. Further, the choice task makes it possible to measure errors as well as RTs and, thus, to compute the speed–accuracy trade-off function. When RT in a choice task is manipulated between experimental blocks (e.g., by enforcing a progressively stricter response deadline) and then plotted against accuracy (i.e., percent correct), the result is a monotonically increasing, negatively accelerated function (Pachella, 1974; Pew, 1969). In other words, up to some asymptotic level, as RT increases so does the percentage of correct responses.

The interpretation of the speed–accuracy trade-off depends on the specific model of CRT one chooses to endorse. With the exception of the fast-guess model (Yellott, 1971), the two major classes of models vying to account for CRTs—evidence-accumulation models and random-walk models (see Luce, 1986, for an extensive review)—interpret errors as a result of decisions based on insufficient information. Both classes of models assume that responses depend on central decision processes that operate on evidence about the stimulus that is acquired over time. Both assume that the accumulated evidence is inherently probabilistic, but they differ in the way that they assume the evidence is used in making decisions. The evidence-accumulation model (which is an extension of the SRT model presented earlier) assumes that information regarding each stimulus alternative is simply aggregated and that a decision is made as soon as information regarding one alternative reaches its predetermined criterion. Because it takes time to accumulate evidence, higher criteria will result in longer RTs but fewer errors. According to this model, the speed–accuracy trade-off simply reflects a shift in decision criteria.

In the random-walk model, information about stimulus alternatives is haphazardly acquired over time until, eventually, evidence favors one of the alternatives. The decision to respond is based on a relative criterion rather than an absolute one. When the evidence favoring one alternative exceeds the evidence in favor of the other by a critical amount, a response is initiated. RT depends on this critical amount: If its value is high, RT will be great, but few errors will be made because the decision will be based on a relatively large amount of evidence. Decreasing the critical value promotes faster responses but increases the likelihood of errors because the response decision is based on less information. According to the random-walk model, the speed–accuracy trade-off reflects a shift in the critical value by which one alternative must exceed another for a response to be initiated.

Critically for our goals, in both evidence-accumulation and random-walk models, the speed–accuracy trade-off is a hallmark of shifting response criteria at the decision stage of processing and not of changes in the sensory representations of the stimuli. Using this framework, we examined whether changes in CRT within a loudness-recalibration paradigm can be characterized as changes in response criteria. In Experiment 3, listeners classified tones as low (500 Hz) or high (2500 Hz) in frequency. The two experimental conditions were identical to those of the SRT task (Experiment 2), with the 2500-Hz tones having relatively high SPLs and the 500-Hz tones having relatively low SPLs in Condition A and the assignment of SPLs to frequencies being reversed in Condition B.

We assessed the relation between speed and accuracy at each frequency and SPL by having listeners participate in two sessions. In one session, the instructions emphasized accuracy (unspeeded session), and in the other session, the instructions emphasized speed (speeded session). In both sessions, we expected CRT to reveal loudness recalibration as follows: Listeners should classify the 500-Hz tones more quickly than the 2500-Hz tones in Condition A but more slowly in Condition B. Further, for a given

stimulus (combination of frequency and SPL), we could derive four trade-off functions: two functions based on measures obtained with speeded and unspeeded instructions—the speed–accuracy trade-off function (SATF)—and another two functions based on measures obtained in Conditions A and B—the recalibration trade-off function (RTF). If, on the one hand, the RTF is identical to the SATF, then loudness recalibration may be attributed to decisional processes. If, on the other hand, the two functions differ, and in particular if changes in RT under recalibration are not accompanied by corresponding changes in errors, then the results imply that the changes in RT cannot be attributed to decisional processes but are more likely the result of changes in responsiveness (representations of intensity) in the auditory system.

In summary, the present study contains three experiments. In the first, we measured recalibration in judgments of loudness. In the second, we adapted the stimulus regimen of recalibration to a simple detection task, in which listeners responded as quickly as possible to the onset of a stimulus. To anticipate, the results (SRTs) showed recalibration, much like that observed in loudness judgments. Then, in the third experiment, we applied a similar stimulus regimen to a choice task, in which listeners classified tones as quickly as possible as low or high in frequency. To anticipate again, these results (CRTs) also showed recalibration, and, more important, they showed that the changes in RT were not accompanied by corresponding changes in accuracy, thereby suggesting that recalibration involves modification of the underlying perceptual representations of auditory intensity.

Experiment 1

Experiment 1 used a magnitude estimation task to replicate and quantify loudness recalibration with the stimuli that would be used in the subsequent RT experiments. The need to obtain many responses per stimulus in the RT experiments, so as to ensure reliable measures of RTs and errors, dictated the use of a small number of SPLs at each signal frequency—in the SRT and CRT experiments, three SPLs at each frequency, at a brief duration of 250 ms. Although these parameters depart from those used by Marks (1988), who presented eight SPLs at each frequency, at a duration of 1 s, we selected stimulus conditions (Marks, 1988, 1992, 1993, 1994) to ensure that loudness recalibration would be reliable and substantial.

Method

Participants. All participants in this and the subsequent experiments were 18–40 years old and were recruited from the Yale University community. Three men and five women, who self-reported to have normal hearing, were paid to participate.

Apparatus and stimuli. The stimuli in this and all subsequent experiments were produced by a Tucker-Davis System 3 Real Time processor (Alachua, FL) driven by a Matlab program (Natick, MA) running on a Pentium III PC. The 500-Hz and 2500-Hz signals, appropriately attenuated (Tucker-Davis PA5 module) and gated (10-ms rise and decay), were delivered binaurally for 250 ms through calibrated TDH-49 headphones mounted in MX41/AR cushions (Farmingdale, NY). Two experimental conditions were created. In Condition A, the SPLs of the 500-Hz signal were relatively low (35, 50, and 65 dB) and those of the 2500-Hz signal were relatively high (45, 60, and 75 dB); in Condition B, these relations were reversed, with the SPLs at 500 Hz being relatively high (50, 65, and 80 dB) and those at 2500 Hz being relatively low (30, 45, and 60 dB). Thus, at each frequency, the SPLs differed across the two conditions by 15 dB, and because the sets shifted in opposite directions, the overall difference between Conditions A and B was 30 dB. The SPLs at 500 Hz were set 5 dB above those at 2500 Hz overall to compensate for the auditory system's greater sensitivity at 2500 Hz.

Procedure. Participants were tested individually in a sound-treated booth. The method was magnitude estimation with no specified modulus. The participants were instructed to assign to the first tone whatever number they deemed appropriate to stand for its loudness and then to assign to succeeding stimuli other numbers in proportion. Whole numbers, decimals, and fractions were permitted. A response of "zero" would reflect the absence of sound, but no such responses were reported. Each participant received four blocks of randomly presented tones, with experimental conditions alternating between blocks. Half of the participants received the order A, B, A, B, whereas the other half received the order B, A, B, A. All six tones (three SPLs at each frequency) in a given experimental condition were presented a total of 16 times to generate the 96 trials in each block (and 384 trials in the session). After each tone, the participant was prompted to type in the response on a computer keyboard.

Results and Discussion

The magnitude estimates given by each participant to each stimulus were averaged arithmetically within sessions, and these means were then averaged geometrically across participants in each condition. Geometric averaging across participants helps to correct for different listeners using different scales; arithmetic averaging within participants averts problems with geometric averaging when there are occasional responses of "zero"—which, fortunately, did not occur in the present study. Figure 2 shows clearly that the set of SPLs exerted a substantial effect on the judgments of loudness. In Condition A, in which the SPLs at 500 Hz were low and those at 2500 Hz were high, the 45-dB and 60-dB signals at 2500 Hz were judged softer than the corresponding 50-dB and 65-dB signals at 500 Hz. But in Condition B, in which the SPLs at 500 Hz were high and those at 2500 Hz were low, just the reverse occurred, with the 45-dB and 60-dB signals at 2500 Hz now being judged louder than the corresponding 50-dB and 65-dB signals at 500 Hz.

To assess these results statistically, we entered the log-transformed loudness judgments into a repeated measures analysis of variance (ANOVA) with condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (45 and 60 dB at 2500 Hz, the nominally equivalent 50 and 65 dB at 500 Hz) as within-subject variables. As we expected, the Condition $\times$ Frequency interaction was statistically significant, $F(1, 7) = 25.45$, $p < .01$, verifying the change in relative loudness evident in Figure 2. Subsequent analysis showed that the 500-Hz tones were judged significantly louder than the 2500-Hz tones in Condition A, $t(7) = 3.55$, $p < .05$, but significantly softer than the 2500-Hz tones in Condition B, $t(7) = -2.60$, $p < .05$.

Apart from the ubiquitous main effect of intensity, $F(1, 7) = 53.60$, $p < .01$, only two other terms reached statistical significance: the interactions of Condition $\times$ Intensity, $F(1, 7) = 6.20$, $p = .04$, and of Frequency $\times$ Intensity, $F(1, 7) = 7.30$, $p = .03$. The former reflects the finding that weak signals in Condition A received slightly lower ratings on average than weak signals in Condition B but strong signals did not, and the latter reflects the finding that the 45-dB signal at 2500 Hz was judged louder than the (corresponding) 50-dB signal at 500 Hz, whereas the 60-dB signal at 2500 Hz was judged softer than the corresponding 65-dB

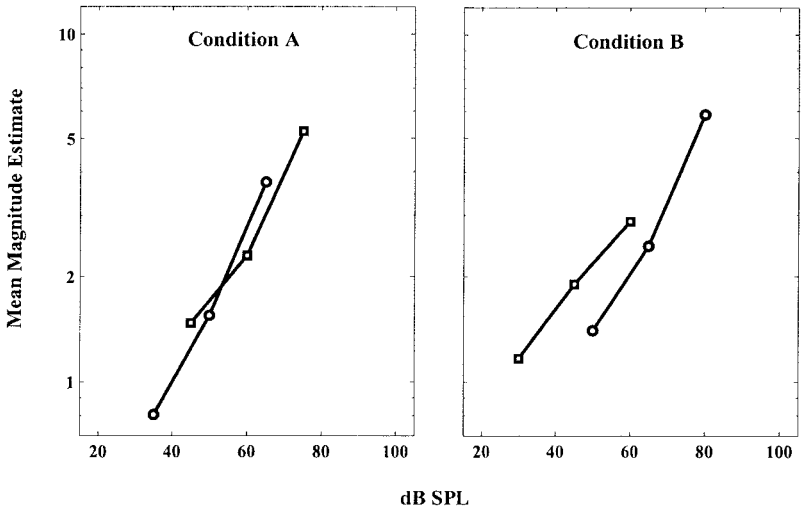

Figure 2. Mean magnitude estimates of the loudness of 500-Hz tones (circles) and 2500-Hz tones (squares), plotted against sound pressure level (SPL) for Experiment 1. Condition A comprised low SPLs at 500 Hz and high SPLs at 2500 Hz, and Condition B comprised high SPLs at 500 Hz and low SPLs at 2500 Hz.

signal at 500 Hz. The Frequency × Intensity interaction notwithstanding, the 500-Hz and 2500-Hz tones were judged, on the whole, roughly equal in loudness, as indicated by the lack of a reliable main effect for frequency, $F(1, 7) = 1.65$, $p > .10$.

It is instructive to calculate loudness recalibration in terms of the SPLs needed at 500 Hz and 2500 Hz to match in loudness. To this end, we used the method described by Stevens and Marks (1980) for deriving matching stimulus values from magnitude estimates. In essence, this matching is accomplished by the following procedure. For each SPL at 500 Hz (value on the abscissa) in Figure 2, we project the resulting magnitude estimate (value on the ordinate) horizontally onto the 2500-Hz function (where interpolation is needed, points are connected by a straight line segment) and then read off the corresponding SPL. The pair of SPLs constitutes an implicit loudness match. An additional set of matches is obtained by reversing the procedure, projecting each mean estimate for 2500 Hz onto the function for 500 Hz.

Applying this procedure to the data obtained in Condition A produces almost perfect stimulus matches between signals at 500 and 2500 Hz. For example, in Condition A, a 500-Hz tone at 50 dB was judged as loud as (i.e., was given the same magnitude estimate as) a 2500-Hz tone at 48 dB. But in Condition B, the same 500-Hz tone at 50 dB was judged as loud as a 2500-Hz tone at 40 dB—a shift in matching SPLs of 8 dB. On average, the SPL had to be 0.5 dB greater at 2500 Hz than at 500 Hz for loudness to be equal in Condition A (about 4.5 dB less than the amount predicted by measures of equal loudness in "neutral," nonrecalibration conditions; Fletcher & Munson, 1933) but 8 dB greater in Condition B. Thus, the overall shift in implicit loudness matches was 7.5 dB. Given that the physical shift between Stimulus Sets A and B amounted to 30 dB overall (15-dB shift at each frequency), a change of 7.5 dB in loudness matches constitutes 25% of the physical shift. This value is smaller than the 44% reported by Marks (1988).

Several major differences between the present study and that of Marks (1988) may account for the smaller percentage change in the present loudness matches. First, Marks used more stimulus levels at each frequency (eight rather than three)—although it was not clear a priori that this should have been of consequence. Second, and more important, Marks used SPLs that extended to higher levels (90 dB at 500 Hz and 85 dB at 2500 Hz); Marks (1994) and Mapes-Riordan and Yost (1999) have shown recalibration to depend primarily on more intense stimuli. Third, Marks (1988) used signals that lasted four times longer (1000 vs. 250 ms); longer durations might evoke greater recalibration. Fourth, and perhaps most important, participants in the current study received alternating blocks of trials containing the two experimental conditions in a single session, whereas Marks's (1988) participants received each condition in a different session separated by at least 24 hr. Marks (1994) found that changing conditions within sessions rather than across sessions reduces the magnitude of recalibration, mainly due to carryover effects from one block of trials to another.

To explore this last possibility, we also looked at just the 48 trials constituting the second half of each block and recomputed loudness matches. When this was done, the overall shift in loudness matches between the two conditions amounted to 11.5 dB, or about 38% of the 30-dB difference between the stimuli. Although this percentage is still smaller than the 44% reported by Marks, it is nevertheless a substantial quantity.

In summary, Experiment 1 revealed loudness recalibration with the stimulus parameters to be used in subsequent RT experiments. Although the recalibration was somewhat smaller in magnitude than those previously reported, it is still substantial, and it provides a baseline for the next experiment, which asks whether recalibration can be demonstrated by using SRT as a surrogate for loudness.

Experiment 2

Method

Participants. Four men and four women, who self-reported to have normal hearing, were paid to participate. None had participated in Experiment 1.

Apparatus and stimuli. The apparatus and experimental conditions were the same as those used in Experiment 1, except for the apparatus to collect RTs. For this purpose, we used a Tucker-Davis RB-25 response box for an RP-2 processor, with a sampling frequency of 50 kHz, which supplied better-than-millisecond accuracy.

Procedure. Participants were tested individually in a sound-treated booth. They were asked to press a key on the response box as soon as they detected a sound. Each trial began with a warning signal, a black square at the middle of the screen, followed by a variable foreperiod prior to onset of the test stimulus. To minimize the influence of expectancy and time estimation on participants' performance, the length of the foreperiod was governed by a constant hazard function (Thomas, 1967). With a constant hazard function, the probability of the signal appearing in the next time interval remains constant, regardless of the length of the foreperiod to that point. In the present case, the probability of presenting the signal within each successive 80-ms interval was 10%, which gave the foreperiod an exponential distribution with a theoretical mean of 800 ms.

The experimental session consisted of seven blocks with 96 trials in each. The first block was considered as practice and was discarded. The experimental conditions alternated between blocks, with half of the participants receiving the order A, B, A, B, A, B and the other half receiving the order B, A, B, A, B, A. Matlab software recorded RTs as well as any premature responses given during the foreperiod. These responses were treated as false alarms and analyzed separately. An average session lasted about 30 min.

Results and Discussion

In this and all subsequent analyses of RTs, values more than 2 standard deviations above and below participants' means were considered outliers and were not included in the analyses. Figure 3 presents the average RTs given to sound signals in the two experimental conditions. The similarities to the magnitude estimation data presented in Figure 2 are striking. First, SRT varied monotonically with signal intensity. At each frequency in each condition, as SPL increased RT diminished. Second, SRT changed across conditions at 500 Hz and 2500 Hz in a fashion that mimicked loudness recalibration. In Condition A, in which SPLs at 500 Hz were low and those at 2500 Hz were high, the 45-dB and 60-dB signals at 2500 Hz were detected more slowly than the corresponding 50-dB and 65-dB signals at 500 Hz. In Condition B, however, in which SPLs at 500 Hz were high and those at 2500 Hz were low, 45-dB and 60-dB signals at 2500 Hz were detected more quickly than the corresponding 50-dB and 65-dB signals at 500 Hz. Given that SRT is monotonically and inversely related to loudness, this pattern is the hallmark of recalibration.

An ANOVA performed on the RTs with condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (50 and 65 dB at 500 Hz, 45 and 60 dB at 2500 Hz) as within-subject variables confirmed what is evident from the data in Figure 3. The Condition × Frequency interaction was significant, $F(1, 7) = 18.65$, $p < .01$, again reflecting recalibration. Subsequent analysis showed that responses to the 500-Hz tones were significantly faster than responses to the 2500-Hz tones in Condition A, $t(7) = 2.51$, $p < .05$; in Condition B, the difference just bordered on significance, $t(7) = -2.17$, $p = .06$, with responses to the 500-Hz tones slower than responses to the 2500-Hz tones. Thus, the frequency at which signals were detected relatively quicker (akin to being judged relatively louder in Experiment 1) depended on the experimental condition. No other statistical term reached significance, apart from the main effect of intensity, $F(1, 7) = 21.90$, $p < .01$.

The measures of RT are amenable to the same analysis, in terms of matching SPLs, that we applied to the magnitude estimates of

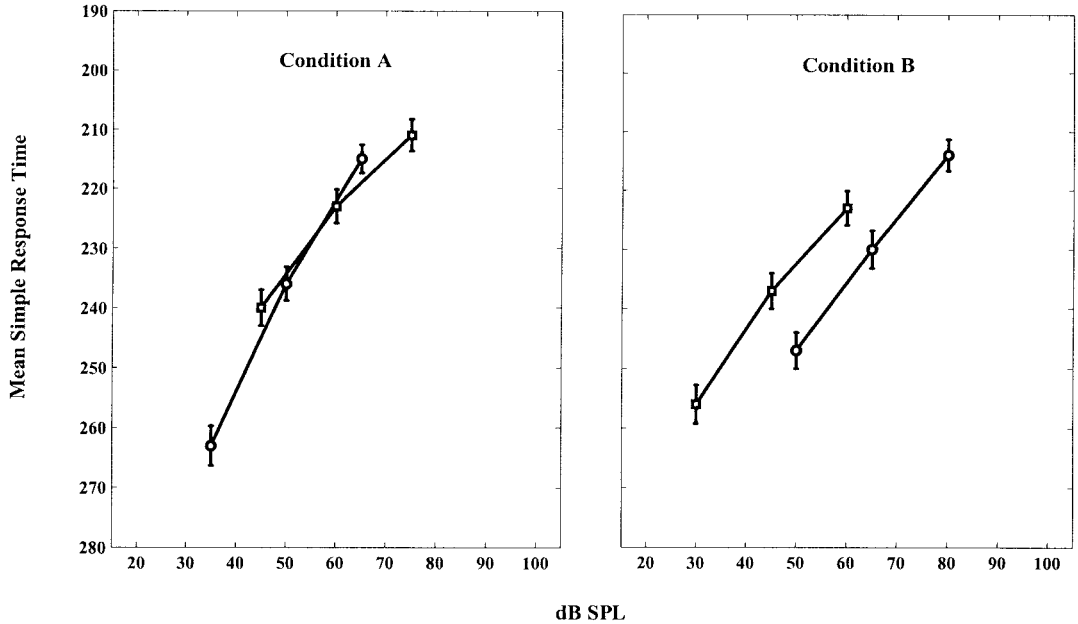

Figure 3. Mean simple response times (in milliseconds; ±1 *SEM*) to detect 500-Hz (circles) and 2500-Hz (squares) tones in Conditions A and B, plotted against sound pressure level (SPL) for Experiment 2. To facilitate comparison with Figure 2, the *y*-axis is reversed.

loudness in Experiment 1. Here, however, the criterion for a match was equal RT rather than equal loudness judgment. In Condition A, participants responded almost identically to 500-Hz and 2500-Hz signals at equal SPLs: On average, a tone at 2500 Hz had to be 0.5 dB less intense than a corresponding tone at 500 Hz to produce the same SRT in Condition A but 12.5 dB more intense in Condition B. Thus, the overall change in matching SPLs between Condition A and Condition B was 13 dB, which amounts to 43% of the overall difference of 30 dB in the stimulus intensities in the two experimental conditions. Although this value is very close to the 44% shift in matching SPLs reported by Marks (1988) using loudness judgments, the similarity may be fortuitous. More important, the present value is somewhat larger than the one reported in Experiment 1.

The magnitude of the recalibration observed here is impressive. Apparently, insofar as SRT relies on intensity-processing mechanisms similar to those that underlie the perception of loudness, SRT, like loudness, can also manifest changes in matching intensities wrought by exposing listeners to different stimulus ensembles. Furthermore, computing matching intensities for only the second half of each experimental block (last 48 of the 96 trials) yielded a contextual shift of 15.8 dB, which is more than 50% of the overall shift between the stimulus ensembles in the two conditions. The larger shift in the second part of the experimental block resembles the pattern obtained in Experiment 1 with magnitude estimation.

As most models of SRT agree, changes in RT may stem from two sources, sensitivity and criterion (Grice, 1968), with the latter usually accompanied by changes in false-alarm rates. Do the shifts in RT from one condition to another reflect changes in sensitivity or in response criterion? Two major difficulties limit a useful analysis of the criterion (false alarms) in Experiment 2. First, the rate of false alarms was small—less than 1%. Second, to explain loudness recalibration in terms of a shifting response criterion, it is necessary to assume that in a given condition the response criterion was placed differently for each frequency (see Figure 1). A false alarm in a detection task is, by definition, a response made before the signal is presented and thus before information about its frequency has registered. Thus, it is unlikely that false-alarm rates could be analyzed meaningfully within the framework of the present loudness-recalibration paradigm.

Experiment 3 was designed to overcome this limitation by determining whether changes in RT across the two experimental conditions indeed reflect shifts in response criteria. Switching to a task requiring the listener on each trial to choose between two possible responses—classifying tones as high or low in pitch—enabled us to measure error rates as well as RTs and, consequently, to compute SATFs. The presence of a speed–accuracy trade-off between experimental conditions could indicate that the observed changes in CRT, and by inference in loudness, are a result of a shifting response criterion, which promotes faster responses although at the cost of greater error rates.

Experiment 3

Method

Participants. Two men and six women, who self-reported to have normal hearing, were paid to participate. None had participated in the previous experiments.

Apparatus and stimuli. The apparatus and experimental conditions were the same as those used in Experiment 2.

Procedure. The procedure and stimulus presentation were the same as those used in Experiment 2, but the listeners' task was different. Listeners were asked to classify each auditory signal as high or low in pitch by pressing the appropriate button on the response box. Listeners served in two sessions that differed in their instructions. In one session, listeners were asked to respond as quickly as possible but also to minimize the error rate. In the other session, listeners were asked to respond as quickly as possible regardless of the resulting error rate. We anticipated that these instructions would induce two different criteria for response, with the criterion being more stringent in the unspeeded session and more lenient in the speeded session. Each session consisted of seven blocks with 96 trials in each. The first block was considered as practice and was discarded. The blocks were presented in two possible orders: A, B, A, B, A, B or B, A, B, A, B, A. Each session lasted about 30 min, and the two sessions were performed in succession with a 10-min break between them. Counterbalanced over participants were the order of sessions, the assignment of keys to high and low pitch, and the order of the blocks of conditions.

Results and Discussion

Figure 4 presents the average RTs for correct responses given to all of the signals in the two experimental conditions. Two features of the data are salient. First, RTs in the speeded session were indeed shorter than RTs in the unspeeded session, indicating that listeners complied with the instructions. Second, loudness recalibration, similar to that observed in Experiments 1 and 2, is clearly evident. Whereas in Condition A RTs to signals at 500 Hz were slightly smaller than the RTs to signals at 2500 Hz, in Condition B they were greater. Thus, for the first time we have demonstrated recalibration using a choice task in which listeners base their response on the frequency rather than the intensity of the tone. But even though listeners responded to the frequency of the auditory signals, the intensity of the signals affected pitch-classification times—and, by implication, loudness perception.

An omnibus ANOVA with session (speeded, unspeeded), condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (50 and 65 dB at 500 Hz, 45 and 60 dB at 2500 Hz) as within-subject variables revealed three significant terms. One was a main effect of session, $F(1, 7) = 14.70$, $p < .01$, reflecting the fact that, on average, listeners were 41 ms faster to classify the tones in the speeded session than they were in the unspeeded session (mean RTs of 278 and 319 ms, respectively). Recalibration was evident in the second significant term, the Condition × Frequency interaction, $F(1, 7) = 22.40$, $p < .01$. Notably, the three-way Session × Condition × Frequency interaction did not reach statistical significance, $F < 1$, indicating that recalibration did not differ between the two sessions. The only other statistical term to reach significance was the omnipresent main effect of intensity, $F(1, 7) = 43.80$, $p < .01$.

To examine further the loudness recalibration, we ran two separate ANOVAs, one on each session, with condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (50 and 65 dB at 500 Hz, 45 and 60 dB at 2500 Hz) as within-subject variables. For data obtained in both sessions, the crucial Condition × Frequency interaction was statistically significant for the speeded session, $F(1, 7) = 9.10$, $p < .05$, and for the unspeeded session, $F(1, 7) = 63.60$, $p < .01$. More specifically, in the speeded session the 500-Hz tones were classified faster than the 2500-Hz tones in Condition A, $t(7) = -4.10$, $p < .01$, but slower than the 2500-Hz

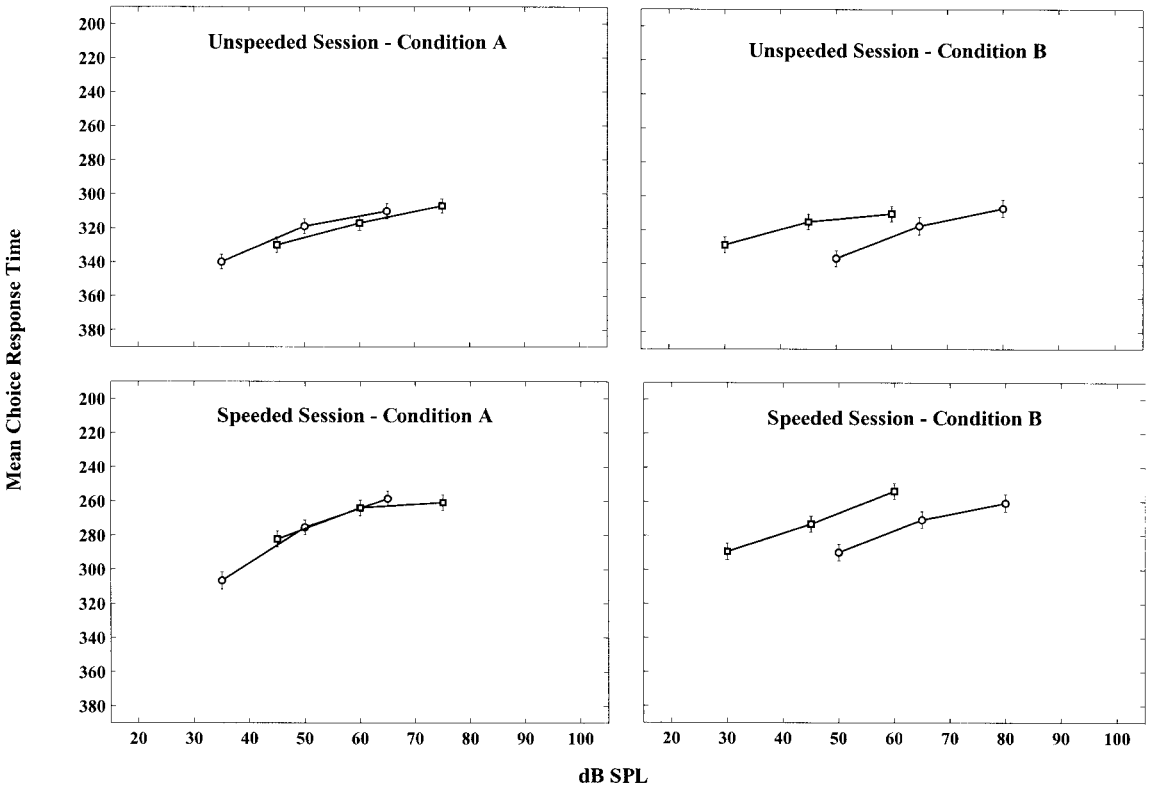

Figure 4. Mean choice response times (in milliseconds; ±1 *SEM*) to classify 500-Hz (circles) and 2500-Hz (squares) tones in Conditions A and B in the speeded and unspeeded sessions, plotted against sound pressure level (SPL) for Experiment 3. To facilitate comparisons with Figures 2 and 3, the *y*-axis is reversed.

tones in Condition B, $t(7) = 3.20$, $p < .05$. A similar pattern was obtained for the unspeeded session: The 500-Hz tones were classified faster than the 2500-Hz tones in Condition A, $t(7) = -4.70$, $p < .01$, but slower in Condition B, $t(7) = 4.10$, $p < .01$.

Curiously, shifts in implicit matches derived from these RT functions were the greatest found in this study. For the unspeeded session, in Condition A the 500-Hz tones had to be 4.5 dB smaller in SPL than the 2500-Hz tones to produce the same RT. In Condition B, however, the 500-Hz tones had to be 21.5 dB greater in SPL than the 500-Hz tones to produce the same RT. These shifts in implicit-matching intensities amount to an overall value of 25 dB, which nearly equals the overall value of 30 dB between stimulus levels in the two conditions. For the speeded session, in Condition A the 500-Hz and 2500-Hz tones matched almost perfectly (equal RT at equal SPL), whereas in Condition B the 2500-Hz tones had to be 18.8 dB greater than the 500-Hz tones to produce the same RT.

Table 1 presents the error rates computed for each stimulus in each session and experimental condition. The most salient feature is the difference in error rates obtained with unspeeded and speeded instructions—5.6% and 18.5%, respectively. This disparity, coupled with the difference in CRTs reported above, indicates that performance under the two instructions underwent a marked speed–accuracy trade-off. With instructions emphasizing speed, listeners paid the cost of more errors for the gain of extra-quick response.

The more important question, and the one that lies at the heart of this study, is whether the changes in RT across the two exper-

Table 1
Error Rates (Percentages) for Classifying 500-Hz and 2500-Hz Tones in Experimental Conditions A and B Under Unspeeded and Speeded Instructions in Experiment 3

	Condition A						Condition B					
	500 Hz			2500 Hz			500 Hz			2500 Hz		
Instructions	35 dB	50 dB	65 dB	45 dB	60 dB	75 dB	50 dB	65 dB	80 dB	30 dB	45 dB	60 dB
Unspeeded	3.4	3.6	5.5	5.5	5.9	5.5	6.8	6.8	7.8	6.3	4.9	6.0
Speeded	21.9	16.9	18.8	15.9	17.6	20.1	20.3	20.8	23.2	16.7	13.8	18.5

imental conditions—that is, the recalibration—were the product of a speed–accuracy trade-off. The data in Table 1 show no evidence that this was so. Consider, for example, the RTs and error rates for the 500-Hz tones at 50 and 65 dB in the unspeeded session. Participants classified the pitch of these tones more quickly in Condition A than Condition B (RTs of 319 and 310 ms in Condition A and 336 and 317 ms in Condition B at 50 and 65 dB, respectively). A criterial (speed–accuracy trade-off) explanation of recalibration predicts that the reduction in RTs in Condition A would be accompanied by an elevation in error rates. But this was not the case. In fact, error rates for pitch classification in Condition A were somewhat lower than error rates in Condition B (3.6% and 5.5% in Condition A and 6.8% and 6.8% in Condition B at 50 and 65 dB, respectively). Overall, the error rates at a given frequency were very similar across the two conditions. In the unspeeded session, the mean error rate at 500 Hz was 4.2% in Condition A and 6.8% in Condition B, and at 2500 Hz it was 5.7% for both conditions. In the speeded session, the mean error rate at 500 Hz was 19.2% in Condition A and 21.4% in Condition B, and at 2500 Hz it was 17.2% and 16.3% for Conditions A and B, respectively.

Subjecting the error rates to an ANOVA with session (speeded, unspeeded), condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (50 and 65 dB at 500 Hz, 45 and 60 dB at 2500 Hz) as within-subject variables revealed only one statistically significant term: a main effect of Session, $F(1, 7) = 43.77$, $p < .01$, reflecting the difference in accuracy between the instructions. We also ran two additional ANOVAs, one on each session, with condition (A, B), frequency (500 Hz, 2500 Hz), and intensity (50 and 65 dB at 500 Hz, 45 and 60 dB at 2500 Hz) as within-subject variables. For both analyses, the Frequency $\times$ Condition interaction did not approach statistical significance, Fs < 1.

To examine further this critical issue, we compared directly the between-sessions SATFs and the between-conditions RTFs. For this analysis, we used the four stimuli common to the two conditions (500 Hz at 50 and 65 dB and 2500 Hz at 45 and 60 dB). For each stimulus, we plotted the average error rate against the average CRT in the speeded and unspeeded sessions and in Conditions A and B, thereby producing four points. Each set of four points could then be connected in two meaningful ways, as shown in an idealized manner in Figure 5. A line connecting two points obtained in a given condition with speeded and unspeeded instructions reveals the classic SATF (dashed line in Figure 5). Alternatively, a line connecting two points obtained in a given session with different stimulus conditions, A and B, reveals the RTF (solid line in Figure 5).

The SATF and the RTF quantify the magnitude of the trade-off between speed and accuracy when, in the one case, listeners shift criterion in response to instructions and, in the other, listeners change responsiveness due to recalibration. Figure 5A shows a theoretical example in which the SATF and the RTF differ substantially. The slopes of the SATF are negative, meaning that shorter RTs are accompanied by higher error rates. This pattern is commonly interpreted as reflecting a criterion shift, whereby lowering the criteria to hasten response results in accumulation of less information and, thus, in more errors. The slopes of the RTF in this example are slightly positive, indicating that shorter RTs are accompanied by slightly lower error rates. This pattern implies that the change in RT is not a result of criterion shift. Figure 5B shows an alternate theoretical example, in which the SATF and the RTF are virtually identical. This pattern implies that loudness recalibration as reflected in CRT can be explained by the SATF and may therefore reflect a shift in response criterion.

Figure 6 shows the SATF and the RTF computed from the data of Experiment 3. The data plotted in the four panels of Figure 6 largely resemble the theoretical alternative shown in Figure 5A. With the 500-Hz signal at 50 dB and 65 dB, and with the 2500-Hz signal at 45 dB, the SATF and the RTF differed markedly; indeed, the slopes of the RTF were either positive or flat, whereas those of the SATF were clearly negative. Overall, the average slope for the SATF was -0.28, and the average slope for the RTF was 0.16. We compared the mean slopes of the SATF and the RTF by using an exact Wilcoxon signed rank test (to compensate for small sample size, $n = 8$). The test showed that the two quantities were reliably different (Wilcoxon $T = 1$, $p = .016$, two-tailed). Taken together, these analyses discount changes in criterion (speed–accuracy trade-off) as a plausible explanation of the recalibration found in CRTs. The data suggest instead that the listeners experienced a change in suprathreshold responsiveness across conditions, brought about by the different levels of stimulation to which the listeners were exposed at the two signal frequencies.

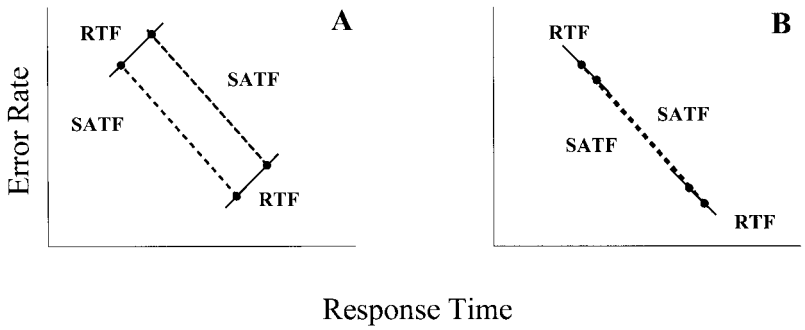

Figure 5. Possible relations between the speed–accuracy trade-off function (SATF) and the recalibration trade-off function (RTF). In Panel A, the slopes of the SATF and the RTF differ substantially, with the former being negative and the latter being positive. In Panel B, the slopes of the two functions coincide.

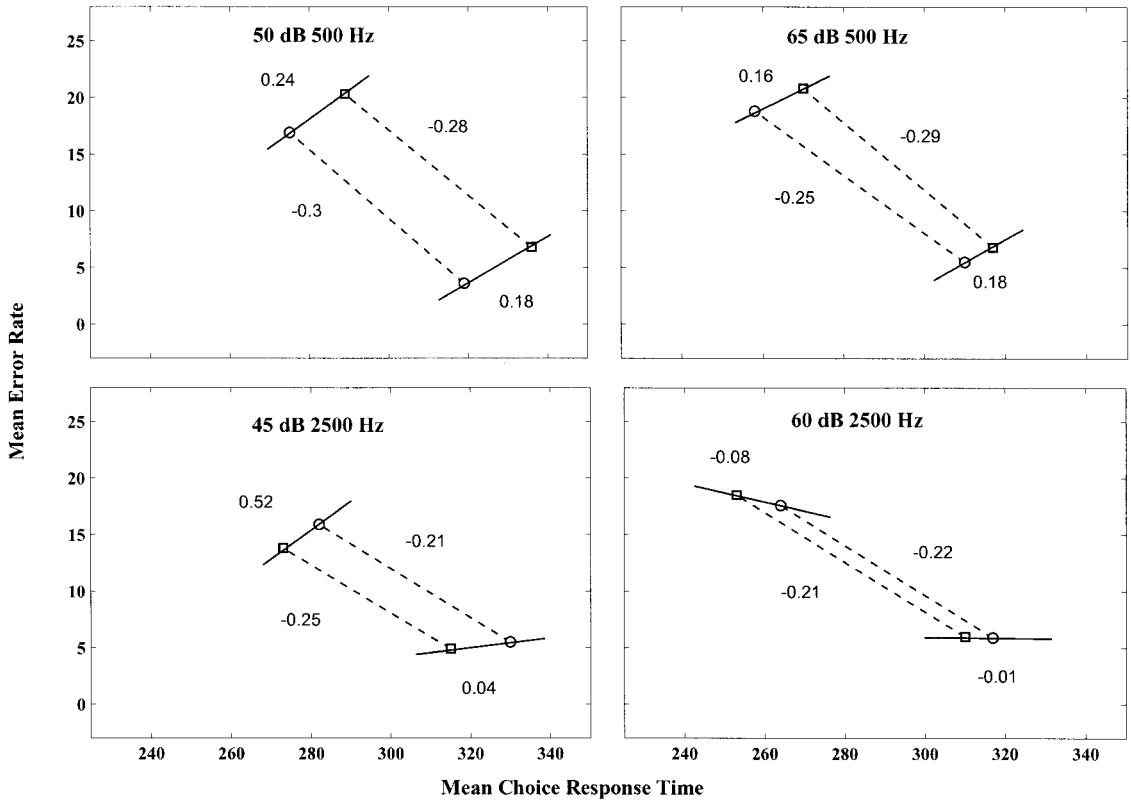

Figure 6. Mean error rates (percentages) plotted against mean choice reaction times (RTs; in milliseconds) for the 500-Hz signals at 50 and 65 dB and for the 2500-Hz signals at 45 and 60 dB. Dashed lines indicate the speed–accuracy trade-off function, which gives the relation between errors and RT across speeded and unspeeded instructions; within each panel, data obtained in Conditions A and B are shown by circles and squares, respectively. Solid lines indicate the recalibration trade-off function, which gives the relation between errors and RT across Conditions A and B; within each panel, data obtained with speeded and unspeeded instructions appear in the upper left and the lower right, respectively.

General Discussion

Let us summarize the study's two main findings: First, the phenomenon of recalibration, initially revealed in ratings of loudness and confirmed here in Experiment 1, also reveals itself in both SRTs (Experiment 2) and CRTs (Experiment 3). Second, in the case of choice responses, a speed–accuracy analysis—which allows us to distinguish sensory and decisional components—implies that recalibration is independent of decisional processes. Thus, we conclude that it represents stimulation-induced changes in the underlying processing and representations of sensory intensity.

The main characteristics of recalibration suggest that it can be characterized as an adaptation-like phenomenon: a decrease in responsiveness at a particular stimulus frequency resulting from prior intense stimulation. What distinguishes recalibration from other adaptation-like processes in hearing, such as auditory fatigue, is the evidence that recalibration requires at most only a few transient exposures (Marks, 1993)—perhaps just a single such exposure (e.g., Mapes-Riordan & Yost, 1999). Consider, for comparison, auditory fatigue. To raise the absolute threshold, it is necessary to present a fatiguing stimulus for several minutes at a relatively high SPL (>80 dB; Botte et al., 1993; Botte & Mönikheim, 1994). Some properties of recalibration and fatigue are similar, however; thus, recalibration, like fatigue, seems to depend on relatively strong stimulation (in hearing, SPLs > approximately 75 dB), although brief stimuli lasting approximately 0.25 s are sufficient to affect the loudness of subsequent stimuli (recalibration).

Numerous studies have elucidated various properties of recalibration, but a fundamental question until now remained unresolved: Does recalibration reflect a change in the underlying sensory representations of loudness (i.e., reductions in loudness at those frequencies receiving higher stimulus levels) or a change in decisional criteria (i.e., differential shifts in criteria at different signal frequencies for eliciting a given response)? Although earlier findings suggested that recalibration does not require overt ratings of loudness (Marks, 1992, 1993, 1994; Schneider & Parker, 1990), none of these studies provided direct tests of sensory versus decisional models. The present study aimed to compare sensory and decisional models directly, under two assumptions: first, that measures of RT, which often follows loudness, also show recalibration and, second, given this, that the mechanism underlying

recalibration in RT is the same as the mechanism that underlies recalibration in loudness.

To this end, Experiment 1 showed recalibration in judgments of loudness, using the same stimuli that we used in Experiments 2 and 3 to measure RTs. As we expected, the results of Experiment 1 showed substantial changes in the relative loudness judgments of 500-Hz and 2500-Hz tones when the ensemble of stimulus levels changed across conditions: Tones of 500 Hz were judged relatively louder than comparable tones of 2500 Hz when the listeners heard relatively soft 500-Hz tones and relatively loud 2500-Hz tones (Condition A), but they were judged relatively softer when the listeners heard relatively loud 500-Hz tones and relatively soft 2500-Hz tones (Condition B). Consistent with the general finding that SRTs to the onsets of sounds are inversely related to loudness, the results of Experiment 2 showed recalibration in SRTs. That is, SRTs to tones of 500 Hz were smaller in Condition A than in Condition B (consistent with greater loudness in Condition A), and SRTs to tones of 2500 Hz were smaller in Condition B than in Condition A (consistent with greater loudness in Condition B). Thus, SRTs, like judgments of loudness, exhibit recalibration.

Finally, the results of Experiment 3 revealed recalibration in RTs measured in a choice procedure in which participants categorized tones as low or high in frequency while the SPLs varied irrelevantly and took on different values in different conditions. Of crucial importance was the finding that the changes in RT measured in the choice task cannot be attributed to changes in decision criteria. If recalibration reflects frequency-specific changes in criteria, then shifts in CRT should be accompanied by corresponding shifts in error rates; that is, it should be possible to account for the recalibration in CRT in terms of the speed–accuracy trade-off. But the results suggest otherwise.

For each combination of frequency and SPL that was common to the two experimental conditions, we computed the relation between speed and accuracy (a) in the speeded session, in which participants were instructed to respond as quickly as possible with regard to accuracy, and in the unspeeded session, in which speed and accuracy were stressed equally, and (b) in experimental Conditions A and B within each session. Changing instructions across sessions produced the classic pattern of speed–accuracy trade-off. Participants paid for extra speed with elevated error rates, a pattern commonly interpreted as reflecting a shift in decision criteria. Presumably, in the speeded session, listeners lowered their criteria for what constituted enough evidence for one response alternative to be selected. Requiring less information for a decision may promote faster response, but it also increases the likelihood of error.

Changing the stimulus conditions, however, produced quite another speed–accuracy relation. When the stimulus set varied between Conditions A and B, and RTs shifted accordingly, these shifts in RTs were not accompanied by corresponding changes in error rates. We take the lack of speed–accuracy trade-off between experimental conditions as evidence against the hypothesis that recalibration can be attributed to differential shifts in criteria. Our findings are more readily accommodated by assuming that the representations of auditory intensity themselves were modified by the changes in stimulus levels at the two signal frequencies.

Recalibration was also evident in the simple detection task. When listeners were asked to press a key as soon as they heard a tone, they were faster at detecting tones for which frequencies were endowed with the lower SPLs. Sanford (1972) asked listeners to respond as quickly as possible to weak (72-dB) and strong (86-dB) noise bursts. In one condition, the strong noise was presented in 75% of the trials, and in another condition, the weak noise was presented in 75% of the trials. Listeners were faster at detecting both levels of the noise in the latter condition (see also Murray, 1970). Sanford interpreted this shift in RT as reflecting a shift in criteria across conditions resulting from the different presentation probabilities. Note, however, another possible explanation. The average SPL in the first condition, in which the strong noise prevailed, was greater than the average SPL in the second condition, in which the weak noise prevailed. Perhaps the longer RT reported in the second condition represented, at least in part, adaptation-like recalibration, similar to the recalibration found in the present study.

In reviewing studies concerning the effect of intensity on information processing, Nissen (1977) noted that the hallmark of criterial shifts in SRT tasks is the modulation of intensity effects across experimental conditions. This result follows directly from Grice's (1968) evidence-accumulation model. If weak and strong stimuli are represented by two information-accumulation functions, each with its own accumulation rate, then the functions will diverge in time. The closer the cutoff point to the function's origin (e.g., the lower the response criteria), the smaller the difference in RT between the strong and weak signals.

Did the effect of intensity vary as a function of experimental condition in our study? Comparing the differences in SRT to weak and strong signals at 500 Hz and at 2500 Hz across the two experimental conditions (Experiment 2), we saw that the change in conditions had almost no effect. The difference between SRTs at 500 Hz was 19 ms in Condition A and 17 ms in Condition B, and the difference at 2500 Hz was 17 ms in both conditions. The lack of interaction between the intensity effect and the experimental condition agrees nicely with the main results of Experiment 3 in that both show no evidence of substantial shifts in criteria across conditions.

It is especially worth noting that recalibration occurs not only in hearing but in several other sensory modalities, including vision (Armstrong & Marks, 1997), haptic perception (Marks & Armstrong, 1996), taste (Rankin & Marks, 1991, 2000), and olfaction (Rankin & Marks, 2000). Further, as in hearing, recalibration has been reported with direct comparison methods as well as rating methods (cf. Armstrong & Marks, 1997; Marks & Armstrong, 1996). In every case, recalibration seems to represent some kind of stimulus-specific adaptation: When people judge or compare the magnitudes of two kinds of stimuli (loudness of sounds at different frequencies, lengths of lines in different orientations, perceived extent of arm movement in different directions, taste intensity or olfactory intensity of different compounds), recalibration follows a simple, uniform principle: Whichever of the two kinds of stimuli is presented at greater levels of physical intensity has its perceived magnitude reduced. So, in hearing, when most 500-Hz tones are higher in SPL than most 2500-Hz tones, a 500-Hz tone is perceived as softer and a 2500-Hz tone as louder than they are when the relative intensities at the two frequencies are switched (e.g., Experiment 1). Similarly, when most horizontal lines are long relative to vertical lines, a given vertical line is perceived as longer and a horizontal line as shorter than they are when the physical lengths in the two orientations are switched (Armstrong & Marks,

1997; Marks & Armstrong, 1996). Recalibration appears to be a widespread characteristic of sensory–perceptual processing.

It is also noteworthy, however, that recalibration is far from universal, being absent from judgments of the pitch of tones (presented at different levels of loudness), of length of lines (presented in different colors), and of duration of tones (presented at different pitches; Marks, 1992). Its lack of universality speaks against the hypothesis that recalibration arises from a general decisional process.

The precise conditions in which recalibration occurs, and in which it does not occur, may provide significant clues to the underlying mechanisms. Consider the explanation for recalibration in loudness perception offered by Schneider and Parker (1990; see also Parker, Murphy, & Schneider, 2002; Parker & Schneider, 1994). These investigators hypothesized the presence of a nonlinear gain control mechanism in the auditory system that operates in a top-down fashion. The operation of such a device hinges on the presence of loud tones in a particular auditory critical band; the device might attenuate auditory signals in their presence, or it might amplify auditory signals in their absence. Either way, the model could account for the changes in loudness matches. Although we find it more convenient to describe recalibration as a reduction in loudness in a frequency band that has received intense stimulation, it is equivalent, although perhaps more awkward, to propose that frequency bands are amplified under weak stimulation, or no stimulation, but lose the amplification under intense stimulation.

Other findings speak to the sensory or perceptual channels in which recalibration arises. Consider recalibration in the perception of visual perceived length of line segments. Whereas the perceived length of lines presented in different orientations depends on the physical distribution of line lengths in those orientations, the perceived length of lines presented in different colors, with orientation fixed, does not depend on the distribution of lines in those colors (Marks, 1992). These results are consistent with the hypothesis that line-length recalibration results from an adaptation-like process that is orientation specific (in at least the early stages of processing line length in the visual cortex, representations are segregated by orientation) but not color specific (with no color-dependent channels processing length). Recently, Arieh and Marks (2002) showed that orientation-dependent recalibration of perceived line length does not transfer between the eyes or between adjacent regions of the retina. These findings too are consistent with processes located early in the visual pathway, and they are readily compatible with the main conclusion of the present study: namely, that auditory recalibration reflects changes based on sensory and not decisional processes.

A similar interpretation applies readily to recalibration in the chemical senses. Rankin and Marks (2000) had participants judge the intensities of chemosensory compounds for which similarity or dissimilarity was in some cases independent of the particular chemosensory modality (gustatory or olfactory) that was activated. This dissociation was accomplished by having participants sip water that could contain a solute that stimulates the gustatory system (e.g., sucrose or citric acid) or one that retronasally stimulates the olfactory system (e.g., vanillin or orange). Because some gustatory and olfactory stimuli (e.g., sucrose and vanillin) are perceived as similar, whereas some pairs of gustatory stimuli (e.g., sucrose and citric acid) are perceived as dissimilar, similarity could be dissociated from communality of sensory processing. The results showed substantial recalibration in intensity judgments whenever different modalities (receptors) were stimulated, even when the stimuli were judged qualitatively similar, leading Rankin and Marks to conclude that chemosensory recalibration reflects changes in sensory systems and not the application of different response criteria to perceptually dissimilar stimuli.

Recalibration takes place in vision, when lines in different orientations take on different mean physical lengths; in the chemical senses, when gustatory and olfactory stimuli activating different quality channels take on different mean concentrations; and in hearing, when different sound frequencies (in different critical bands, preferentially activating different populations of auditory receptors) take on different mean SPLs. Assuming that the sensory channel processing the stimuli at the higher intensities undergoes more adaptation, the observed pattern of results can be predicted. Presenting greater SPLs at 500 Hz than at 2500 Hz preferentially adapts the auditory system to signals in the region of 500 Hz, so that, relative to 2500 Hz, both loudness and speed of processing diminish. Given the overriding similarity of results across different modalities, it is not unreasonable to propose, at least tentatively, a general hypothesis: Recalibration results not from decisional processes but from adaptation-like processes operating within separate channels of a sensory system.

References

Arieh, Y., & Marks, L. E. (2002). Context effects in visual length perception: Role of ocular, retinal, and spatial location. *Perception & Psychophysics, 64,* 478–492.

Armstrong, L., & Marks, L. E. (1997). Stimulus context, perceived length, and the horizontal–vertical illusion. *Perception & Psychophysics, 59,* 1200–1213.

Botte, M. C., Charron, S., & Bouayad, H. (1993). Temporary threshold and loudness shifts: Frequency patterns and correlations. *Journal of Acoustical Society of America, 93,* 1524–1534.

Botte, M. C., & Mönikheim, S. (1994). New data on short-term effects of tone exposure. *Journal of Acoustical Society of America, 95,* 2598–2605.

Cattell, J. M. (1886). The influence of the intensity of the stimulus on the length of reaction time. *Brain, 8,* 512–515.

Chocholle, R. (1940). Variations des temps de réaction auditifs en fonction de l'intensité à diverses fréquences [Variations in auditory reaction times as a function of intensity at different frequencies]. *L'Année Psychologique, 41,* 65–124.

Chocholle, R., & Greenbaum, H. B. (1966). La sonie de sons purs partiallement masqués: Étude comparative par une méthode d'égalisation et par la méthode des temps de reaction [Loudness of partially masked pure tones: Comparative study by an equalization method and by the reaction time method]. *Journal de Psychologie Normale et Pathologique, 63,* 387–414.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of Acoustical Society of America, 5,* 82–108.

Green, D. M., & Luce, R. D. (1971). Detection of auditory signals presented at random times. *Perception & Psychophysics, 9,* 257–268.

Grice, G. R. (1968). Stimulus intensity and response evocation. *Psychological Review, 75,* 359–373.

Keuss, P. J. G., & van der Molen, M. W. (1982). Positive and negative effects of stimulus intensity in auditory reaction tasks: Further studies on immediate arousal. *Acta Psychologica, 52,* 61–72.

Kohfeld, D. L. (1971). Simple reaction time as a function of stimulus in

decibels of light and sound. *Journal of Experimental Psychology, 88,* 251–257.
Kohfeld, D. L., Santee, J. L., & Wallace, N. D. (1981). Loudness and reaction time: I. *Perception & Psychophysics, 29,* 535–549.
Luce, R. D. (1986). *Response times.* New York: Oxford University Press.
Luce, R. D., & Green, D. M. (1972). A neural timing theory for response time and the psychophysics of intensity. *Psychological Review, 79,* 14–57.
Mapes-Riordan, D., & Yost, W. A. (1999). Loudness recalibration as a function of level. *Journal of Acoustical Society of America, 106,* 3506–3511.
Marks, L. E. (1988). Magnitude estimation and sensory matching. *Perception & Psychophysics, 43,* 511–525.
Marks, L. E. (1992). The slippery context effects in psychophysics: Intensive, extensive, and qualitative continua. *Perception & Psychophysics, 51,* 187–198.
Marks, L. E. (1993). Contextual processing of multidimensional and unidimensional auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 227–249.
Marks, L. E. (1994). "Recalibrating" the auditory system: The perception of loudness. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 382–396.
Marks, L. E. (1996). Recalibrating the perception of loudness: Interaural transfer. *Journal of the Acoustical Society of America, 100,* 473–480.
Marks, L. E., & Armstrong, L. (1996). Haptic and visual representations of space. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI: Integration information in perception and communication* (pp. 263–287). Cambridge, MA: MIT Press.
Marks, L. E., & Warner, E. (1991). Slippery context effect and critical bands. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 986–996.
Murray, H. G. (1970). Stimulus intensity and reaction time: Evaluation of a decision-theory model. *Journal of Experimental Psychology, 84,* 383–392.
Nissen, M. J. (1977). Stimulus intensity and information processing. *Perception & Psychophysics, 22,* 338–352.
Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. Kantrowitz (Ed.), *Human information: Tutorials in performance and cognition* (pp. 41–82). Hillsdale, NJ: Erlbaum.
Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.
Parker, S., Murphy, D. R., & Schneider, B. A. (2002). Top-down gain control in the auditory system: Evidence from identification and discrimination experiments. *Perception & Psychophysics, 64,* 598–615.
Parker, S., & Schneider, B. A. (1994). The stimulus range effect: Evidence for top-down control of sensory intensity in audition. *Perception & Psychophysics, 56,* 1–11.
Pew, R. W. (1969). The speed–accuracy operating characteristics. *Acta Psychologica, 30,* 16–26.
Pins, D., & Bonnet, C. (1996). On the relation between stimulus intensity and processing time: Piéron's law and choice reaction time. *Perception & Psychophysics, 58,* 390–400.
Rankin, K. M., & Marks, L. E. (1991). Differential context effects in taste perception. *Chemical Senses, 16,* 617–629.
Rankin, K. M., & Marks, L. E. (2000). Chemosensory context effects: Role of perceived similarity and neural commonality. *Chemical Senses, 25,* 747–759.
Sanford, A. J. (1972). Criterion effects in simple reaction time: Results with stimulus intensity and duration manipulations. *Journal of Experimental Psychology, 95,* 370–374.
Scharf, B. (1983). Loudness adaptation. In J. V. Tobias & E. D. Schubert (Eds.), *Hearing: Research and theory* (Vol. 2, pp. 1–56). New York: Academic Press.
Schneider, B., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgment? *Perception & Psychophysics, 48,* 409–418.
Stevens, J. C., & Marks, L. E. (1980). Cross-modality matching functions generated by the method of magnitude estimation. *Perception & Psychophysics, 27,* 379–389.
Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Stevens's law. *American Journal of Psychology, 86,* 3–27.
Thomas, E. A. C. (1967). Reaction time studies: The anticipation and interaction of responses. *British Journal of Mathematical and Statistical Psychology, 20,* 1–29.
van der Molen, M. W., & Keuss, P. J. G. (1979). The relationship between reaction time and auditory intensity in discrete auditory tasks. *Quarterly Journal of Experimental Psychology, 31,* 95–102.
Ward, W. D. (1973). Adaptation and fatigue. In J. Jerger (Ed.), *Modern developments in audiology* (2nd ed., pp. 301–344). New York: Academic Press.
Yellott, J. I., Jr. (1971). Correction for guessing and the speed–accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology, 8,* 159–199.

Received January 18, 2002
Revision received July 23, 2002
Accepted October 9, 2002 ■