

Chapter 2

Measurement of Loudness, Part I: Methods, Problems, and Pitfalls

Lawrence E. Marks and Mary Florentine

2.1 Introduction

It is a matter of everyday experience that sounds vary in their perceived strength, from the barely perceptible whisper coming from across the room to the overwhelming roar of a jet engine coming from the end of an airport runway. Loudness is a salient feature of auditory experience, closely associated with measures of acoustical level (energy, power, or pressure) but not identical to any of them. It is a relatively straightforward matter for a person to note whether one sound is louder or softer than another, or to rank order a set of sounds with regard to their loudness. To measure loudness, however, in the typical sense of “measuring,” requires more than just ranking the experiences from softest to loudest. It entails quantifying how much louder (e.g., determining whether the ratio or difference in the loudness of sounds A and B is greater or smaller than the ratio or difference in loudness of sounds C and D).

The quantitative measurement of loudness in this sense is important both to basic research and to its applications – important to scientists seeking to understand neural mechanisms and behavioral processes involved in hearing and to scientists, engineers, and architects concerned with the perception of noise in factories and other industrial settings, in the streets of urban centers, and in residences located along flight paths and near airports. As Laird et al. (1932) wrote more than three-quarters of a century ago, in an article describing one of the earliest attempts to quantify the perception of loudness,

When a considerable amount of money is to be appropriated for making a work place quieter, for instance, the engineer can say that after acoustical material is added the noise level will be reduced by five or ten decibels. “But how much quieter will that make the office,” is likely to be the inquiry. “A great deal” is not only an unsatisfactory but an unscientific answer. (p. 393)

L.E. Marks(✉)

John B. Pierce Laboratory, Department of Epidemiology and Public Health, Yale University School of Medicine, and Department of Psychology, Yale University New Haven, CT 06519, USA
e-mail: marks@jbpierce.org

What is called for both scientifically and practically is a quantitative assessment of the change in loudness, such as knowing that reducing the physical level of environmental noise by a specified amount will reduce its perceived strength, its loudness, by 50%.

The present chapter focuses on methods for the measurement of loudness. Luce and Krumhansl (1988) pointed out that the psychophysical analysis of sensory measurement may operate at any one of three distinct levels. One level is mathematical, and it deals with the development of appropriate axioms for the numerical representations entailed by scales of sensory measurement. The second level is theoretical, and it deals with the structure of relations among scales of measurement. The third and last level is empirical, and it deals with the sensory relations expressed through the measurements. This third level treats sensory/perceptual measurement from a functional and pragmatic perspective, and it lies at the heart of the present chapter. From this perspective, the measurement of loudness is useful and valuable to the extent that it sheds light on basic mechanisms of hearing or makes it possible to predict responses to sounds in real-world settings.

Research over the past century and a half has developed and refined several approaches to measure loudness. This chapter summarizes the main approaches, evaluating the principles that underlie the application of each method and assessing the theoretical and practical problems that each approach faces – in essence, identifying the strengths and weaknesses of each approach. The chapter does not attempt to review the long-standing, often philosophically oriented, debates as to whether and how perceptual experiences may be quantified, but operates on the pragmatic assumption that quantification is not only possible but also scientifically meaningful and important; readers interested in the debates over quantification are directed elsewhere (see Savage 1970; Laming 1997; Marks and Algom 1998). The chapter starts with a brief history of loudness measurement in the nineteenth and twentieth centuries. Understanding this history is important because many of the concepts developed in the twentieth century resound in current scientific literature. Errors made in the interpretation of loudness data in the twenty-first century may arise from ignorance of these basic concepts regarding methods of measuring loudness. After the historical review, the reader is introduced to the theoretical, empirical, and practical constraints on loudness measurement.

2.2 A Brief History of Loudness Measurement

The history of loudness measurement is divided into two parts. The first part covers nineteenth century work by Fechner, Delboeuf, and others that raised the psychophysical problem of measuring loudness. The second part covers early twentieth century attempts to measure loudness by Piéron, Richardson and Ross, and Stevens.

2.2.1 *Measurement of Loudness: Recognizing the Psychophysical Problem*

The first steps toward measuring the loudness of sounds, and the magnitudes of other perceptual events, came in the second half of the nineteenth century with increased awareness on the part of sensory physiologists, psychologists, and physicists of what might be called “the psychophysical problem of intensity” – that the perceived magnitude of a perceptual experience need not be quantitatively proportional to the magnitude of the physical stimulus that evokes the experience. The experience of loudness is distinct from the physical measure of the stimulus. Nevertheless, it was sometimes assumed that that physical magnitude and perceived strength were commensurate, such that loudness was directly proportional to the physical magnitude of a sound. For example, Johann Krüger (1743) derived a simple rule of proportionality between the intensity of sensations and the intensity of the physical stimuli that produce the sensations. A century later, in his *Elements of Psychophysics*, Gustav Fechner recognized that direct proportionality flies in the face of direct experience. “I found it very interesting to hear the statement,” wrote Fechner (1860/1966), “... that a choir of 400 male voices did not cause a significantly stronger impression than one of 200” (p. 152). The average (root-mean-square) acoustic power associated with a choir of 400 voices should be, in principle, about twice that associated with a comparable one of 200 voices. Yet the difference in the experience of loudness is not nearly so great as two-to-one.

To be sure, Fechner was not the first to make or recognize a distinction between perceptual experiences and the corresponding properties of stimulus events responsible for producing those experiences; the distinction goes back more than two millennia, at least as far as Democritus’s famous dictum in the fifth century BCE, which states, “Sweet exists by convention, bitter by convention, color by convention; atoms and Void (alone) exist in reality.... We know nothing accurately in reality, but (only) as it changes according to the bodily condition, and the constitution of those things that flow upon (the body) and impinge upon it” (Freeman 1948, p. 110). Two millennia later, Locke (1690) noted that what he called secondary physical qualities of objects are not the same as our perceptions of them. Locke’s distinction underlies the philosophical problem of sensory qualia – a topic that falls outside the scope of the present chapter (for a scientifically informed philosophical account, see Clark 1993).

Fechner was among the first, however, to recognize that there may be quantitative as well as qualitative differences between stimuli and sensations. In particular, Fechner pointed to quantitative differences between changes in the physical intensity of a stimulus and corresponding quantitative changes in the perceptual experience of it. He addressed the question of how perceived strength depends on physical intensity, stating that the intensity of sensation is proportional to the logarithm of physical intensity, when physical intensity is reckoned in units equal to the absolute threshold. This was his famous psychophysical law.

The first inklings of Fechner’s logarithmic law came to him from philosophical and, later, mathematical intuition. He saw how he could derive the law, and hence

derive measures of perceived magnitude, including loudness, from measures of the ability to discriminate two sounds. In fact, Fechner described how one could both derive the logarithmic law formally, from Weber's law of intensity discrimination, with the help of subsidiary theoretical and mathematical assumptions, and reveal the law empirically, by what is essentially a graphical procedure for summing discrimination thresholds [just-noticeable-differences (JNDs) in stimulation].

Fechner's proposal to construct scales of sensation from measures of discrimination rested in part on his view that sensation magnitudes could not be assessed accurately, in numerical fashion, by direct introspection, at least not in a scientifically meaningful way (although contemporaries of Fechner did take small steps in this direction, e.g., Merkel 1888). Fechner did, however, consider the possibility that intervals or differences in sensation magnitude might be compared directly, and investigators in the late nineteenth and early twentieth centuries began to develop and test several methods for producing sensory scales with equal-appearing intervals (e.g., Delboeuf 1873). One of these methods came to be called the method of bisection. In the method of bisection, a subject adjusts the level of a stimulus to appear midway between fixed upper and lower stimulus levels. Without modern technology, however, it was difficult to create an experiment in which subjects could adjust the physical levels of sounds in a controlled, continuous fashion, and it was especially difficult to measure the resulting sound levels even if one could vary them. The development of vacuum-tube technology in the early decades of the twentieth century provided the needed impetus.

As elegant as it is, Fechner's approach to sensory measurement in general and to the measurement of loudness in particular has not proven especially useful. The approach is exceedingly laborious to apply – a criticism that also applies to the approach of Thurstone (1927), which requires many pairwise comparisons of relative intensity of all possible pairs of stimuli. Thurstonian measurement is not reviewed here, but the interested reader is directed to other summaries (Marks and Algom 1998; Marks and Gescheider 2002), as well as to evaluations of Thurstone's conceptualizations in the development of sensory measurement (Luce 1994). Even more importantly, Fechner's approach often produces results that fail tests of internal consistency. If the number of JNDs above absolute threshold can serve as a fixed unit of loudness, as discussed in the next section, then all pairs of sounds that lie equal numbers of JNDs above threshold should be equally loud. Considerable evidence contradicts this principle. Nevertheless, because modern versions of Fechner's approach still have proponents (e.g., Falmagne 1985; Link 1992; Dzhafarov and Colonius 2005), a review and analysis are appropriate.

2.2.1.1 Fechner's Law and Fechnerian Measurement

Fechner (1860) reported that on the morning of 22 October 1850, he first conjectured that a logarithmic function might relate the magnitude of sensation to physical intensity. This conjecture actually preceded his discovery of empirical evidence supporting it. Having come to the putative insight that sensation increases as a

logarithmic function of stimulus intensity, Fechner then came upon Weber's work on sensory discrimination, and this discovery led Fechner to develop both a mathematical derivation for the logarithmic law and a more general, empirical method for generating quantitative psychophysical functions, logarithmic or otherwise. Although Fechner's law was eventually replaced, as described later, the proposal of the law itself marked a watershed moment in the history of psychophysics, which the International Society of Psychophysics celebrates every year at its annual meeting.

To derive the logarithmic law mathematically, Fechner relied first on the generalization that has come to be called Weber's law, and second on two auxiliary mathematical assumptions. Extensive experimentation by both Weber and Fechner on intensity discrimination focused on measures of the JND, that is, the smallest difference between stimulus intensities that a person is able to distinguish. (The JND is also known as the difference limen [DL].) Fechner showed how JNDs could provide the building blocks for scales of sensation magnitude. Much of the data reported by Weber and Fechner conforms at least loosely to Weber's law, which states that if I is the baseline intensity from which a change in the stimulus is made, then the minimal change in I that is perceptible, ΔI (the JND), is proportional to I . That is,

$$\Delta I = k_1 I \quad (2.1)$$

An assumption critical to both Fechner's mathematical approach and his experimental approach is the subjective equality of JNDs – the assumption that all JNDs have the same psychological magnitude. If L is sensation magnitude such as loudness, then, for every JND, ΔL is constant, that is,

$$\Delta L = k_2 \quad (2.2)$$

A second assumption, critical to the mathematical derivation, though not to the empirical approach, is that one can convert the difference equations (2.1) and (2.2) into differential equations. Converting (2.1) and (2.2) and rearranging the terms leads to:

$$dI / (k_1 I) = 1 \quad (2.3a)$$

$$dL / k_2 = 1 \quad (2.3b)$$

Combining (2.3a) and (2.3b) and integrating in turn leads to Fechner's law:

$$L = k \log I + k_3 \quad (2.4)$$

where $k = k_2/k_1$.

Fechner further assumed that sensation magnitude takes on positive values only when the intensity of the stimulus, I , exceeds the absolute threshold. Consequently, by measuring I in terms of the absolute threshold, I_0 , $k_3 = 0$, and one can write:

$$L = k \log (I / I_0) \quad (2.5)$$

Measured in this way, loudness, L , would have the properties of a ratio scale (Stevens 1946): A sound having a loudness, L , of 10 units (ten JNDs above threshold) would be twice as loud as a sound having a loudness, L , of 5 units (five JNDs above threshold). Without fixing the starting-point of the sensation scale, one would only be able to compare differences or intervals along the scale, but not ratios.

Fechner's model is elegant, but as Luce and Edwards (1958) showed, his general approach provides mathematically consistent results only when intensity discrimination (level discrimination) follows a limited number of formulas, such as Weber's law ($\Delta I = k_1 I$) and its linearization ($\Delta I = k_1 I + \text{constant}$). The approach fails mathematically, for example, when auditory intensity discrimination follows what has been called a "near miss" to Weber's law, as shown in results of many studies (e.g., McGill and Goldberg 1968; Jesteadt et al. 1977; Florentine et al. 1987; see Parker and Schneider 1980; Schneider and Parker 1987). The near miss may be written as

$$\Delta I = k I^b \quad (2.6)$$

where b is smaller than 1.0, often having a value around 0.8–0.9.

The empirical approach to Fechnerian measurement, however, avoids these complications because the approach may be used to generate a Fechnerian scale from any set of intensity-discrimination data, regardless of whether Weber's law holds. Taking the empirical approach, one would proceed as follows: first, define as L_0 the sensation magnitude (e.g., loudness) associated with baseline intensity I_0 . Second, measure the JND, ΔI_1 , from baseline I , and then define the sensation magnitude of intensity $I_2 (= I + \Delta I_1)$ as $L_0 + 1$. Next, starting from intensity I_2 , measure the subsequent JND, ΔI_2 , and define the sensation magnitude of $I_2 (= I_0 + \Delta I_1 + \Delta I_2)$ as $L_0 + 2$; and so forth. This approach essentially builds up a measurement scale, under the assumption that each additional step of stimulus intensity, calculated as a JND, adds another unit of sensation magnitude.

Most studies of intensity discrimination do not use Fechner's adaptive approach, but measure JNDs using a predetermined set of starting intensities. Nevertheless, given a fixed set of stimulus intensities, it is possible to derive a reasonable empirical approximation to a Fechnerian function by interpolating values along the empirical discrimination function and then summing the inferred JNDs.

Figure 2.1 shows an example – a Fechnerian loudness scale derived from intensity-discrimination data at sound frequencies of 200, 400, 600, 800, 1,000, 2,000, 4,000, and 8,000 Hz, as reported by Jesteadt et al. (1977). In their experiment, Jesteadt et al. measured intensity discrimination at each of the eight frequencies at intensity levels of 5, 10, 20, 40, and 80 dB above threshold (sensation level, SL), omitting 80-dB SL at 200 Hz. The entire ensemble of results could be described by a single equation consistent with the "near miss" to Weber's law given in (2.6):

$$\Delta I = 0.463 (I / I_0)^{0.928} \quad (2.7)$$

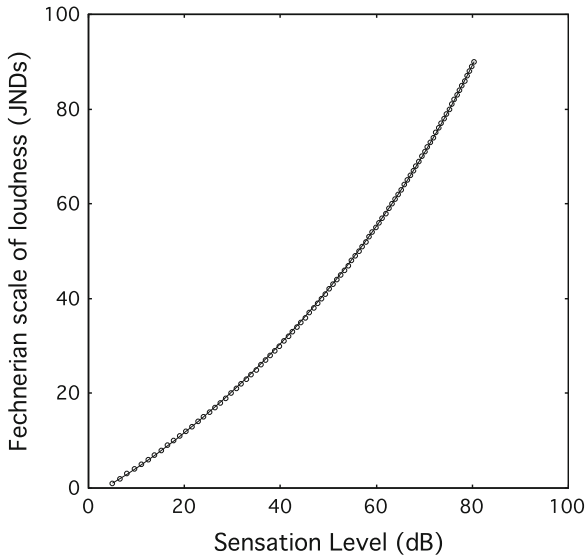


Fig. 2.1 A scale for loudness constructed from measures of just-noticeable differences in sound intensity at eight sound frequencies over the range 200–8,000 Hz (based on data and analysis of Jesteadt et al. 1977)

where I_0 is the reference for SL at each frequency. The Fechnerian function shown in Fig. 2.1 was constructed empirically, by summing JNDs as calculated from (2.7). If the discrimination data were consistent with Weber’s law instead of its near miss, then the Fechnerian function derived by summing JNDs would follow a straight line. Instead, because of the “near miss,” the derived function curves upward when plotted against SL in decibels. These derived data can be fitted, as shown, by a power function with an exponent of 0.129 (re: sound pressure; 0.0645 re: sound energy).

2.2.1.2 Fechnerian Loudness and the Principle of Equality

In Fechner’s terms, the function shown in Fig. 2.1 would characterize the relation between loudness and sound intensity, applicable over a wide range of sound frequencies. Because the function is based on (2.7), which applies to frequencies from 200 to 4,000 Hz (see Florentine et al. 1987), Fechnerian loudness would vary directly with the ratio I/I_0 at all frequencies, which means that loudness would vary directly with sensation level (SL, i.e., the number of decibels above threshold), given that SL equals $10 \log(I/I_0)$. This is to say, that if the Fechnerian function shown in Fig. 2.1 represents loudness, then, according to the principle of equality, all sounds at a given SL (at least between 200 and 4,000 Hz) should be equally loud.

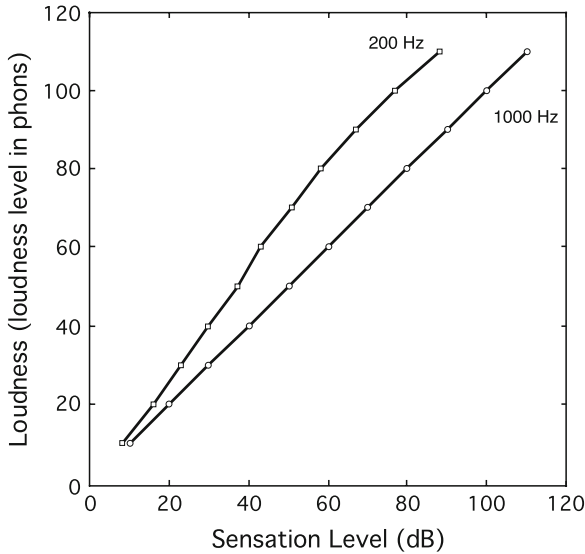


Fig. 2.2 The level in decibels above threshold of a 1,000-Hz tone (ordinate) that sounds as loud as test tone of 200 and 1,000 Hz at various levels above their threshold (abscissa). The data points are taken from curves appearing in Fig. 3 of Fletcher and Munson (1933). These failure of the points at 200 and 1,000 Hz to overlap contradicts the conjecture that decibels above threshold can serve as a uniform scale to quantify loudness

This inference is incorrect, as shown by equal-loudness relations determined across the sound spectrum. It has long been known that loudness depends on acoustic frequency as well as sound level (Fletcher and Munson 1933; Robinson and Dadson 1956; see Chap. 5, or ISO standard 226 2003). Figure 2.2 uses a subset of Fletcher and Munson’s data to show how SL, or decibels above threshold, fails to meet the criterion of internal consistency, hence fails to provide an adequate measure of loudness. For a decibel to serve as a universal unit of loudness, the principle of equality requires that all acoustic signals 20 dB above threshold appear equally loud, all signals 30 dB above threshold appear equally loud, and so forth.

Figure 2.2 shows one reason why this prediction fails. A tone having a frequency of 1,000 Hz and a level that is 60 dB above its threshold would be assigned a loudness that equals, by definition, 60 (loudness = decibel) units. But, as determined by equal loudness matching, a tone having a frequency of 200 Hz that lays 60 dB above its threshold appears much louder, equal to 80 loudness units. In general, increasing sound intensity by a fixed number of decibels above threshold produces greater increments in loudness at 200 Hz than at 1,000 Hz. Thus, loudness matches obtained across the spectrum make it possible to eliminate one possible method for measuring loudness – in terms of the number of decibels above threshold.

Considered across sound frequency (Newman 1933; Ozimek and Zwislacki 1996), across masking conditions (Hellman et al. 1987; Johnson et al. 1993), and

across normal hearing and hearing loss (Zwislocki and Jordan 1986; Stillman et al. 1993), JNDs fail to provide a constant unit of loudness. This failure was recognized early by Riesz (1933), who proposed a possible solution to the failure of JNDs to provide a constant unit of loudness across sound frequency. Riesz suggested that, at every sound frequency, one may ascertain the range of loudness from bottom to top and then determine the number of JNDs in this range. Once this is done, loudness at each frequency, according to Riesz, would depend directly on the fraction of the total number of JNDs to that point. This has been called a proportional-JND hypothesis, a view later considered by Lim et al. (1977). The status of this modified Fechnerian hypothesis, however, remains uncertain (see Houtsma et al. 1980). It has not been rigorously determined, for example, whether or to what extent the proportional-JND hypothesis could account for loudness of tones heard in quiet and in masking noise, or tones heard by listeners with normal hearing and with hearing losses characterized by abnormally rapid or slow loudness growth of loudness with increasing level (Florentine et al. 1979). In any case, while of theoretical interest, the approaches using Fechnerian and Thurstonian methods are impractical.

2.2.2 Early Attempts to Measure Loudness in the Twentieth Century

Four approaches to measuring loudness in the early twentieth century are noteworthy. These approaches, described in the following sections, are: (1) measurement through decibels, (2) measurement through reaction times, (3) measurement through additivity, and (4) measurement through judgments of ratios or magnitudes.

2.2.2.1 Fechner's Law and the Use of Decibels to Measure Loudness

Fechner's approach, and in particular his logarithmic law, helped propel the study of loudness measurement in the early decades of the twentieth century – especially with the widespread use of the decibel notation for representing relative values of sound intensity or sound pressure. The decibel (dB) scale is a logarithmic transformation of stimulus power or pressure, as is Fechner's scale of sensations. By implication, if Fechner were correct, then the decibel scale might serve as a scale or measure of loudness. As Fletcher and Munson (1933) noted, "In a paper during 1921 one of us suggested using the number of decibels above threshold as a measure of loudness...." (p. 82). Indeed, with zero decibels (0 dB) set at the absolute threshold, a decibel scale of loudness should have numerical ratio properties: A sound 80 dB above threshold would have twice the loudness of a sound 40 dB above threshold.

All of this seemed reasonable enough at first, except that direct experience contradicted the inference. As Churcher (1935) wrote, "... the experience of the

author and his colleagues over many years is that the numbers assigned by the decibel scale to represent sensation magnitudes are not acceptable to introspection as indicating their relative magnitudes.... The loudness of the noise of a motor assessed at 80 dB above threshold ... is, to introspection, enormously greater than twice that of a motor assessed at 40 dB” (p. 217). Whereas a choir of 400 voices appears only slightly louder than a choir of 200, a motor producing 100,000,000 (threshold) units of acoustical power (80 dB above threshold) sounds far more than twice as loud as motor producing 10,000 units (40 dB above threshold). Of course, the preceding analysis is predicated on the assumption, among others, that loudness is zero at absolute threshold. For several reasons, this is highly unlikely. Evidence indicates that threshold-level sounds have small positive values of loudness (see Buus et al. 1998). Even so, Churcher’s point remains valid: Decibels serve poorly as direct indicators of loudness.

Psychoacoustic research in the early decades of the twentieth century, and especially from 1930 onward, sought to quantify loudness in ways that would be commensurate with direct experience and that also would satisfy basic scientific principles of measurement. Three subsequent approaches were important, each of which sought in its own unique way to develop a score of loudness: (1) using speed of response as a surrogate measure for loudness, (2) building a scale on the basis of additivity, and (3) building a scale from overt judgments of loudness ratios. A fourth approach, estimating perceived magnitudes, originated during this same period and became important only in the second half of the last century. Each approach is described in the following sections.

2.2.2.2 Measuring Loudness from Response Times: Piéron’s Law

One measure of sensory performance is the speed of response to a stimulus. Beginning at least with the report of Cattell (1886), it has been clear that as the level of a stimulus increases, the response time decreases. Nearly a century ago, Piéron (1914) suggested that response speed, the inverse of response time, might serve as a surrogate measure of sensation intensity (see Piéron 1952, for a later summary; for recent reviews, see Wagner et al. 2004 and Chap. 4). Piéron reported the results of a systematic study of the way that response time varies with physical intensity in several modalities, including hearing. In each case, Piéron concluded that response time decreased as a power function of stimulus intensity, writing an equation of the form

$$RT - R_0 = al^{-m} \quad (2.8)$$

where RT is the response time for the particular stimulus and modality, m is the exponent, and R_0 is the “irreducible minimum” RT , representing the asymptote of the function as I becomes very large. The parameter R_0 presumably represents the minimal time needed to prepare and execute the response. Subsequent research has confirmed that a power function of the form expressed in (2.8) provides a good description

to measures of simple *RTs* to acoustic stimuli varying in level (e.g., McGill 1961; Kohfeld 1971; Luce and Green 1972; Kohfeld et al. 1981a, b). Luce and Green developed a mathematical model to show how loudness and *RT* could be related through a hypothesized dependence of both variables on mechanisms of neural timing. McGill (1961) pointed out, however, that the values of exponents fitted to functions for auditory *RT* generally differ markedly from the values of exponents derived from direct estimates of loudness, especially magnitude estimations, although the exponents derived from *RT* agree better with exponents estimated from measures of loudness derived from judgments of differences or intervals. Exponents derived from measures of *RT* generally have values around 0.3 when the stimulus is reckoned in terms of sound pressure, 0.15 when reckoned in terms of sound energy or power (see Marks 1974b, 1978).

As we asked about decibel measures, so too may we ask about *RT*: Do sounds that are equally loud produce the same response times? Often, this is approximately the case. But violations of the principle of equality have been reported, for instance, in the *RTs* given to tones heard in the quiet vs. backgrounds of masking noise (Chocholle and Greenbaum 1966) and in the *RTs* given to tones of different frequencies (Kohfeld et al. 1981a; Epstein and Florentine 2006b). In particular, Kohfeld et al. reported that equally loud, low intensity tones gave similar *RTs*, but not identical ones.

2.2.2.3 Measuring Loudness by Additivity: Fletcher and Munson's Loudness Scale

Fletcher and Munson (1933) offered a novel approach to the measurement of loudness, which served as a powerful conceptual alternative to Fechner's. Fletcher and Munson sought to create a scale for loudness that was both internally consistent and grounded in a principle of additivity. Internal consistency was ensured empirically by matching all sounds in loudness to a common yardstick, a tone at 1,000 Hz. Additivity was assumed, on the basis of the postulate that acoustic stimuli that activate separate populations of auditory receptors will produce component loudnesses that in turn would combine by simple linear summation. Fletcher and Munson identified two conditions for independent activation and, hence, for presumed linear addition of loudness: stimulation of the two ears vs. one (binaural vs. monaural stimulation) and stimulation of the same ear with acoustic stimuli containing two (or more) widely separated tones vs. a stimulus containing a single tone.

Fletcher and Munson's procedure for measuring loudness contained, therefore, two steps: One starts by matching the loudness of a 1,000-Hz tone to the loudness of every acoustic stimulus of interest – to individual tones or tone complexes, presented to one or both ears. For every possible test stimulus, therefore, one determines the SPL of a matching 1,000-Hz tone – that is, the loudness level in phons. Subsequently, one may construct a scale of loudness by comparing, for example, the level in phons of a given sound presented binaurally and monaurally. Given the assumption of additivity, the sound will be twice as loud when heard by two ears

compared to one. Similarly, the loudness of two equally loud tones, spaced sufficiently in frequency, will be twice as loud when played together as either tone alone. If, for example, an acoustic signal has a loudness level of 70 phons when heard binaurally but 60 phons when heard monaurally, then the increase in SPL from 60 to 70 dB at 1,000 Hz constitutes a doubling of loudness.

Although Fletcher and Munson were able to perform a limited number of empirical tests of the adequacy of the principle of additivity, this critical principle remained largely an assumption of the system. Methods such as magnitude estimation, discussed below, can be used to ask, for example, whether subjects judge binaural sounds to be twice as loud as monaural sounds; the results can depend, however, on the ways that subjects make numerical judgments (see Algorn and Marks 1984). Methods of conjoint measurement (Luce and Tukey 1964) and functional measurement (Anderson 1970, 1981) provide additional mathematical and statistical tools for assessing additivity (for reviews, see Marks and Algorn 1998; Marks and Gescheider 2002). Results using these approaches have produced both some support for additivity (e.g., Levelt et al. 1972; Marks 1978), at least with narrow-band stimuli (Marks 1980), but also evidence against it (e.g., Gigerenzer and Strube 1983; Hübner and Ellermeier 1993). There is now considerable evidence indicating that a sound heard by two ears can be less than twice as loud as a sound heard by one (see Chaps. 7 and 8). Most pertinently here, however, as discussed in Sect. 2.2.3, Fletcher and Munson's loudness scale, based on the principle of additivity, is close to the scale that Stevens (1955, 1956) would later propose.

2.2.2.4 Measuring Loudness by Judging Ratios: The Original Sone Scale

Several contemporaries of Fletcher and Munson sought to measure loudness by instructing their subjects to make quantitative (numerical) assessments of relative values of loudness – an approach that aimed at ensuring that the measures of loudness would agree better than decibels-above-threshold with direct experience. In 1930, Richardson and Ross reported the results of a pioneering study in which they asked eleven subjects to estimate numerically the loudness values of tones that varied in both frequency and level, all of the loudness judgments being made relative to a standard tone assigned the value of 1.0. This method is essentially a version of magnitude estimation, which Stevens (1955) would reinvent and elaborate nearly three decades later.

Richardson and Ross's study marked the beginning of a spate of experiments on loudness scaling. Many of these experiments used what came to be called "ratio methods" (Stevens 1958b), in that the subjects were instructed, in one way or another, to assess the ratio or proportionality between the loudness of one sound and another, or to produce sounds that fall in a specified loudness ratio. One ratio method often used in the 1930s was fractionation. In fractionation, subjects are instructed to adjust the level of one tone to make its loudness appear one-half, or some other fraction, of the loudness of a standard tone (e.g., Ham and Parkinson 1932; Laird et al. 1932; Geiger and Firestone 1933).

By 1936, Stevens was able to pull together several sets of findings and use them to construct a scale of loudness that he called the sone scale. Richardson and Ross had inferred from their measurements that, on average, loudness increased as a simple power function of the stimulus – with an exponent of 0.44. Like Fletcher and Munson’s scale, the 1936 sone scale resembles the loudness scale that Stevens would later propose.

2.2.3 *Sone Scale of Loudness and Stevens’s Law*

Two decades later, Stevens (1955, 1956) proposed a revision of the sone scale, which, like Richardson and Ross’s loudness scale, follows a power function. According to Stevens, power functions characterize the general relationship between perceptual magnitudes and stimulus intensities, a relationship that applies to audition and to most, if not all, sensory modalities. Although Stevens mustered evidence in favor of a general power law, often designated as Stevens’s law, a lion’s share of his effort went to the measurement of loudness, and to the establishment of the new sone scale and its relation to the sound pressure or energy of the stimulus. In Stevens’s formulation, loudness in sones, LS , follows a power function of the form

$$LS = I^\beta \quad (2.9)$$

where the unit of measurement of I equals the sound pressure or energy of a 1,000-Hz tone at 40-dB SPL and the tone is presented simultaneously to both ears.

The exponent of the power function describing the new sone scale is 0.6 re: sound pressure (0.3 re: sound pressure or power), which is about one third larger than the value reported by Richardson and Ross – and in its overall form, the new sone scale broadly resembles both the earlier sone scale and the scale of Fletcher and Munson, despite the departure of both of the latter scales from a simple power-law representation.

Figure 2.3 plots Stevens’s (1955, 1956) new sone scale, which has served as the modern scale of loudness until fairly recently, together with his 1936 sone scale and with Fletcher and Munson’s (1933) loudness scale. Stevens inferred that the original sone scale of 1936 departed from a power function largely because of biases inherent in the method of fractionation, the method used to generate much of the data that contributed to the scale (for recent critiques, see Ellermeier and Faulhammer 2000; Zimmer 2005). Lacking independent evidence regarding which methods are biased, how they are biased, and to what extent they are biased, it is also possible that the “true” loudness function at 1 kHz actually falls closer to the original sone scale than to the revised scale, that the departures from a power function evident in the original sone scale accurately represent loudness. Indeed, by 1972, Stevens would acknowledge the possibility of systematic deviations of loudness from a power function, a notion confirmed by subsequent findings of Florentine et al. (1996) and Buus et al. (1997), who came to this conclusion using a different conceptual framework (for

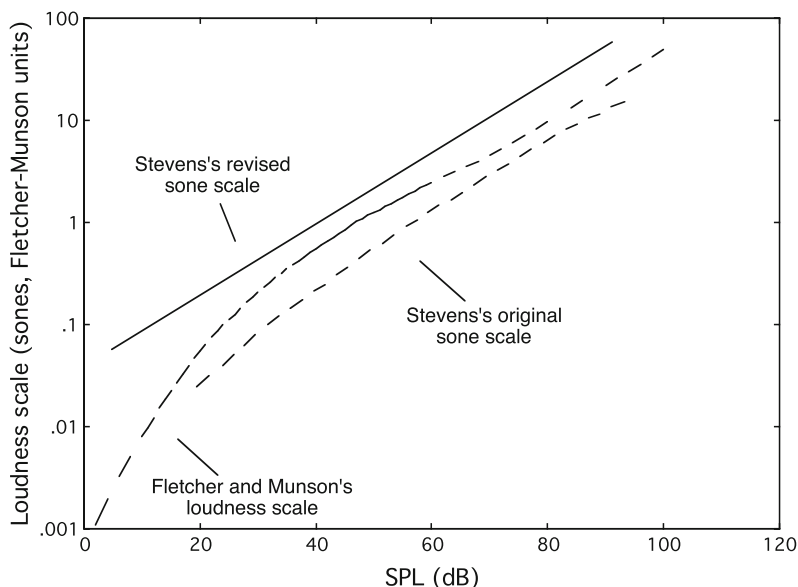


Fig. 2.3 Fletcher and Munson's (1933) loudness scale, Stevens's (1936) original sone scale, and Stevens's (1956) subsequent revision of the sone scale. All three scales are plotted on logarithmic axes, the decibel scale being itself logarithmic. The modern sone scale is defined explicit by a power function (*straight line* in these axes), whereas Fletcher and Munson's scale and the original sone scale only approximate power functions. Note that for clarity of display, Stevens's original sone scale is displaced downward by multiplying the values in sones by one-third

review, see Buus and Florentine 2001). Evidence that the log-log slope (exponent) of the loudness function is smaller at moderate SPLs, 25–60 dB, than at lower or higher ones, suggests the need to modify Stevens's simple power function with a more complex function. Such a function has been proposed by Florentine et al. (1996) and Buus et al. (1997) and termed the inflected exponential (InEx) function (see Florentine and Epstein 2006 and Chap. 5).

Note that Stevens (1956) derived the new sone scale largely on the basis of data obtained with magnitude estimation (the method used by Richardson and Ross 1930), as well as with data obtained using magnitude production, a method that inverts magnitude estimation. In magnitude estimation, the experimenter presents a series of sounds and the subject assigns numbers in proportion to the loudness of each; in magnitude production, the experimenter presents a series of numbers and the subject's task is to adjust the loudness of each to match. To revise the sone scale, Stevens included data obtained with both estimation and production methods.

This revised sone scale maintained the definition of 1 sone as the loudness of a binaurally heard tone at 40-dB SPL (see Chap. 5). The revised sone scale is a simple power function, and it was subsequently accepted by the ISO as the standard for the measurement of loudness (ISO 1959). Over the past half century, the sone scale has served as a touchstone for the measurement of loudness, as other approaches have

been developed and investigated. This work has been critical in pointing to the ways that different psychophysical methods can give different results, and to the problems and potential pitfalls associated with the application of different psychophysical methods to measuring the magnitudes of sensations, including loudness.

2.3 Contemporary Approaches to Measuring Loudness

A number of methods are currently used to assess how loudness depends on various stimulus parameters. Modern approaches to the measurement of loudness rely primarily on several kinds of ratings or estimations of loudness, using variants of methods used by, for example, Richardson and Ross (1930) and Gage (1934). These have been reviewed in the previous sections. Each method has strengths and limitations; there is no perfect method for measuring loudness. Loudness researchers need to choose the best measurement method from what is available, while keeping in mind its limitations. The purpose of this section is to summarize issues of relevance when choosing a method of measurement. There are two broad types of measurement methods that are currently used: equal loudness matching and scaling methods. Each of these will be described in turn.

Whatever method is chosen to measure loudness, it must meet the basic requirement of yielding internally consistent measurements (see, e.g., Marks 1974b). A test of internal consistency can be defined in terms of loudness matches or comparisons (cf., Buus 2002). Acceptable methods for measuring loudness provide data conforming to two principles. The first is an ordinal indicant of relative loudness. If sound A has a measured loudness greater than that of sound B, then sound A is louder than sound B, and sound B is softer than sound A. Further, whenever two (or more) sounds are equally loud, the system must assign to them the same value in loudness. The second principle is that loudness equalities must be transitive: If acoustic signal A1 is as loud as signal A2, and A2 is as loud as A3, then A1 must be as loud as A3. The topic of internal consistency of loudness measurements will be revisited at various points in this section as it pertains to specific methods.

Before discussing specific methods, a word of caution is in order regarding their classification. Some authors have designated modern approaches as “direct” or “indirect.” This has led to some confusion, because all methods for measuring sensory magnitudes are indirect, although it is fair to say that some are more indirect than others. The term “direct” has been used to denote approaches in which subjects are instructed to judge or rate loudness itself, often on a scale that has putative quantitative or quasi-quantitative properties. The designation of several approaches as “direct” is also intended to contrast with “indirect” approaches, such as that of Fechner, who sought to infer sensation magnitudes from measures of discrimination. Nevertheless, use of the adjective “direct” in this way remains something of a misnomer. The process for measurement involves not only the task that is set forth to the subject – for instance, to rate loudness on a discrete, bounded scale containing a fixed number of categories, or on a continuous, open-ended magnitude-estimation scale – but also

involves a set of explicit or implicit mathematical assumptions that the experimenter makes so as to infer quantitative measures of loudness from the rating responses. To prevent a potential source of confusion, the practice of labeling methods as “direct” and “indirect” should be avoided.

2.3.1 Equal Loudness Matching

Equal loudness matching has been used extensively to assess how loudness depends on various stimulus parameters. It uses listeners as null-detectors to obtain measurements of stimulus parameters leading to the point of subjective equality (i.e., the level at which one sound is as loud as the other). Equal loudness matching needs only to assume that listeners can judge identity along a particular dimension, such as loudness, while ignoring differences along other dimensions (e.g., pitch, timbre, apparent duration, etc.). This axiom has never been seriously questioned (Zwislocki 1965; Chap. 1) and there is a general consensus among psychoacousticians that equal-loudness measurements continue to be the “gold” standard to which results obtained by other methods must conform. Loudness-matching (loudness-balance) measurements do not provide direct information about how loud a particular stimulus sounds. They provide information only about the level of a comparison sound judged as loud as the stimulus under investigation. Of course, if the loudness function for the comparison is known, the loudness function for the test stimulus can be constructed.

The measure known as “loudness level” was developed to construct a system in which loudness could be set equal to a common currency: in terms of the SPL of a 1-kHz tone whose loudness matches the loudness of any given test tone. The unit of loudness level is a phon, so that the loudness level of N phons is as loud as a 1-kHz tone at N -dB SPL [see Chap. 5, or the international standard (ISO 226, 2003)].

In several respects, loudness level in phons serves as a useful tool for assessing loudness: The specification of loudness level in decibel (phons) provides both a nominal indicant of loudness – all acoustical signals that are equal in loudness are, by definition, equal in loudness level – and also an ordinal indicant of relative loudness described earlier. The contention that all acoustical signals that have the same loudness should have the same loudness level points to a basic constraint on any method for measuring loudness. Whenever two (or more) sounds are equally loud, the system must assign to them the same value in loudness.

Loudness-balance measurements almost always determine the sound levels at which a test stimulus and a comparison stimulus appear equally loud. These measurements usually require that the level of one stimulus (the comparison) be varied in some manner to ascertain the level at which it is as loud as another stimulus (the standard). The variation in stimulus level can be accomplished in several ways, depending on the psychophysical procedure used to measure the point of subjective equality. The most frequently used psychophysical procedures are the method of adjustment and the modern adaptive procedures, which are described in the following

sections. The method of constant stimuli, often used to measure loudness in classic research (e.g., Fletcher and Munson 1933), is highly inefficient and has been replaced by modern adaptive psychophysical procedures. For a description of the method of constant stimuli and other psychophysical procedures, see Gescheider (1997), Gulick et al. (1989), or Gelfand (2004).

2.3.1.1 Measuring Equal Loudness with the Method of Adjustment

In the method of adjustment, a listener is presented two sounds that alternate in time and is given direct control of the level of one of the sounds. The listener is instructed to adjust the variable sound to be equal in loudness to the sound that is fixed in level. Usually, the listener is asked to use a bracketing procedure, that is, to adjust the variable stimulus alternately louder and softer than the fixed stimulus so as to “home in” on the point of equality. One measurement of the point of subjective equality is taken to be the level produced by the final setting of the attenuator.

Although this procedure is conceptually simple, systematic errors may distort the results unless they are minimized through careful experimental design. For example, listeners tend to judge the second of two successive identical sounds as louder or softer than the first, depending on the interstimulus interval between the two (Stevens 1955; Hellström 1979). These time–order errors can be minimized if the order of presentation of the fixed and variable stimuli is randomized. More importantly, listeners tend to overestimate the loudness of the fixed stimulus. An additional bias of the adjustments toward comfortable listening levels may reinforce the overestimation for measurements at low levels, but reduce it at high levels (Stevens 1955). Thus, listeners will tend to set the variable stimulus too high in level in measurements at low and moderate levels, whereas this bias often appears small at high levels (e.g., Zwicker et al. 1957; Zwicker 1958; Scharf 1959, 1961; Hellman and Zwislocki 1964). These adjustment biases may also depend on the mechanical and electrical characteristics of the device used to control the variable stimulus (Guilford 1954; Stevens and Poulton 1956). Averaging the results by having the listeners adjust both the test stimulus to the comparison and the comparison to the test stimulus may minimize the effect of these adjustment biases. Because markings and steps on the adjusted attenuator may produce intractable biases in the adjustments, the variable stimulus should be controlled via an unmarked, continuously variable attenuator.

2.3.1.2 Measuring Equal Loudness with Adaptive Methods

The widespread availability of computers to control psychoacoustic experiments has led many investigators to use adaptive procedures for loudness-balance measurements (e.g., Jesteadt 1980; Hall 1981; Silva and Florentine 2006; for an introduction to adaptive procedures, see Gelfand 2004). In these procedures, the listener is presented two stimuli in sequence with a pause between them and is asked to respond which

of the two is louder. The listener's response determines the presentation level of the variable stimulus on the next trial, according to rules that generally make the variable level approach, from both above and below, the level required for equal loudness. In many of the procedures, the critical values are the reversal points – stimulus levels at which the response to the variable changes from “softer” to “louder” or from “louder” to “softer.” The complexity of the rules varies from a simple up–down procedure (e.g., Levitt 1971; Jesteadt 1980; Florentine et al. 1996) to complex procedures based on maximum-likelihood estimates of the psychometric function (e.g., Hall 1981; Takeshima et al. 2001).

Although a number of adaptive procedures have been used to measure absolute threshold (e.g., see Leek 2001), the simple up–down procedure is without doubt the most frequently used adaptive procedure to measure equal loudness. The amount of change in the level of the variable stimulus on each trial is determined by the experimenter and is often reduced as the point of subjective equality is approached. For example, a 5- or 6-dB step size may be used until the second reversal in direction of the level, with a 2-dB step size used thereafter (e.g., Zeng and Turner 1991; Buus and Florentine 2002). The entire series of trials over which the signal level varies according to a single adaptive algorithm is called an “adaptive track” and it results in a single measurement.

The stopping rules for adaptive tracks vary among laboratories and are usually based on a predetermined number of reversals. In general, there is a trade-off between the number of trials and the variability in the data: the more measurements, the less variability. On simple statistical grounds, the standard error of the mean across repeated measurements should be inversely proportional to the square root of the number of observations. But requiring subjects to make large numbers of tedious judgments may produce fatigue, which in turn is likely to increase variability over time. For this reason, it is essential that the psychophysical procedure be efficient and that subjects take breaks from listening to prevent fatigue, especially in long experiments.

Care must be taken to eliminate sources of bias in adaptive procedures that may distort judgments. In addition to the time-order errors mentioned earlier, adjustment biases might affect results obtained with adaptive procedures. Although the control over stimulus levels in adaptive procedures is indirect, the listener may nevertheless become aware of which stimulus is varied and attempt to “adjust” the level by responding in particular ways – for instance, by either perseverating or changing responses. Moreover, responses may be affected if the listener compares the perception of the current stimulus to the memory of stimuli on previous trials. Some of these biases can be minimized by randomizing the order of the test stimulus and comparison on every trial and by interleaving multiple adaptive tracks in which the test stimulus and the comparison are varied (Buus et al. 1998; for a general discussion of possible biases and the use of interleaved tracks, see Cornsweet 1962). Using concurrent tracks with the fixed-level stimulus presented at different levels creates additional, apparently random, variation in overall loudness, which forces the listeners to base their responses only on the loudness judgments presented in a trial. However, caution should be used when roving the stimulus level due to context effects, such as induced loudness reduction, described by Arieih and Marks in Chap. 3.

Comparisons across studies using adaptive procedures show large variability in the resulting equal-loudness matches. In part, the variability is likely due to individual differences; it appears also to result from characteristics of the measurement procedures themselves. In many experiments, the goal is to take measurements at several different levels of intensity (or some other parameter of the acoustic stimulus). Different experimenters may use different experimental designs to determine the sequence of the presentation levels, and the sequence may affect the results. Experimenters may opt to vary stimulus intensity in several ways: increasing level across blocks of trials (Ascending Across Blocks [(AAB)]), decreasing level across blocks of trials (Descending Across Blocks [(DAB)]), randomizing level across blocks of trials (Random Across Blocks [(RAB)]), or randomizing level within blocks (Random Within Blocks [(RWB)]). Most contemporary studies use an RAB paradigm, but all four of the aforementioned designs have been used in one investigation or another. Unfortunately, some studies failed to report the stimulus sequence (for review, see Silva and Florentine 2006).

Researchers have long known that “measurement bias” can affect equal-loudness matching data. For example, Stevens and Greenbaum (1966) found that when listeners adjust the level of stimulus B to match several fixed intensity levels of A, and also adjust A to match several levels of B, the results commonly show a so-called “regression effect”: The slope of the function plotting adjusted B against A is flatter than the slope of the function plotting B against adjusted A. This might occur due to the preferences that subjects have for listening to sounds at a comfortable loudness. The implication of the regression effect is that the loudness of the variable sound is “over-estimated” near threshold and “under-estimated” at high levels. Regression-type biases in comparison and matching are ubiquitous. Florentine et al. (1996, 1998) observed a regression effect in an adaptive two-interval, two-alternative forced-choice RAB procedure, originally developed by Jesteadt (1980), when they measured the loudness of two stimuli having different durations. An example of this regression effect is shown in Fig. 2.4.

To examine how different stimulus sequences affect loudness matches measured in an adaptive procedure, Silva and Florentine (2006) compared four different sequences in a study of temporal integration. Specifically, they obtained loudness matches between 1-kHz tones having two durations (5 and 200 ms) in each of six listeners, asking whether different sequences of stimuli might affect the magnitude of temporal integration. Three of the sequences varied the level of the fixed tone either sequentially (AAB, DAB) or randomly (RAB) across blocks of trials. The fourth sequence (RWB) randomized the level within blocks. As shown in Fig. 2.5, when the short-duration tone was fixed, there was a significant difference between the magnitude of temporal integration obtained using the RWB procedure vs. the other three procedures, at moderate levels (50–60-dB SL). When comparing loudness matches obtained over a wide range of levels in different experimental studies, therefore, it is important to consider the sequence of stimulus levels presented within each study.

Methods of measuring equal loudness vary among research laboratories and some of the methods have not been fully evaluated with regard to internal consistency.

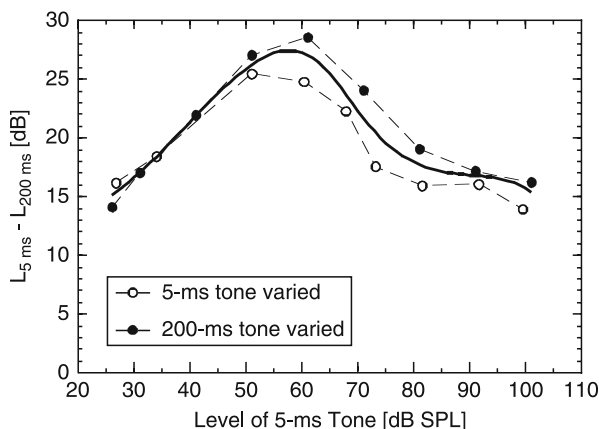


Fig. 2.4 Amount of temporal integration as a function of level. Data of Florentine et al. (1998) for the level difference between equally loud 5- and 200-ms tones at 1 kHz are plotted as a function of the SPL of the 5-ms tone. The filled points show data obtained by a simple up-down method when the 200-ms tones were varied; the unfilled points show the data obtained when the 5-ms tones were varied. Differences between the filled and unfilled points reflect judgment biases, which cause the level of the variable tone to migrate toward a comfortable loudness. The solid line shows the difference in level obtained between the 5- and 200-ms loudness functions (the figure, from Buus (2002, Fig. 19), is reproduced with permission. It was published in Tranebjærg L, Christensen-Dalsgaard J, Andersen T, Poulsen T (eds): *Genetics and the Function of the Auditory System*. Proceedings of the 19th Danavox symposium, Kolding, Denmark. Danavox Jubilee Foundation, ISBN 87-982422-9-6, Copenhagen, 2001)

When a question exists regarding the viability of a particular method, whenever feasible, it is wise to include a check of internal consistency in the experimental design. To be sure, additional testing of consistency can be laborious and time-consuming, but in many circumstances it is critical to ensure that one understands how methodological decisions may affect the results, and therefore the conclusions drawn from them. An example of an experimental design containing a test of consistency can be found in Florentine et al. (1978).

2.3.2 Loudness Scaling

Measures such as loudness level serve as a kind of intervening variable, to use the terminology of MacCorquodale and Meehl (1948). Loudness level captures information about a perceptual attribute, indicating, for instance, that any sound having a specific loudness level has the same loudness as any other sound of a specific loudness level. Loudness level also tells us about rank order of loudness, in that loudness level increases as loudness increases. Loudness level indicates nothing more. As an intervening variable, loudness level specifies loudness equivalence – and, by the

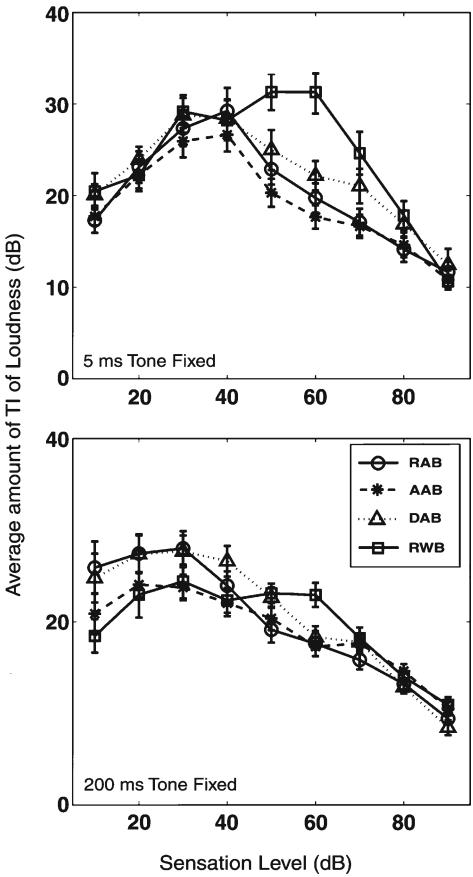


Fig. 2.5 Average difference in level required for equal loudness between 5- and 200-ms tones, plotted as a function of sensation level across six listeners for four adaptive procedures. The random across blocks (RAB), ascending across blocks (AAB), and descending across blocks (DAB) procedures varied the level of the fixed tone in a random, increasing, and decreasing order, respectively, across blocks of trials. The random within blocks (RWB) procedure presented only two blocks of trials, where the level of the fixed tone varied randomly across a range of levels. The error bars represent plus and minus one standard error (the figure is a reproduction of Fig. 4 from: Silva and Florentine 2006)

addition of an empirical measure of order, about rank order – but it does not specify to what extent the loudness of one sound exceeds that of another. This is insufficient; we want to know how loudness itself depends on loudness level.

A set of quantitative measures of loudness, per se, can serve as a hypothetical construct – another term used by MacCorquodale and Meehl (1948), this one referring to unobserved variables that go beyond summarizing empirical relations but presumably contain additional information. Quantitative scales of loudness, obtained by methods such as category rating or magnitude estimation, contain all of the information that is available in intervening variables such as loudness level,

and more. These scales, when they are free of biases, inform not only about equivalences and rank orders, but also have the potential to inform about quantitative relations, such as how loud one sound is relative to others.

It is convenient to characterize loudness rating scales in terms of two main attributes that help distinguish them: whether the scale is bounded or unbounded, and whether the scale is discrete or continuous. On one end of the spectrum are traditional loudness rating scales (i.e., categorical loudness scales), which are both bounded and discrete. Within a fixed range (bounds), these scales provide the listener with a relatively small number of categorical labels (discrete), such as the integers from 1 through 9, or descriptive labels from “extremely soft” to “extremely loud.” On the other end of the spectrum are magnitude-estimation scales (described more fully in the next section), which are unbounded and continuous. In principle, responses on a magnitude-estimation scale may be infinitesimally small or infinitely large.

Hybrid scales are also possible, a common modern adaptation being the so-called visual analog scales. These hybrids are bounded, continuous scales that are often presented as line segments and labeled numerically or adjectivally at their ends (and sometimes at various points between). Hybrid scales allow essentially continuous response along the line segment with continuity being limited only by the precision of the response or its measurement. Visual analog scales are attractive because people generally find it easy to use spatial length or position as a “metaphor” for perceived strength (see Lakoff and Johnson 1980); children as young as 3–4 years readily make graded responses on visual analog scales (e.g., Anderson and Cuneo 1978; Cuneo 1982; Marks et al. 1987). As we shall see, visual analog scales avoid some of the pitfalls of many discrete rating scales, especially those associated with the use of small numbers of discrete labels or categories.

2.3.2.1 Category Scales

The use of categorical rating scales has a long history, going back to the nineteenth century. These scales have been deployed to study not only loudness (and other sensory responses), but just about anything that people are able to judge. Dawes (1972), for example, noted that in the year 1970, about 60% of all of the experimental articles published in the *Journal of Personality and Social Psychology* reported measures made on these scales. Category Loudness Scales (CLS) are ubiquitous in clinical settings for fitting hearing aids and elsewhere, because they are easy to administer and measurements can be obtained quickly (for clinical usages, see Chap. 9). In this procedure, a listener is presented a series of sounds. After each sound, the listener assigns one of a number of possible categories to its loudness.

To ensure proper use and interpretation of CLS, it is important to understand their limitations and the assumptions that underlie them. It is often assumed that each successive number or adjective on a discrete rating scale marks off, or should mark off, a uniform difference or interval in the quantity being measured. For example, in rating loudness on a nine-point scale, the categories might be the integers from 1 through 9, or they might be nine descriptive labels, such as “extremely soft,”

“very soft,” “soft,” “somewhat soft,” “medium,” “somewhat loud,” “loud,” “very loud,” and “extremely loud.” Either way, it is commonly assumed that the step from 1 and 2 on the numerical scale, or from “extremely soft” to “very soft” on the adjectival scale, represents the same difference in loudness as the step from 3 to 4 or from 8 to 9, or from “soft” to “somewhat soft” or from “very loud” to “extremely loud.” Thus, for computational purposes, successive categories on an adjectival scale are commonly assigned successive integers. The assumption of uniformity implies that the resulting measurements are made on an interval scale.

It is often assumed further that adjectival and numerical labels provide similar quantitative information. Judgments made on rating scales are often relativistic, as subjects tend to use all of the categories equally often, commonly assigning the lowest category to the weakest stimulus presented and the highest category to the strongest (see Chap. 3). The lowest and highest stimulus levels, in turn, may serve as anchors, so responses to the lowest and highest levels often show much less variability than responses to stimulus levels in between – an example of what has been called the “edge resolution effect” (e.g., Berliner and Durlach 1973; Berliner et al. 1977). The tendency toward relativistic judgment and the presence of edge effects have important consequences for any attempt to compare directly category ratings made by different groups of subjects, for example, subjects with normal hearing and subjects with hearing loss. A person who cannot hear very soft sounds (e.g., with “softness imperception” caused by a hearing loss) may label a sound close to threshold as “very soft,” not because it is perceived with the same loudness as a person with normal hearing, but because it is the softest sound the person is capable of perceiving.

An explicit method for using descriptive categories in the measurement of loudness was proposed by Heller (1985). In Heller’s scheme, the measurement procedure involves two phases: First, in response to a test sound, the listener selects from five broad descriptive categories that cover the range of possible loudness from very soft to very loud. And second, in response to a repetition of the test sound, the listener then selects from ten levels within the initial category. Thus, the overall scale contains 50 possible response categories in all – a sufficiently large number to avoid the biases inherent in the use of small numbers of response alternatives. As a shortcut, one can combine the two steps into one, presenting the subject with all 50 alternatives. Other modifications have also been offered, such as the adaptive method of Brand and Hohmann (2002). A new ISO standard, 16832 (ISO 2006), proposes conditions to help ensure reliability in the use of categorical methods to study loudness. These methods may be useful in applied research (e.g., audiology or environmental noise).

There are different types of CLS and a number of factors to consider when choosing a CLS for a given task, such as stimulus spacing and the number of response categories. Stevens and Galanter’s (1957) classic study compared the ratings of loudness on category scales to the corresponding ratings of the same stimuli on unbounded magnitude-estimation scales, and the critical finding – discussed later, in the review of unbounded scales – was the nonlinear relation between judgments on the two scales. For present purposes, however, it is sufficient to note that Stevens and Galanter obtained ratings in several experiments in which the authors

compared, albeit somewhat unsystematically, (1) numerical and adjectival labels, (2) different numbers of available response categories, and (3) different sized steps between successive stimulus levels. Presenting numbers vs. adjectives did not have a major effect on the ratings for their normal-hearing subjects, but variations in the number of available response categories and variations in stimulus spacing had more substantial effects. The greater the number of response categories available to the subject, for example, and hence the more nearly continuous the rating scale, the more linear the relation between the resulting responses on the rating scale and responses on a fully continuous magnitude-estimation scale. Heller's (1985) scheme, described in the preceding text, capitalizes on the availability of a relatively large number of response categories.

Research following the study of Stevens and Galanter (1957) has shown systematic effects of both number of categories and spacing of stimulus levels on rating-scale responses, although these studies have largely investigated sensory dimensions other than loudness. Marks (1968) reported results of a systematic study of the perception of brightness of flashes of light, examining the effects of both stimulus spacing (size of the log intensity step between successive stimuli) and number of available numerical categories. The ratings could be described by a power function of stimulus intensity, having the form

$$C - C_0 = cS^a \quad (2.10)$$

where C is the average rating on the category scale and a is the fitted exponent (the additive constant C_0 is necessary in order to adjust for the scale's arbitrary zero-point). The value of the exponent a increased with increasing stimulus spacing and with increasing number of available responses. Results of Stevens and Galanter (1957) suggest that category ratings of loudness would behave similarly.

Given the historic popularity of seven-point category scales, it may be tempting to assume that a relatively small number of categories is sufficient. Incorrect though it is, this temptation may be increased by the evidence, famously reviewed by Miller (1956) that the channel capacity for absolute identification of stimuli on a univariate continuum, such as loudness, is roughly seven items. It is important to keep in mind that the "magical number seven," as Miller dubbed it, does not mean that performance ceases to improve by presenting more than seven stimuli and seven response categories. The channel capacity of seven refers to the level of asymptotic performance when the number of stimuli – and the number of possible responses – is considerably larger than seven. Reanalysis (Marks 1996) of the category ratings reported by Marks (1968) suggests that increasing the number of response categories from 4–20 to 100 increases the amount of information transmitted, a measure of the mutual discriminability among the stimuli (see also Garner 1960). When bounded scales are appropriate or desirable, therefore, it is crucial that the scale contain a sufficient number of categories, on the order of 15 or more. An alternative to a discrete category scale is a visual analog scale, a line scale that permits the subject virtually continuous response between the end points (e.g., Anderson, 1981).

Category rating scales appear to be especially sensitive to the selection of the stimulus and response alternatives. Half a century ago, Stevens (1958a) examined the role of stimulus spacing in categorical judgments of loudness. In particular, Stevens compared uniform decibel spacing to two kinds of nonuniform spacing: spacing with stimuli bunched at the lower end and spacing with stimuli bunched at the upper end. Spacing exerted a substantial effect on judgment: Over the region in which stimulus levels were bunched together, subjects tended to spread out their responses to a greater extent than they did over comparable ranges where stimuli were spaced more sparsely. Because the scale is bounded, if a subject “uses up” more categories where spacing is narrow, the subject must perforce change the relation of the ratings to the stimuli in the remaining region of the stimulus range, where spacing is broader. With unbounded scales, however, this constraint could disappear, or at least diminish. Not surprisingly, in the same study, Stevens found that stimulus spacing exerted a much smaller effect on magnitude estimations of loudness than it did on category ratings.

Effects similar to those of stimulus spacing, just mentioned, arise when the ensemble of stimulus levels remains constant, but the frequency of presentation of the various stimuli changes across conditions. When fixing the stimulus levels, one might present the lower level twice as often as the higher ones, or the higher ones twice as often as the lower ones. From the perspective of the subject, the effect is much like bunching stimulus levels at the low and high ends, respectively, and the resulting patterns of response are similar. From results of this sort, Parducci (1965, 1974) developed a range-frequency model of categorical judgment, using as one of his primary principles the notion that, with discrete scales, subjects tend to use all of the available responses equally often. This tendency underlies the effects of stimulus spacing and frequency of presentation. Given this tendency, the “ideal” or “unbiased” function would be one that spaced the stimuli uniformly with regard to loudness, so that successive categories marked off uniform changes in loudness. Pollack (1965a, b) has shown how one can use an iterative experimental method to reduce equal-response tendencies – capitalizing on evidence that these tendencies do not wholly determine the results. In Pollack’s method, one uses the results of an initial experiment to adjust the spacing so as to try to increase the uniformity of the subjects’ responses in a subsequent experiment, and the procedure is repeated until a uniform scale is achieved.

One concern about using categorical and other rating scales is that they may not provide adequate measures of internal consistency. Relatively few studies have addressed the question whether rating scales provide results consistent with the principle of equality. One pair of studies did address the question, albeit indirectly, by examining binaural summation of loudness using several methods, including loudness matching, loudness scaling on a visual analog scale, and magnitude estimation, which is described in the next section (Marks 1978, 1979). All three methods gave comparable measures of summation, quantified in terms of matching the loudness of monaural and binaural sounds. It is possible, however, that some rating scales may fail this “test of internal consistency.” Support for this contention comes from two small studies in vision that obtained category ratings (on 9-point

and 11-point scales, respectively) in order to determine how brightness depends on both the duration and the luminance of flashes of light (Raab et al. 1961; Lewis 1965). Results of both studies failed to show the presence of peaks in brightness as a function of duration (Broca-Sulzer effect); these peaks are readily shown with other methods, including direct matching (Aiba and Stevens 1964) and magnitude estimation (Raab 1962; Stevens and Hall 1966). Lewis's study also failed to show the level-dependent change in critical duration for integration, another visual phenomenon revealed in both matches and magnitude estimations. These findings suggest that category scales, especially ones using relatively small numbers of possible responses, may give results incompatible with the principle of equality.

Another potential problem with using CLS is that rating scale responses are typically averaged and treated as if they provide reasonably uniform (interval scale) measures of the underlying perceptual representations. Yet, as discussed in detail by Arieh and Marks in Chap. 3, the pervasive effects of stimulus spacing, presentation frequency, and number of available responses suggest that decisional processes play a substantial role in determining categorical judgments, hence in determining the relation of mean category judgment to stimulus level.

2.3.2.2 Magnitude Estimation Scales of Loudness

Magnitude estimation is a type of unbounded, continuous scaling procedure. In the method of magnitude estimation, a listener is presented a series of stimulus levels in random order. After each stimulus presentation the listener is asked to respond with a number that matches its loudness. Any positive number that seems appropriate to the listener may be used. Stevens (1956, 1975), Hellman (1991), and others (e.g., Zwislocki 1983) have argued that this type of unbounded response scale is most effective in producing responses that are approximately proportional to loudness.

Magnitude estimation comes in several varieties. In Stevens's earliest version of the method, subjects were presented at the start of a session with a standard stimulus of fixed sound level, together with a numerical modulus assigned to represent each stimulus. The standard typically came from somewhere in the middle range of levels, and the modulus commonly had a value of "10," a numeral deemed neither "too large" nor "too small." Subjects were instructed to assign numbers to the loudness of other sounds in proportion – that is, to maintain the appropriate ratio between numbers and sounds. If another sound was twice as loud as the standard, it should receive a response of "20." If it was one fifth as loud, it should receive a response of "2." Some investigators may omit the standard stimulus, but continue to emphasize the relative, ratio relations of responses, by explicitly asking subjects to judge the loudness of the current stimulus in terms of the loudness of the previous stimulus; if the previous stimulus was assigned the numeral "5" and the current stimulus appears three times as loud, the subject should assign it the numeral "15" (e.g., Luce and Green 1974). Luce and Green dubbed this method "ratio magnitude estimation." Ratio magnitude estimation is likely to enhance sequential (contextual) effects, that is, the way that stimuli and responses on trial n affect responses on trial $n + 1$ (for a

more thorough discussion of sequential effects, see Arieih and Marks, Chap. 3). Although useful for studying decisional processes, ratio magnitude estimation is probably not a method of choice when the goal is to measure loudness in ways that minimize such sequential effects.

Eventually, Stevens (e.g., 1956) abandoned the use of both a standard stimulus and numerical modulus. Following many earlier studies on the topic, Hellman and Zwischlocki (1961) found that the values of both standard and modulus affected the numerical responses, and, in particular, affected the observed exponent of the power function relating judgments of loudness to sound intensity. Most notably, the exponent remained constant if both standard and modulus increased or decreased in tandem, but not if either standard or modulus changed while the other remained constant. This pattern of results suggested the possible existence of a “natural” connection between sensation magnitude and numerical response, and hence the possibility that the experimenter’s arbitrary choice of standard and modulus may induce biases in responses (Hellman and Zwischlocki 1963, 1964; Hellman 1991). This eventually led the way to the development of a method known as “absolute magnitude estimation,” in which instructions avoid any reference to ratio relations, but instead encourage subjects to assign numerals to stimuli such that the “perceived magnitude of the numbers match the perceived magnitudes of the sensations.” Subjects may be allowed to hear a stimulus as often as desired before rendering a judgment (Cross 1973; Hellman 1976).

An example of instructions using absolute magnitude estimation follows:

You are going to hear a series of sounds. Your task is to specify how loud each sound is by assigning numbers. Louder sounds should be assigned larger numbers. You are free to use any positive numbers that seem appropriate—whole numbers, decimals, or fractions. Do not worry about running out of numbers; there will always be a smaller number than the smallest you use and a larger number than the largest you use. If you do not hear a sound, please assign it zero, otherwise all numbers should be larger than zero. Do not worry about the number you assigned to previous sounds, simply try to match the appropriate number to each sound regardless of what number you may have assigned the pervious sound.

Although there appear to be conditions in which the method of absolute magnitude estimation encourages subjects to map their numerical responses to sensations in a way that, per the method’s label, is “absolute” (Zwischlocki and Goodman 1980; Zwischlocki 1983), absolute magnitude estimation shows at least some of the contextual effects in the judgment of loudness that are shown by other methods, such as category rating and ratio magnitude estimation (Ward 1987).

The range and spacing of the stimuli presented to the subject also influence magnitude estimates of loudness. Although Arieih and Marks (Chap. 3) discuss the role of contextual effects on loudness, it is important to consider here the role of stimulus range and stimulus distribution. Several investigations have shown that the form of the loudness function, and in particular the exponent of the power function, can vary systematically with the range of test levels: the larger the range, the smaller the exponent (Poulton 1968, 1989; Teghtsoonian 1973). Keep in mind, however, that the effect of stimulus range is typically fairly modest, appearing only when the range of levels becomes very small (smaller than about 20 dB). Over

larger ranges, the exponent is more or less independent of the range (Teghtsoonian 1971, 1973). This effect can be explained, at least in part, by the fact that the slope of the loudness function is shallower at moderate levels than at low and high levels.

Although stimulus range exerts a relatively small effect on power-function exponents, the effect is systematic, and the very presence of the range effect points to the importance of distinguishing between the underlying perception of loudness and the overt responses that listeners give to a particular set of stimuli, in a particular contextual setting, under a particular set of instructions. Overt responses, such as magnitude estimations, represent the end product of at least two sets of processes. The first is the set of sensory processes by which patterns of stimulus energy are transformed into internal representations of sounds, including their loudness. The second is the set of decisional and judgmental processes by which the internal representations of loudness map into the numerical responses (see Gescheider 1997; Marks and Algom 1998; Marks and Gescheider 2002). To explain the effect of stimulus range on the exponent of the loudness function, therefore, one would hypothesize an initial sensory, power-function transformation of sound pressure or energy to loudness, followed by a subsequent decisional, power-function transformation of loudness to numerical response. Only when the exponent of the decisional power function is 1.0 – that is, when the function is linear – would the numerical responses provide “valid” measures of loudness.

The modest size of the range effect contravenes the hypothesis (e.g., Poulton 1968, 1989) that exponents are simply accidental byproducts of the choice of stimuli presented by the experimenter, along with the predilections for particular numerical responses on the part of subjects. Were this so, then the subjects would presumably give the same range of numerical responses regardless of the stimuli presented. This does not occur. Instead, as stimulus range increases, so does the range of numerical responses, implying that stimulus range has only a modest effect on the exponent of the decisional power function (Teghtsoonian 1971). Nevertheless, to help circumvent effects of stimulus range, and other factors that influence decisional processes, one might choose to “calibrate” the subjects in advance of testing, by teaching them a particular stimulus-response function, as suggested by West et al. (2000; see also Marks et al. 1995).

One should be cautious, however, about making the implicit assumption that stimulus range affects only the decisional and judgmental processes that intervene between loudness and overt responses. Algom and Marks (1990) have provided some evidence that stimulus range may have two effects: As already discussed, changing the stimulus range can influence the decisional function relating numerical responses to the underlying values of loudness. But changing range may also affect the sensory function relating the underlying values of loudness to stimulus level. Algom and Marks drew this conclusion from the observation that stimulus range affected the implicit loudness matches between tones heard monaurally and binaurally.

Loudness functions can vary not only with the overall dynamic range of stimuli but also with their spacing and distribution. For example, if the sound levels

are spaced unevenly, with smaller steps between successive levels in one region of the overall range compared to others (or if a subset of levels is presented more frequently than others), the exponent of the power function will tend not to be uniform over the entire stimulus range, but instead will be greater over the local region in which the stimulus levels are bunched (Stevens 1958a).

2.3.2.3 Magnitude Production and Cross-Modality Matching

In some methods, the subject controls the stimulus and sets it to a target loudness, which can be specified in several ways. In magnitude production, the subject hears only the variable stimulus and is instructed on each trial to adjust its loudness to match the number assigned on that trial. If the perceived magnitude of numbers is considered a separate modality, then magnitude production and magnitude estimation become special cases of cross-modality matching (e.g., Stevens 1959; Reynolds and Stevens 1960; Hellman and Zwislöcki 1961, 1963; Hellman 1991).

Some investigators have suggested that the results of magnitude estimation and magnitude production be averaged in order to compensate for biases in each method (Hellman and Meiselman 1993). The combination of these two methods is sometimes called “numerical magnitude balance” (Hellman and Zwislöcki 1963; Hellman 1976). The recommendation to average results obtained by estimation and production assumes that the biases in the two methods are equal and opposite. In this regard, Hellman and Zwislöcki (1961) reported excellent agreement between results obtained by directly matching tones in the absence and presence of masking noise and results obtained by magnitude production alone.

In methods involving “ratio determinations,” the subject is presented a fixed stimulus alternating with the variable stimulus and is instructed to adjust the loudness of the variable to some given ratio (or fraction) of the fixed stimulus’s loudness. Often the subject is asked to halve or double the loudness, but other ratios have also been used. There are undoubtedly biases in these procedures, in that, for example, doubling loudness twice is not the same as quadrupling loudness (see Ellermeier and Faulhammer 2000; Zimmer 2005). In the method of “bisection” the subject is presented two reference stimuli differing in loudness and is instructed to adjust the variable to be midway between them. Although the method of bisection too has its biases (e.g., Gage 1934), carefully measured bisections of loudness (Garner 1954; Carterette and Anderson 1979) produce scales that, like those produced by magnitude estimation and production, can be described as power functions of sound level. The scales obtained by bisection – and by other methods in which subjects judge or compare intervals of loudness (Parker and Schneider 1974; Schneider et al. 1974) – generally have much smaller exponents than do scales obtained by magnitude estimation and production (see Marks 1974a). Because the subject controls the level of the variable, all of these methods are likely to be affected by the adjustment biases described earlier. In fact, Stevens and Poulton (1956) found that results obtained in these adjustment procedures depend on the attenuation characteristics of the device and

advocate the use of a “sone potentiometer,” which is designed to make loudness in sones an approximately linear function of angular position of the unmarked, smoothly rotating knob.

In the method of cross-modality matching, the loudness of a sound is matched to the magnitude of a percept in another modality, such as line length (or string length), brightness, tactile vibration, or the magnitude of the other percept is matched to the loudness. Cross-modality matching between loudness and line length is most common (Teghtsoonian and Teghtsoonian 1983). Cross-modality matches are consistent with results obtained by magnitude estimation in subjects with normal hearing and hearing losses (Hellman 1991). Results obtained in individual subjects are more consistent in cross-modality matching with line length or string length than in magnitude estimation, especially for short-duration sounds (Green and Luce 1974; Hellman and Meiselman 1988; Epstein and Florentine 2005, 2006a).

2.3.2.4 Magnitude Estimation, Magnitude Production, Cross-Modality Matching, and the Principle of Equality

Over the past half-century, the methods of magnitude estimation, magnitude production, and cross-modality matching have shown themselves to be especially versatile, readily applied to study loudness perception of groups of listeners in a variety of settings and under a variety of conditions; the methods have not been tested nearly so thoroughly, however, in individual listeners. To give just a few examples, the method of magnitude estimation in particular has been used to study how loudness is affected by factors such as stimulus duration (Stevens and Hall 1966; Epstein and Florentine 2006a), the presence of masking noise (Hellman and Zwislocki 1964), delivery to one ear or two (Hellman and Zwislocki 1963; Scharf and Fishken 1970; Marks 1978; Epstein and Florentine 2009), and normal hearing vs. hearing loss (e.g., Hellman and Meiselman 1991, 1993; Marozeau and Florentine 2009).

Results obtained with scaling methods potentially provide two kinds of information, information about relative magnitude and information about equality – assuming in each case that one can minimize or take account of the pertinent sources of potential bias. The use of the qualifier “pertinent” is intended to indicate the possibility that certain biases may selectively affect one kind of information but not the other. For example, the so-called regression effect (Stevens and Greenbaum 1966) points to nonlinear relations between numerical judgments, such as magnitude estimations, and stimulus level. According to Stevens and Greenbaum, subjects tend to compress the range of whatever response variable is under their control, compressing the range of numerical responses in magnitude estimation and compressing the range of stimulus levels in magnitude production. For magnitude estimations to be unbiased, the numerical responses must be directly proportional to loudness: Quadruple the underlying loudness, and the subject should give a number four times as great. With a tendency to compress the range of numbers, subjects might only double their numerical responses when loudness quadruples. In this case, the exponent obtained in magnitude estimation would be half the size of the exponent that governs the underlying perceptions of loudness.

Although regression and similar biases affect the quantitative properties of the results, they need not necessarily affect the underlying loudness equalities. Consider the situation in which several sounds have underlying loudness values of X , while other sounds have underlying values of loudness $4X$. Then as long as all of the sounds with loudness X receive the same average judgment of loudness and all of the sounds with loudness $4X$ receive the same average judgment of loudness (whether four or only two times as great), the resulting numerical judgments will preserve the loudness equalities. Simply put, as long as the loudness of every sound is mapped to a single, uniform numerical scale, the loudness judgments will conform to the principle of equality. It is possible, of course, that subjects may use different numerical scales to judge different sounds. Consequently, it is often helpful to obtain converging information about loudness equalities with other methods, such as loudness matching. A few studies have asked to what extent results obtained by scaling methods such as magnitude estimation and magnitude production agree with equal-loudness matches. Hellman and Zwislocki (1964) found excellent agreement between measures of masking of a 1,000-Hz tone by noise as determined by magnitude production and by loudness matching, and Marks (1978) reported good agreement between measures of binaural addition predicted from magnitude estimations and determined directly by loudness matches between tones of equal and unequal SPL to the two ears. Epstein and Florentine (2006a) compared loudness measures for 5- and 200-ms tones, obtaining magnitude estimations and equal-loudness matches from the same subjects. Results indicated that both procedures provide rapid and accurate assessments of group loudness functions for brief tones, although the assessments may not be reliable enough to reveal specific characteristics of loudness in individual subjects. Comparisons of scaling data and direct matches are especially important in studies of individual differences, where magnitude estimations and loudness matches may not give equivalent measures.

Almost all of the studies discussed thus far presented listeners with static, steady-state sounds, that is, with stimuli whose levels remained constant over a single trial (except for initial rise and final decay). Most sounds encountered in the world, however, are dynamic. The levels of speech, music, and environmental noises commonly rise and fall over time, either because the levels emitted from the sources themselves change, or because the sound source, the listener, or both change their spatial locations over time. Assessing the loudness of dynamic sounds poses special questions: Can listeners judge momentary loudness? Overall or average loudness? In judging overall or average loudness, how might the listener weight the loudness experienced at different points in time?

Several experiments have studied what has been called “decruitment” in loudness: the marked decrease in loudness when sound level decreases over time, compared to comparable increases over time (Canévet and Scharf 1990; Teghtsoonian et al. 2000). When a sound decreases steadily from a high level to a low one, at the low-level the sound appears softer than it does when it is presented discretely (statically), following the same high-level sound at a comparable point in time. Testing sounds that increased steadily in their level, Marks and Slawson (1966) asked a rather different kind of question about the perception of dynamic sounds: how linear

do listeners perceive the change in loudness to be when sound level increases as a power function of time? Marks and Slawson tested a wide range of different exponents and found that subjects judged the increase in loudness to be most linear when sound intensity increased as 3.3 power of time: $I = t^{3.3}$. Given that loudness in sones equals (given appropriate units) the 0.3 power of intensity, $LS = I^{0.3}$, this outcome means that the increase in loudness was judged most linear when loudness increased linearly in sones: $LS = (t^{3.3})^{0.3} = t$.

2.3.3 Measuring Loudness of Long-Duration Sounds

Equal-loudness matching and loudness scaling are especially useful in measuring the loudness of steady state and non-steady-state sounds of relatively short duration – usually no more than a few seconds. The methods that are used in laboratories to measure the loudness of short-duration sounds are not generally useful for measuring the subjective impressions of sounds along a sound stream that varies over time and can last for long durations, such as those in daily environments (see Teghtsoonian et al. 2005). Although attempts have been made to adapt category scaling and magnitude estimation to the assessment of long-duration sounds, these methods have not been carefully evaluated. There is a need to develop and rigorously test methods for measuring the loudness of relatively long dynamic stimuli.

Two methods that may be useful are “the method of continuous judgment by category” (see Chap. 6) and the “acoustic menu” (Molino et al. 1979). The original method of continuous judgment by category uses a modified category scale to record subjective judgments over time, but the method has a number of variations. For example, continuous judgments may be made using cross-modality matching of muscular effort (Susini et al. 2002) or line-length (Kuwano and Namba 1990). The acoustic menu method uses an avoidance paradigm to measure the unpleasantness of loud sounds. Kuwano and Namba describe these methods in detail in Chap. 6.

2.4 Evaluative Summary

Most researchers studying loudness are primarily interested in obtaining measurements of loudness and are less interested in details of the methods *per se*. This is understandable given time constraints in research settings, but it is unwise to choose a method without understanding its limitations as well as its strengths. Errors made in the acquisition, treatment, and interpretation of the data can arise from ignorance of basic concepts regarding methods of measuring loudness. Every method, technique, and paradigm designed to measure loudness (or probably anything else) rests tacitly or explicitly on a set of underlying assumptions, hypotheses, or theoretical principles. For example, it had long been assumed that loudness at threshold is zero. This assumption influenced models of loudness in people with

normal hearing and hearing losses. When Buus et al. (1998) actually measured loudness at threshold, the data showed a small but positive value. Models of loudness (see Chap. 10) and standards (e.g., ANSI S3.4-2007) are now being revised in light of this new finding. The collapse of this old assumption about loudness at threshold has opened the door to question other assumptions (see Chap. 1).

To measure loudness means, *ipso facto*, to be able to determine how loudness depends, quantitatively, on all of the variables that affect it: not only on level of an acoustical signal, but also on its other stimulus variables (such as frequency, spectral content, duration, presence of background sounds, etc.). The fact that loudness depends on a multiplicity of physical, psychological, and physiological factors – that an enormous number of different stimuli and conditions can produce the same loudness – sets a minimal empirical requirement for any method to measure loudness adequately. To measure loudness adequately, the method must provide measures that are internally consistent. That is, the method must assign the same value to all of the different conditions of stimulation that produce a given level of loudness. In other words, acoustical signals that have the same loudness should have the same loudness level. In addition, loudness equalities must be transitive: If acoustic signal A is as loud as signal B, and B is as loud as C, then A must be as loud as C.

In theory, once an adequate system for loudness measurement is established, the system itself will be able to provide information about loudness and known sources of bias can be taken into account. Unfortunately, psychophysical methods in their many forms have not been tested for all potential sources of bias, and the design of very few experiments permits the ready assessment of internal consistency in the data. In such cases, it is wise to ensure that the experimental designs include checks of internal consistency.

Ignorance of the limitations of a measurement method is only one of a number of pitfalls that an experimenter must avoid. Measurements are determined not only by the experimental method, but also by the way the data are treated. A review of all the possible errors is not possible, given the many potential pitfalls in data analysis, so an example will have to suffice.

It is well known that the distributions of magnitude estimations typically are highly skewed and often log normal, leading many investigators, appropriately, to use geometric averages. This approach becomes problematic, however, if a few subjects occasionally give judgments of “zero,” because the geometric mean of a distribution containing a value of zero will be zero. An investigator may be tempted to try to circumvent the problem by adding a positive constant to all of the magnitude estimations, calculate geometric averages, then subtract out the constant. This may be satisfactory if the data have appropriate statistical properties, but these properties must be ascertained. Other, simpler, solutions include calculating medians.

The final pitfall discussed in this chapter lies in errors in the interpretation of the data. For example, whereas loudness level provides a useful scale, it informs only how the loudness of a given sound compares to that of a 1-kHz tone. That is, loudness level provides information only about loudness equalities and rank order. Importantly, loudness level does not correspond directly to the subjective magnitude of the perception. Loudness level is not the same as loudness. For example, a sound

with the loudness level of 100 phons is much more than twice as loud as a sound with the loudness level of 50 phons.

The implementation of every psychophysical method is based on a set of underlying assumptions. So too is every analytical and statistical treatment of the data and so is every interpretation of the results. Progress in every discipline of science, including psychoacoustics, comes with advances in technology, methodology, and conceptualizations. But progress also requires a firm understanding of the assumptions that underlie interpretation, analyses, and, notably, the methods.

References

- Aiba RS, Stevens SS (1964) Relation of brightness to duration and luminance under light- and dark-adaptation. *Vision Res* 4:391–401.
- Algom D, Marks LE (1984) Individual differences in loudness processing and loudness scales. *J Exp Psychol Gen* 113:571–593.
- Algom D, Marks LE (1990) Range and regression, loudness processing and loudness scales: Toward a context-bound psychophysics. *J Exp Psychol Hum Percept Perform* 16:706–727.
- Anderson NH (1970) Functional measurement and psychophysical judgment. *Psychol Rev* 77:153–170.
- Anderson NH (1981) *Foundations of Information Integration Theory*. New York: Academic Press.
- Anderson NH, Cuneo DO (1978) The height + width rule in children's judgments of quantity. *J Exp Psychol Gen* 107:335–378.
- ANSI-S3.4 (2007) *American National Standard Procedure for the Computation of Loudness of Steady Sounds*. New York: American National Standards Institute.
- Berliner MH, Durlach NI (1973) Intensity perception. IV. Resolution in roving-level discrimination. *J Acoust Soc Am* 53:1270–1287.
- Berliner JE, Durlach NI, Braida LD (1977) Intensity perception. VII. Further data on roving-level discrimination and the resolution and bias edge effects. *J Acoust Soc Am* 61:1577–1585.
- Brand T, Hohmann F (2002) An adaptive procedure for categorical loudness scaling. *J Acoust Soc Am* 112:1597–1604.
- Buus S (2002) Psychophysical methods and other factors that affect the outcome of psychoacoustic measurements. In: Tranebjærg L, Christensen-Dalsgaard J, Andersen T, Poulsen T (eds), *Genetics and the Function of the Auditory System: Proceedings of the 19th Danavox Symposium*. Copenhagen, Denmark: Holmens Trykkeri, pp. 183–225.
- Buus S, Florentine M (2001) Modifications to the power function for loudness. In: Summerfield E, Kompuss R, Lachmann T (eds), *Fechner Day 2001. Proceedings of the 17th Annual Meeting of the International Society for Psychophysics*. Berlin: Pabst, pp. 236–241.
- Buus S, Florentine M (2002) Growth of loudness in listeners with cochlear hearing losses: Recruitment reconsidered. *J Assoc Res Otolaryngol* 3:120–139.
- Buus S, Florentine M, Poulsen T (1997) Temporal integration of loudness, loudness discrimination, and the form of the loudness function. *J Acoust Soc Am* 101:669–680.
- Buus S, Müsch H, Florentine M (1998) On loudness at threshold. *J Acoust Soc Am* 104:399–410.
- Canévet G, Scharf B (1990) The loudness of sounds that increase and decrease continuously in level. *J Acoust Soc Am* 88:2136–2142.
- Carterette EC, Anderson NH (1979) Bisection of loudness. *Percept Psychophys* 26:265–280.
- Cattell JMcK (1886) The influence of the intensity of the stimulus on the length of the reaction time. *Brain* 8:510–515.

- Chocholle R, Greenbaum HB (1966) La sonie de sons purs partiellement masqués: Étude comparative par une méthode d'égalisation et par la méthode des temps de réaction [Loudness of partially masked pure tones: Comparative study by an equalization method and by the reaction time method]. *J Psychol Norm Pathol* 63:387–414.
- Churcher BG (1935) A loudness scale for industrial noise measurements. *J Acoust Soc Am* 6:216–225.
- Clark A (1993) *Sensory Qualities*. Oxford: Oxford University Press.
- Cornsweet TN (1962) The staircase-method in psychophysics. *Am J Psychol* 75:485–491.
- Cross DV (1973) Sequential dependencies and regression in psychophysical judgments. *Percept Psychophys* 14:547–552.
- Cuneo DO (1982) Children's judgments of numerical quantity: A new view of early quantification. *Cogn Psychol* 14:13–44.
- Dawes RM (1972) *Fundamentals of Attitude Measurement*. New York: Wiley.
- Delboeuf JR (1873) Étude psychophysique: Recherches théorétiques et expérimentales sur la mesure des sensations, et spécialement des sensations de lumière et de fatigue [Psychophysical study: Theoretical and experimental research on the measurement of sensations, especially sensations of light and fatigue]. *Mémoires de l'Académie Royale de Belgique* 23:3–115.
- Dzhafarov EN, Colonius H (2005) Psychophysics without physics: A purely psychological theory of Fechnerian scaling in continuous stimulus spaces. *J Math Psychol* 49:1–50.
- Ellermeier W, Faulhammer G (2000) Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Percept Psychophys* 62:1505–1511.
- Epstein M, Florentine M (2005) A test of the equal-loudness-ratio hypothesis using cross-modality matching functions. *J Acoust Soc Am* 118:907–913.
- Epstein M, Florentine M (2006a) Loudness of brief tones measured by magnitude estimation and loudness matching. *J Acoust Soc Am* 119:1943–1945.
- Epstein M, Florentine M (2006b) Reaction time to 1- and 4-kHz tones as a function of sensation level in listeners with normal hearing. *Ear Hear* 27:424–429.
- Epstein M, Florentine M (2009) Binaural loudness summation for speech and tones presented via earphones and loudspeakers. *Ear Hear* 30:234–237.
- Falmagne JC (1985) *Elements of Psychophysical Theory*. Oxford: Oxford University Press.
- Fechner GT (1860/1966) *Elemente der Psychophysik*. Leipzig, Germany: Breitkopf und Härtel. [Elements of Psychophysics, Adler HE (trans), Howes DH, Boring EG (eds). New York: Holt, Rinehart, and Winston].
- Fletcher H, Munson WA (1933) Loudness, its definition, measurement, and calculation. *J Acoust Soc Am* 5:82–108.
- Florentine M, Epstein M (2006) To honor Stevens and repeal his law (for the auditory system) In: Kornbrot DE, Msetfi RM, MacRae AW (eds), *Fechner Day 2006. Proceedings of the 22nd Annual Meeting of the International Society for Psychophysics*. St. Albans, England: ISP, pp. 37–42.
- Florentine M, Buus S, Bonding P (1978) Loudness of complex sounds as a function of the standard stimulus and the number of components. *J Acoust Soc Am* 64:1036–1040.
- Florentine M, Reed C, Durlach NI, Braida LD (1979) Intensity discrimination and loudness matches in subjects with sensorineural hearing loss. In: Wolf JJ, Klatt DH (eds), *Speech Communication Papers*. New York: Acoustical Society of America. *J Acoust Soc Am Speech Commun Papers*, pp. 575–578.
- Florentine M, Buus S, Mason CR (1987) Level discrimination as a function of level for tones from 0.25 to 16 kHz. *J Acoust Soc Am* 81:1528–1541.
- Florentine M, Buus S, Poulsen T (1996) Temporal integration of loudness as a function of level. *J Acoust Soc Am* 99:1633–1644.
- Florentine M, Buus S, Robinson M (1998) Temporal integration of loudness under partial masking. *J Acoust Soc Am* 104:999–1007.
- Freeman K (1948) *Ancilla to the Pre-Socratic Philosophers: A Complete Translation of the Fragments in Diels Fragmente der Vorsokratiker*. Oxford: Blackwell.

- Gage FH (1934) An experimental investigation of the measurability of auditory sensation. *Proc R Soc Lond* 116B:103–122.
- Garner WR (1954) Context effects and the validity of loudness scales. *J Exp Psychol* 48:218–224.
- Garner WR (1960) Rating scales, discriminability, and information transmission. *Psychol Rev* 67:343–352.
- Geiger PH, Firestone FA (1933) The estimation of fractional loudness. *J Acoust Soc Am* 5:25–30.
- Gelfand SA (2004). *Hearing – An Introduction to Psychological and Physiological Acoustics* (4th Ed). New York: Marcel Dekker.
- Gescheider G (1997) *Psychophysics: The Fundamentals*. Mahwah, NJ: Lawrence Erlbaum.
- Gigerenzer G, Strube G (1983) Are there limits to binaural additivity of loudness? *J Exp Psychol* 9:126–136.
- Green DM, Luce RD (1974) Variability of magnitude estimates: A timing theory analysis. *Percept Psychophys* 15:291–300.
- Guilford JP (1954) *Psychometric Methods* (2nd Ed). New York: McGraw-Hill.
- Gulick WL, Gescheider GA, Frisina RD (1989) *Hearing: Physiological Acoustics, Neural Coding, and Psychoacoustics*. New York: Oxford University Press.
- Hall JL (1981) Hybrid adaptive procedure for estimation of psychometric functions. *J Acoust Soc Am* 69:1763–1769.
- Ham LB, Parkinson JS (1932) Loudness and intensity relations. *J Acoust Soc Am* 3:511–534.
- Heller O (1985) Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung (KU) [Listening field audiometry by the process of categorical subdivision (KU)]. *Psychologische Beiträge* 27:478–493.
- Hellman RP (1976) Growth of loudness at 1000 and 3000 Hz. *J Acoust Soc Am* 60:672–679.
- Hellman RP (1991) Loudness measurement by magnitude scaling: Implications for intensity coding. In: Bolanowski SJ Jr, Gescheider GA (eds), *Ratio Scaling of Psychological Magnitude*. Hillsdale, NJ: Lawrence Erlbaum, pp. 215–228.
- Hellman RP, Meiselman CH (1988) Prediction of individual loudness exponents from cross-modality matching. *J Speech Hear Res* 31:605–615.
- Hellman RP, Meiselman CH (1991) Loudness relations for individuals and groups in normal and impaired hearing. *J Acoust Soc Am* 86:2596–2606.
- Hellman RP, Meiselman CH (1993) Rate of loudness growth for pure tones in normal and impaired hearing. *J Acoust Soc Am* 93:966–975.
- Hellman RP, Zwislocki JJ (1961) Some factors affecting the estimation of loudness. *J Acoust Soc Am* 33:687–694.
- Hellman RP, Zwislocki JJ (1963) Monaural loudness function at 1000 cps and interaural summation. *J Acoust Soc Am* 35:856–865.
- Hellman RP, Zwislocki JJ (1964) Loudness function of a 1000-cps tone in the presence of a masking noise. *J Acoust Soc Am* 36:1618–1627.
- Hellman RP, Scharf B, Teghtsoonian M, Teghtsoonian R (1987) On the relation between growth of loudness and the discrimination of intensity of pure tones. *J Acoust Soc Am* 82:448–453.
- Hellström A (1979) Time errors and differential sensation weighting. *J Exp Psychol Hum Percept Perform* 5:460–477.
- Houtsma AJM, Durlach NI, Braida LD (1980) Intensity perception. XI. Experimental results on the relation of intensity resolution to loudness matching. *J Acoust Soc Am* 68:807–813.
- Hübner R, Ellermeier W (1993) Additivity of loudness across critical bands: A critical test. *Percept Psychophys* 54:185–189.
- International Organization for Standardization (1959) *ISO/R 131:1959 Acoustics. Expression of the Physical and Subjective Magnitudes of Sound*. Geneva: International Organization for Standardization.
- International Organization for Standardization (2003) *ISO 226:2003 Acoustics. Normal Equal-Loudness Contours*. Geneva: International Organization for Standardization.
- International Organization for Standardization (2006) *ISO 16832 Acoustics. Loudness Scaling by Means of Categories*. Geneva: International Organization for Standardization.

- Jesteadt W (1980) An adaptive procedure for subjective judgments. *Percept Psychophys* 28:85–88.
- Jesteadt W, Luce RD, Green DM (1977) Sequential effects in judgments of loudness. *J Exp Psychol Hum Percept Perform* 3:92–104.
- Johnson JH, Turner CW, Zwislowski JJ, Margolis RH (1993) Just noticeable differences for intensity and their relation to loudness. *J Acoust Soc Am* 93:983–991.
- Kohfeld DL (1971) Simple reaction time as a function of stimulus intensity in decibels of light and sound. *J Exp Psychol* 88:251–257.
- Kohfeld DL, Santee JL, Wallace ND (1981a) Loudness and reaction time: I. *Percept Psychophys* 29:535–549.
- Kohfeld DL, Santee JL, Wallace ND (1981b) Loudness and reaction time: II. Identification of detection components at different intensities and frequencies. *Percept Psychophys* 29:550–562.
- Krüger JG (1743) *Naturlehre* [Lectures on nature]. Halle-Magdeburg: Hemmerde.
- Kuwano S, Namba S (1990) Continuous judgment of loudness and annoyance. In: Müller F (ed), *Fechner Day 90. Proceedings of the 6th Annual Meeting of the International Society for Psychophysics*. Würzburg, Germany: ISP, pp. 129–134.
- Laird DA, Taylor E, Wille HH Jr (1932) The apparent reduction of loudness. *J Acoust Soc Am* 3:393–401.
- Lakoff G, Johnson M (1980) *Metaphors We Live By*. Chicago, IL: University of Chicago Press.
- Laming DRJ (1997) *The Measurement of Sensation*. Oxford: Oxford University Press.
- Leek MR (2001) Adaptive procedures in psychophysical research. *Percept Psychophys* 63:1279–1292.
- Levelt WJM, Riemersma JB, Bunt AA (1972) Binaural additivity of loudness. *Br J Math Statist Psychol* 25:51–68.
- Levitt H (1971) Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49:467–477.
- Lewis MF (1965) Category judgments as functions of flash luminance and duration. *J Opt Soc Am* 55:1555–1560.
- Lim LS, Rabinowitz WM, Braida LD, Durlach NI (1977) Intensity perception. VIII. Loudness comparisons between different types of stimuli. *J Acoust Soc Am* 62:1256–1267.
- Link SW (1992) *The Wave Theory of Difference and Similarity*. Hillsdale, NJ: Lawrence Erlbaum.
- Locke J (1690) *An Essay Concerning Humane Understanding*. London: Basset.
- Luce RD (1994) Thurstone and sensory scaling: Then and now. *Psychol Rev* 101:271–277.
- Luce RD, Edwards W (1958) The derivation of subjective scales from just noticeable differences. *Psychol Rev* 65:222–237.
- Luce RD, Green DM (1972) A neural timing theory for response times and the psychophysics of intensity. *Psychol Rev* 79:14–57.
- Luce RD, Green DM (1974) The response ratio hypothesis for magnitude estimation. *J Math Psychol* 11:1–14.
- Luce RD, Krumhansl CL (1988) Measurement, scaling, and psychophysics. In: Atkinson RC, Herrnstein RJ, Lindzey G, Luce RD (eds), *Stevens' Handbook of Experimental Psychology* (2nd Ed), Vol. 1. New York: Wiley, pp. 3–74.
- Luce RD, Tukey JW (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. *J Math Psychol* 1:1–27.
- MacCorquodale K, Meehl PE (1948) On a distinction between hypothetical constructs and intervening variables. *Psychol Rev* 55:95–107.
- Marks LE (1968) Stimulus-range, number of categories, and form of the category-scale. *Am J Psychol* 81:467–479.
- Marks LE (1974a) On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as science. *Percept Psychophys* 16:358–375.
- Marks LE (1974b) *Sensory Processes: The New Psychophysics*. New York: Academic Press.
- Marks LE (1978) Binaural summation of the loudness of pure tones. *J Acoust Soc Am* 64:107–113.

- Marks LE (1979) A theory of loudness and loudness judgments. *Psychol Rev* 86:256–285.
- Marks LE (1980) Binaural summation of loudness: Noise and two-tone complexes. *Percept Psychophys* 27:489–498.
- Marks LE (1996) Psychophysics in the scientific market-place: Peer review of grant applications. In: Masin S (ed), *Fechner Day 96. Proceedings of the 12th Annual Meeting of the International Society for Psychophysics*. Padua, Italy: ISP, pp. 329–334.
- Marks LE, Algom D (1998) Psychophysical scaling. In: Birnbaum MH (ed), *Measurement, Judgment, and Decision Making*. San Diego, CA: Academic Press, pp. 81–178.
- Marks LE, Gescheider GA (2002) Psychophysical scaling. In: Wixted J, Pashler H (eds), *Stevens's Handbook of Experimental Psychology (3rd Ed)*. Vol. 4. Methodology. New York: Wiley, pp. 91–138.
- Marks LE, Slawson AW (1966) Direct test of the power function for loudness. *Science* 154: 1036–1037.
- Marks LE, Hammeal RJ, Bornstein MH (1987) Perceiving similarity and comprehending metaphor. *Monogr Soc Res Child Dev* 42:1–91.
- Marks LE, Galanter E, Baird JC (1995) Binaural summation after learning psychophysical functions for loudness. *Percept Psychophys* 57:1209–1216.
- Marozeau J, Florentine M (2009) Testing the binaural equal-loudness-ratio hypothesis with hearing-impaired listeners. *J Acoust Soc Am* 126:310–317.
- McGill WJ (1961) Loudness and reaction time: A guided tour of the listener's private world. *Acta Psychologica* 19:193–199.
- McGill WJ, Goldberg JP (1968) Pure-tone intensity discrimination and energy detection. *J Acoust Soc Am* 44:576–581.
- Merkel J (1888) Die Abhängigkeit zwischen Reiz und Empfindung [The relation between stimulus and sensation]. *Philosophische Studien* 4:541–594.
- Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev* 63:81–97.
- Molino JA, Zerdy GA, Lerner ND, Harwood DL (1979) Use of the “acoustic menu” in assessing human response to audible (corona) noise from electric transmission lines. *J Acoust Soc Am* 66:1435–1445.
- Newman E B (1933) The validity of the just noticeable difference as a unit of psychological magnitude. *Trans Kans Acad Sci* 36:172–175.
- Ozimek E, Zwislocki JJ (1996) Relationships of intensity discrimination to sensation and loudness levels: Dependence on sound frequency. *J Acoust Soc Am* 100:3304–3320.
- Parducci A (1965) Category judgment: A range-frequency model. *Psychol Rev* 72:407–418.
- Parducci A (1974) Contextual effects: A range-frequency analysis. In: Carterette EC, Friedman MP (eds), *Handbook of Perception*, Vol. 2. Psychophysical Judgment and Measurement. New York: Academic Press, pp. 127–141.
- Parker S, Schneider B (1974) Non-metric scaling of loudness and pitch using similarity and difference estimates. *Percept Psychophys* 15: 238–242.
- Parker S, Schneider B (1980) Loudness and loudness discrimination. *Percept Psychophys* 28:398–406.
- Piéron H (1914) Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices [Research on the laws of variation of sensory latency as a function of excitatory intensity]. *L'Année Psychologique* 20:2–96.
- Piéron H (1952) *The Sensations: Their Functions, Processes and Mechanisms*. New Haven: Yale University Press.
- Pollack I (1965a) Iterative techniques for unbiased rating scales. *Q J Exp Psychol* 17:139–148.
- Pollack I (1965b) Neutralization of stimulus bias in the rating of grays. *J Exp Psychol* 69:564–578.
- Poulton EC (1968) The new psychophysics: Six models for magnitude estimation. *Psychol Bull* 69:1–19.
- Poulton EC (1989) *Bias in Quantifying Judgments*. Hove, England: Lawrence Erlbaum.

- Raab DH (1962) Magnitude estimation of the brightness of brief foveal stimuli. *Science* 135:42–44.
- Raab D, Fehrer E, Hershenson M (1961) Visual reaction time and the Broca-Sulzer phenomenon. *J Exp Psychol* 61:193–199.
- Reynolds GS, Stevens SS (1960) Binaural summation of loudness. *J Acoust Soc Am* 32:1337–1344.
- Richardson LF, Ross JS (1930) Loudness and telephone current. *J Gen Psychol* 3:288–306.
- Riesz RR (1933) The relationship between loudness and the minimum perceptible increment of intensity. *J Acoust Soc Am* 5:211–216.
- Robinson DW, Dadson RS (1956) A re-determination of the equal-loudness relations for pure tones. *Brit J Appl Phys* 7:166–181.
- Savage CW (1970) *The Measurement of Sensation: A Critique of Perceptual Psychophysics*. Berkeley: University of California Press.
- Scharf B (1959) Loudness of complex sounds as a function of the number of components. *J Acoust Soc Am* 31:783–785.
- Scharf B (1961) Loudness summation under masking. *J Acoust Soc Am* 33:503–511.
- Scharf B, Fishken D (1970) Binaural summation of loudness: Reconsidered. *J Exp Psychol* 86:374–379.
- Schneider B, Parker S (1987) Intensity discrimination and loudness for tones in notched noise. *Percept Psychophys* 41:253–261.
- Schneider B, Parker S, Stein D (1974) The measurement of loudness using direct comparisons of sensory intervals. *J Math Psychol* 11:259–273.
- Silva I, Florentine M (2006) Effect of adaptive psychophysical procedure on loudness matches. *J Acoust Soc Am* 120:2124–2131.
- Stevens JC (1958a) Stimulus spacing and the judgment of loudness. *J Exp Psychol* 56:246–250.
- Stevens JC, Hall JW (1966) Brightness and loudness as a function of stimulus duration. *Percept Psychophys* 1:319–327.
- Stevens SS (1936) A scale for the measurement of a psychological magnitude: Loudness. *Psychol Rev* 43:405–416.
- Stevens SS (1946) On the theory of scales of measurement. *Science* 103:677–680.
- Stevens SS (1955) The measurement of loudness. *J Acoust Soc Am* 27:815–829.
- Stevens SS (1956) The direct estimation of sensory magnitudes – loudness. *Am J Psychol* 69:1–25.
- Stevens SS (1958b) Problems and methods of psychophysics. *Psychol Bull* 55:177–196.
- Stevens SS (1959) Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *J Exp Psychol* 57:201–209.
- Stevens SS (1975) *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects*. New York: Wiley.
- Stevens SS, Galanter EH (1957) Ratio scales and category scales for a dozen perceptual continua. *J Exp Psychol* 54:377–411.
- Stevens SS, Greenbaum HB (1966) Regression effect in psychophysical judgment. *Percept Psychophys* 1:439–446.
- Stevens SS, Poulton EC (1956) The estimation of loudness by unpracticed observers. *J Exp Psychol* 51:71–78.
- Stillman JA, Zwislocki JJ, Zhang M, Cefaratti LK (1993) Intensity just-noticeable differences at equal-loudness levels in normal and pathological ears. *J Acoust Soc Am* 93:425–434.
- Susini P, McAdams S, Smith Benett K (2002) Global and continuous estimation of sounds with time-varying intensity. *Acta Acoustica united with Acustica* 88:536–548.
- Takeshima H, Suzuki Y, Fujii H, Kumagai M, Ashihara K, Fujimori T, Sone T (2001) Equal-loudness contours measured by the randomized maximum likelihood sequential procedure. *Acta Acoustica united with Acustica* 87:389–399.
- Teghtsoonian M, Teghtsoonian R (1983) Consistency of individual exponents in cross-modal matching. *Percept Psychophys* 33:203–214.

- Teghtsoonian R. (1971). On the exponents in Stevens' law and the constants in Ekman's law. *Psychol Rev* 78:71–80.
- Teghtsoonian R (1973) Range effects of psychophysical scaling and a revision of Stevens' law. *Am J Psychol* 86:3–27.
- Teghtsoonian R, Teghtsoonian M, Canévet G (2000) The perception of waning signals: Decruitment in loudness and perceived size. *Percept Psychophys* 62:637–646.
- Thurstone LL (1927) A law of comparative judgment. *Psychol Rev* 34:273–286.
- Wagner E, Florentine M, Buus S, McCormack J (2004) Spectral loudness summation and simple reaction time. *J Acoust Soc Am* 116:1681–1686.
- Ward LM (1987) Remembrance of sounds past: Memory and psychophysical scaling. *J Exp Psychol Hum Percept Perform* 13:216–227.
- West R, Ward M, Khosla R (2000) Beyond magnitude estimation: Constrained scaling and the elimination of idiosyncratic response bias. *Percept Psychophys* 62:137–151.
- Zeng F-G, Turner CW (1991) Binaural loudness matches in unilaterally impaired listeners. *Quart J Exp Psychol* 43A: 565–583.
- Zimmer K (2005) Examining the validity of numerical ratios in loudness fractionation. *Percept Psychophys* 67:569–579.
- Zwicker E (1958) Über psychologische und methodische Grundlagen der Lautheit [On psychological and methodological bases of loudness]. *Acustica* 8:237–258.
- Zwicker E, Flottorp G, Stevens SS (1957) Critical band width in loudness summation. *J Acoust Soc Am* 29:548–557.
- Zwislocki JJ (1965) Analysis of some auditory characteristics. In: Luce RD, Bush RR, Galanter E (eds), *Handbook of Mathematical Psychology*, Vol. 3. New York: Wiley, pp. 1–97.
- Zwislocki JJ (1983) Group and individual relations between sensation magnitudes and their numerical estimates. *Percept Psychophys* 33:460–468.
- Zwislocki JJ, Goodman DA (1980) Absolute scaling of sensory magnitudes: A validation. *Percept Psychophys* 28:28–38.
- Zwislocki JJ, Jordan HN (1986) On the relations of intensity jnd's to loudness and neural noise. *J Acoust Soc Am* 79:772–780.

Loudness

Florentine, M.; Popper, A.N.; Fay, R.R. (Eds.)

2011, XIV, 290 p., Hardcover

ISBN: 978-1-4419-6711-4