Perception & Psychophysics
2005, 67 (4), 569-579

Examining the validity of numerical ratios in loudness fractionation

KARIN ZIMMER
Aalborg Universitet, Aalborg, Denmark

Direct psychophysical scaling procedures presuppose that observers are able to directly relate a numerical value to the sensation magnitude experienced. This assumption is based on fundamental conditions (specified by Luce, 2002), which were evaluated experimentally. The participants' task was to adjust the loudness of a 1-kHz tone so that it reached a certain prespecified fraction of the loudness of a reference tone. The results of the first experiment suggest that the listeners were indeed able to make adjustments on a ratio scale level. It was not possible, however, to interpret the nominal fractions used in the task as "true" scientific numbers. Thus, Stevens's (1956, 1975) fundamental assumption that an observer can directly assess the sensation magnitude a stimulus elicits did not hold. In the second experiment, the possibility of establishing a specific, strictly increasing transformation function that related the overt numerals to the latent mathematical numbers was investigated. The results indicate that this was not possible for the majority of the 7 participants.

Scaling constitutes the assignment of numbers to sensations. In most applications, *direct* estimation (or production) procedures are employed, for which observers are asked to relate a numerical value to the sensation magnitude a stimulus elicits. The number words used in this straightforward assessment are then considered to correspond to the proper (mathematical) numbers they signify, which in turn are interpreted as scale values on the sensation scale in question.

A typical task involves presenting the listeners with two sounds. The loudness of the first sound, the standard, is given a prespecified modulus value, often 10 or 100. The listener is then asked to evaluate the loudness of the second sound by assigning a number that reflects the ratio of loudnesses of the two sounds. Thus, if the second sound is judged to be twice as loud as the standard, it will be assigned the number 20 (or 200); if it is one fourth as loud, the number given is 2.5 (or 25). If the observer assigns the numbers, the method is called *magnitude estimation*; if number words are presented and the corresponding stimulus levels are sought, the procedure is called *magnitude production*. A classical variant of magnitude production, also used in this study, asks observers to adjust fractions, such as ½ or ⅓ of magnitudes, and has been termed *fractionation*, or *ratio production* (e.g., Gescheider, 1997). Because observers are asked to judge the magnitude ratio of two sounds, it is presumed, furthermore, that the numbers assigned (or matched with an appropriate sound level) are valid on a ratio scale.

These procedures were first used to study brightness perception ("Methode der doppelten Reize"; Merkel, 1888) and to describe the strength of auditory imagery and the loudness of sine tones (Richardson, 1929; Richardson & Ross, 1930), and they have been widespread ever since Stevens (e.g., 1956, 1975) propagated them in his seminal work. Assessments obtained by procedures of direct scaling are generally reliable and consistent (for investigations in loudness scaling, see, e.g., Collins & Gescheider, 1989; Hellman, 1976; Teghtsoonian & Teghtsoonian, 1971). Their validity, however, rests on two fundamental assumptions first laid out by Narens (1996)—namely, (1) that observers are able to estimate (or produce) sensation magnitudes that are meaningful on a ratio scale level and (2) that the number words the observers use can be taken "at face value" (i.e., that they are identical to the mathematical numbers they denote).

Narens's (1996) theory of magnitude scaling specifies the fundamental conditions, or axioms, that have to be met in order for these substantial assumptions to hold. The theory covers magnitude estimations or productions with a standard (modulus). It is based on a behavioral ax-

Portions of the data were presented at the First Tutorials and Research Workshop on the Auditory Quality of Systems, April 22–24, 2003, at Mont Cenis, Germany, and at the Nineteenth Annual Meeting of the International Society for Psychophysics, October 19–23, 2003, at Larnaca, Cyprus. This work has been supported in part by the Institut für Kognitionsforschung (now Institut für Psychologie), Universität Oldenburg, Germany, and by the Sound Quality Research Unit, Department of Acoustics, Aalborg Universitet, Denmark. The author thanks Oliver Baumann for collecting the majority of the data and keeping the participants, especially those in Experiment 2, in good humor all the while. The author is much indebted to R. Duncan Luce for suggesting an alternative evaluation of reduction invariance (Experiment 2) that leads to a clear-cut conclusion with respect to the validity of this axiom and for his detailed comments on a previous version of the manuscript. The author also gratefully acknowledges the comments by Ragnar Steingrimsson and an anonymous reviewer during the reviewing process. Correspondence concerning this article should be addressed to K. Zimmer, Leibniz-Institut für die Pädagogik der Naturwissenschaften, Christian-Albrechts-Universität zu Kiel, Olshausenstr. 62, D-24098 Kiel, Germany (e-mail: zimmer@ipn.uni-kiel.de).

iomatization that relates the number words used to mathematical numbers and on a cognitive axiomatization connecting the mathematical numbers to the observer's sensations. Only the behavioral axiom system can be subjected to an empirical test.

In 2002, Luce presented a comprehensive theory of psychophysical scaling that treats, and seeks to unify, its most important methods—namely, the sensory integration of stimuli, magnitude (or proportion) scaling, and cross-modality matching. Luce's theory incorporates an extension of Narens's behavioral axiomatization (Luce, 2002, p. 525), allowing for fractions to be used in ratio assessments. Moreover, he presents a fundamental axiom that leads to a possible form for a transformation function relating the overt numerical judgments to the underlying (mathematical) numbers.

In both axiom systems, the authors distinguish between the physical stimulus (**X**), the ratios to be estimated or produced ($p*$, $q*$, $r*$), and the corresponding mathematical numbers (p, q, r).[1] They then go on to define two fundamental conditions that have to be met empirically, in order to be able to validly use the direct-scaling procedures in the standard way advocated by Stevens. These axioms are termed (1) *commutativity* (Narens, 1996, Axiom 4, p. 114), or *threshold proportion commutativity* (Luce, 2002, Equation 35, p. 525), and (2) *multiplicativity* (Narens, 1996, Axiom 9, p. 119), or *probability reduction property* (Luce, 2002, unlabeled equation on p. 525).

1. Assessments are (threshold proportion) commutative, if the order in which successive adjustments are made is irrelevant: $p * q * \mathbf{X} \sim q * p * \mathbf{X}$.

2. Multiplicativity, or probability reduction, holds, if the loudness of two successive adjustments and the loudness of a single adjustment match, whenever the product of the corresponding mathematical numbers (for the successive adjustments) equals the mathematical number denoted by the ratio to be produced in the single adjustment: $p * q * \mathbf{X} \sim r * \mathbf{X}$; with $\mathrm{p} \cdot \mathrm{q} = \mathrm{r}$.

If (threshold proportion) commutativity holds, together with a number of other assumptions that either are technical in nature or can be assumed to hold experimentally in the domain investigated, it can be proven that the assessments given are indeed valid on a ratio scale. If, moreover, assessments are multiplicative (i.e., fulfill the probability reduction property), the number words used may be interpreted as corresponding to the mathematical numbers they stand for.

In the only experimental investigation of Narens's (1996) axiom system published so far, Ellermeier and Faulhammer (2000) tested commutativity and multiplicativity in the magnitude productions of the loudness of 1-kHz tones, using 2∗, 3∗, and 6∗ adjustments. The authors found commutativity to hold and multiplicativity to be violated in the majority of listeners and thus concluded that although listeners are generally able to make adjustments on a ratio scale level, contrary to established procedures, the number words used cannot be taken at face value.

In cases such as this one, the attempt at scaling is stuck in a quagmire: Although the magnitude judgments are proven to be valid on a ratio scale, the scale values proper cannot be retrieved.

The logical next step is to investigate the *possibility* of finding a functional relationship between the number words used in the magnitude productions and the corresponding mathematical numbers reflecting the sensation magnitude. Narens's (1996) theory is concerned solely with laying the foundations for Stevens's scaling approach and, therefore, does not deal with this issue. Luce (2002), on the other hand, specifies a relevant property, termed *reduction invariance* (Theorem 2, Equations 43 and 44, p. 527), which is a generalization of the probability reduction property. If reduction invariance holds, a particular family of strictly increasing transformations, called *Prelec functions* (Prelec, 1998)—$W(p) = \exp\{-\lambda[-\ln(p)]^{\mu}\}$, λ and μ being positive constants (Luce, 2002, Equation 45, p. 527)—exist, which relate the number word used in describing the ratio to be estimated or produced to the corresponding mathematical number, p.

3. Given that the loudness of successive adjustments, $p * q * \mathbf{X}$, matches the loudness of some single adjustment $s*$ of a stimulus **X**, reduction invariance holds if the match is still valid when, instead of the loudness fractions $p*$, $q*$, and $s*$, their values are raised to the power $l = 2^n \cdot 3^m$, m and n being integers, with two different instantiations of the exponent l: If $p * q * \mathbf{X} \sim s * \mathbf{X}$, then $p^{l_1} * q^{l_1} * \mathbf{X} \sim s^{l_1} * \mathbf{X}$ and $p^{l_2} * q^{l_2} * \mathbf{X} \sim s^{l_2} * \mathbf{X}$.

Note that for the mathematical numbers p, q, and s, which the numerals p, q, and s denote, it is not necessary that $\mathrm{p} \cdot \mathrm{q} = \mathrm{s}$. The number word s used in the fractionation task can refer to any mathematical number. Reduction invariance demands just that it be the same mathematical number, no matter whether the fractions themselves or their power values are employed in the loudness assessment.

In the following, Luce's (2002) theory of proportion scaling will be put to an empirical test by investigating the validity of loudness fractionations of 1-kHz tones. While producing successive integer multiples, such as in magnitude production, tends to quickly run into ceiling problems (Ellermeier & Faulhammer, 2000), producing fractions such as ¼, ⅔, ¾, and so forth, which has not been the focus of an empirical evaluation in axiomatic frameworks before, offers more, and finer, gradations. The present investigation is also the first to address the empirical validity of reduction invariance and, thereby, the adequacy of Prelec's transformation function $W(p)$ relating numerals and numbers.

The first experiment will evaluate the assumptions inherent in Stevens's direct-scaling approach by testing the empirically relevant axioms of threshold proportion commutativity (1) and probability reduction (2), whereas the second experiment will focus on the existence of a specific family of functions that transform the numerals used into (mathematical) numbers by evaluating the reduction invariance property (3).

METHOD

Participants

A total of 13 listeners between 20 and 35 years of age participated in the experiments, 8 of them in the first (20–34 years; median age: 21.5 years), conducted at Oldenburg University, Germany, and 7 in the second (23–35 years; median age: 26.0 years), performed at Aalborg University, Denmark, a year later. The experimenters K.A. and O.L. took part in both experiments. As was established by screening audiometry, all the participants were within 20 dB of normal hearing thresholds (American National Standards Institute, 1996) in the range from 125 to 8000 Hz. The listeners were given credit toward a study requirement or were paid for participation. With the exception of the experimenters, all the participants were ignorant of the goals of the investigation.

Stimuli and Apparatus

The stimuli consisted of 1-kHz sinusoids of 500-msec duration, including 10-msec, $\cos^2$-shaped rise/decay ramps. The signals were generated digitally using a Tucker-Davis Technologies (TDT) System II AP2 signal processor, were converted with a 50-Hz sampling rate to analog signal by a 16-bit (TDT DD1) signal converter, and were passed through a low-pass filter with a cutoff frequency of 20 kHz (TDT FT6). Thereafter, sounds were attenuated to appropriate levels by a sequence of two programmable attenuators (TDT PA4), before being delivered diotically to a headphone amplifier (TDT HB6) and, further, to AKG-K 501 headphones.[2]

In the first experiment, two standard signals, at 72 and 82 dB (SPL), respectively, were employed; in the second experiment, only the higher level standard was used. Starting from around the standard level(s), the participants were instructed to generate certain loudness fractions in order to evaluate the axioms of threshold proportion commutativity and probability reduction in Experiment 1 and probability reduction and reduction invariance in Experiment 2. In Experiment 1, the following loudness fractions had to be adjusted: $\frac{2}{3}*$, $\frac{1}{2}*$, $\frac{1}{3}*$, $\frac{1}{4}*$, and $\frac{1}{6}*$ of a reference tone **X**. In Experiment 2, the fractions were $\frac{4}{5}*$, $\frac{7}{10}*$, $\frac{3}{5}*$, $\frac{1}{2}*$, $\frac{1}{3}*$, $\frac{1}{6}*$, $\frac{1}{10}*$, and $\frac{1}{20}*$ of **X**.

Procedure

Seated in a soundproof chamber, the listener was presented, over headphones, with a pair of sounds, which were separated by a 500-msec pause. The participant's task was to adjust, by a converging sequence of louder/softer judgments, the loudness of the second tone, so that it reached a certain prespecified fraction of the loudness of the first tone—that is, the standard. The participant held a custom-made response box that was equipped with three response buttons and six light-emitting diodes (LEDs). Immediately below each LED, a label was attached, with a certain fraction printed on it. During each trial, one of the LEDs was lit, and the fraction it signified had to be adjusted. The participant indicated, via a buttonpress on the response box, whether the second tone was louder or softer than the fraction to be met, and accordingly, the sound level of the second tone was adjusted before the signal pair was presented again. On the first presentation in every trial, the level of the second tone was randomly chosen from a range between 5 and 15 dB (in 1-dB steps) below the standard level. This was done to avoid the possibility of introducing a response bias by having to turn down the volume at the first in the vast majority of the experimental trials. After the listener's response, the level of the second tone initially changed by 4 dB. After the first, second, and third response reversals, level changes decreased to 2, 1, and 0.5 dB, respectively. When the participant was confident that the loudness of the second sound equaled the prespecified fraction of loudness with respect to the standard, he or she pressed an "OK" button, thereby ending the trial. The final adjustment thus reached was recorded as the outcome of the trial. After a 2-sec interval, the next trial started.

In the first experiment, a full test of all the axioms was performed in every block. In the second experiment, a full test of probability reduction (multiplicativity) and of the premise and one of the conclusions of reduction invariance was performed. The order in which either the premise and Conclusion 1 (called the *square-root condition* in the following) or the premise and Conclusion 2 (the *cube-root condition*) were tested followed a repeated ABBA–BAAB design that was counterbalanced across participants. Thus, when evaluating the square-root condition, adjustments were made for the fractions $\frac{7}{10}*$, $\frac{3}{5}*$, $\frac{1}{2}*$, $\frac{1}{3}*$, $\frac{1}{6}*$, and $\frac{1}{20}*$ of a reference **X**. In the cube-root condition, tones were adjusted to be $\frac{4}{5}*$, $\frac{7}{10}*$, $\frac{1}{2}*$, $\frac{1}{3}*$, $\frac{1}{6}*$, or $\frac{1}{10}*$ as loud as the reference.

In both experiments, the trials were presented in random order within blocks. A session consisted of four blocks and lasted approximately 45–55 min in the first experiment and 35–45 min in the second experiment. After two training sessions, the participants completed 15 blocks in 4 sessions in the first experiment. In the second experiment, in order to increase statistical power, 28 blocks of data were collected in each experimental condition. Thus, every participant ran a total of $2 \times 28 = 56$ blocks in 14 sessions.

Rationale of the Axiom Tests

Experiment 1. In evaluating threshold proportion commutativity, the fractions $\frac{2}{3}$ and $\frac{1}{4}$ were used. If the axiom holds empirically, the order in which adjustments are made is irrelevant: $\frac{2}{3}*\ \frac{1}{4}*\ \mathbf{X} \sim \frac{1}{4}*\ \frac{2}{3}*\ \mathbf{X}$. Therefore, the sound pressure levels of the tones that have been adjusted using both fractions in sequence should be the same, no matter whether they were constructed by first adjusting a tone to be $\frac{2}{3}$ as loud as the standard **X** and, subsequently, making the outcome $\frac{1}{4}$ as loud or by adjusting a tone to have $\frac{1}{4}$ of the loudness of the standard tone first and, thereafter, making the outcome $\frac{2}{3}$ as loud.

The same rationale was followed in testing probability reduction, using the fractions $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$. The axiom is valid if a tone that is adjusted to be $\frac{1}{6}$ as loud as the standard has the same sound pressure level as a tone that is adjusted to be $\frac{1}{2}$ as loud and the outcome is made $\frac{1}{3}$ as loud: $\frac{1}{6}*\ \mathbf{X} \sim \frac{1}{2}*\ \frac{1}{3}*\ \mathbf{X}$.

Experiment 2. Reduction invariance is a generalization of probability reduction that needs to be put to an empirical test only if probability reduction is violated in a given domain. Therefore, in the second experiment, probability reduction was evaluated as well, using the same fractions as those in the first experiment. Whereas probability reduction demands that a tone adjusted to be half as loud as the standard and its outcome one third as loud in order to match, in sound pressure level, a tone that is one sixth as loud as the standard, reduction invariance assumes only that there is some fraction s that corresponds to the successive adjustment: $\frac{1}{2}*\ \frac{1}{3}*\ \mathbf{X} \sim s*\ \mathbf{X}$. With this premise established, reduction invariance holds if the above equation is fulfilled when, instead of the fractions themselves, the square root and the cube root of the fractions are used: $\sqrt{\frac{1}{2}}*\ \sqrt{\frac{1}{3}}*\ \mathbf{X} \sim \sqrt{s}*\ \mathbf{X}$, and $\sqrt[3]{\frac{1}{2}}*\ \sqrt[3]{\frac{1}{3}}*\ \mathbf{X} \sim \sqrt[3]{s}*\ \mathbf{X}$.

From pilot experiments run with a different sample, the loudness adjustment $s*\ \mathbf{X}$ corresponding to the successive adjustment $\frac{1}{2}*\ \frac{1}{3}*\ \mathbf{X}$ was expected to fall between $\frac{1}{10}*\ \mathbf{X}$ and $\frac{1}{20}*\ \mathbf{X}$ for most of the listeners. For the axiom to be empirically valid, the successive adjustments of the square root and the cube root of the fractions $\frac{1}{3}$ and $\frac{1}{2}$ must, therefore, lie in the range between the square root and the cube root of $\frac{1}{10}$ and $\frac{1}{20}$, respectively.

Since the roots of these fractions are irrational numbers (given in the second column of Table 1), they cannot be expressed in true fractions that a listener is able to adjust readily. In order to make the task more easily accessible experimentally, the irrational numbers were approximated by clear-cut fractions of comparable magnitude. Thus, instead of using the square roots of $\frac{1}{3}$ and $\frac{1}{2}$, listeners were asked to adjust tones to be $\frac{3}{5}$ and $\frac{7}{10}$ as loud as the reference, respectively. Similarly, in the cube-root condition, the fractions $\frac{7}{10}$ and $\frac{4}{5}$ were used.

Table 1 gives the fractions used in the experiment, as well as the rounding error thereby introduced. The differences between the exact root values and the approximate fractions employed for the successive adjustments did not exceed a value of $0.071 \times 0.0226 \approx 0.0002$. It was assumed that this magnitude of imprecision exerted no, or only

Table 1
Fractions Used in Experiment 2

Premise: Fraction	Exact Value	Approximation Used	Numerical Difference
	Square-Root Conclusion		
1/2	$\sqrt{1/2} \approx 0.7071$	7/10	0.0071
1/3	$\sqrt{1/3} \approx 0.5774$	3/5	−0.0226
1/10	$\sqrt{1/10} \approx 0.3162$	1/3	−0.0171
1/20	$\sqrt{1/20} \approx 0.2236$	1/6	0.0569
	Cube-Root Conclusion		
1/2	$\sqrt[3]{1/2} \approx 0.7937$	4/5	−0.0063
1/3	$\sqrt[3]{1/3} \approx 0.6934$	7/10	−0.0066
1/10	$\sqrt[3]{1/10} \approx 0.4642$	1/2	−0.0358
1/20	$\sqrt[3]{1/20} \approx 0.3684$	1/3	0.0351

Note—Given are the exact fractions used for the premise and for the conclusion conditions, as well as the (rational) approximations used in evaluating the square-root and cube-root conclusion of the reduction invariance property. The rightmost column lists the numerical difference between the exact and the approximate values.

a negligible, influence on the adjustments elicited. The rounding errors of the acceptable ranges for $\sqrt{s}$ and $\sqrt[3]{s}$ in the conclusion conditions are bigger, reaching up to 0.057. Note that because of these approximations, the ranges used for empirical evaluation are larger than the ranges of acceptance warranted by the reduction invariance axiom itself.

RESULTS

Experiment 1: Evaluating Threshold Proportion Commutativity and Probability Reduction

Each participant made 15 adjustments for every fraction. Overall, 99.0% of the adjustments were made using the final, minimal step size of 0.5 dB; the remaining adjustments were based on a 1-dB step size. For 1 participant, the sound pressure levels of the fractionations were so imprecise that the adjustments of $\frac{2}{3} * \mathbf{X}$ and $\frac{1}{6} * \mathbf{X}$ did not differ statistically. She was therefore excluded from further data analyses. For the remaining 7 listeners, standard errors of the adjustments were similar for both levels of the standard and ranged between 0.63 and 3.17 dB for the single adjustments and between 0.92 and 3.68 dB for the successive adjustments, in which the outcome of one trial was used as the starting level for a subsequent trial.

For all the listeners, the variability of adjustments tended to be greater for the smaller fractions than for the larger ones, which is in accordance with the fact that the just-noticeable difference for loudness increases with decreasing sound pressure level (Jesteadt, Wier, & Green, 1977). Moreover, the distribution of adjustments around their mean is not known a priori. Therefore, in assessing threshold proportion commutativity and probability reduction, nonparametric Mann–Whitney *U* tests were used in the statistical data analyses. The significance level was set to $\alpha = .1$, in order to reduce the risk of failing to detect a true axiom violation (the alternative hypothesis).[3] Data were analyzed separately for every individual.

Threshold proportion commutativity. Table 2 lists the mean sound pressure level of the successive adjustments $\frac{2}{3} * \frac{1}{4} * \mathbf{X}$ and $\frac{1}{4} * \frac{2}{3} * \mathbf{X}$, respectively. For both standard levels **X**, adjustments are statistically different only for 1 listener, E.N. (Mann–Whitney *U* test; $\alpha = .1$). For the other participants, threshold proportion commutativity can be concluded to hold. To illustrate this result, the mean adjustments and standard errors of Listener J.E., which can be considered as typical for the present sample, are given in Figure 1. Here, the sound pressure level of the adjustments in question are denoted by symbols ($\diamond$ and $*$, for the 72-dB SPL and 82-dB SPL standard, respectively) that are connected by horizontal arrows.

Probability reduction. Probability reduction (multiplicativity) was evaluated for the 6 listeners for whom threshold proportion commutativity was found to hold. From Table 3, it can be seen that the mean sound pressure levels of the successive adjustments $\frac{1}{2} * \frac{1}{3} * \mathbf{X}$ clearly differed from the single adjustment $\frac{1}{6} * \mathbf{X}$ in 5 of the 6 participants, leading to a violation of the probability reduction axiom. Only Listener O.L. showed no statistically significant effect (Mann–Whitney *U* test; $\alpha = .1$). For the other participants, the deviations ranged from 3.3 to 13.9 dB and were thus far above the discrimination threshold. In Figure 2, this result is illustrated for Listener J.E.: The sound pressure levels of the successive adjustments lie below the corresponding single adjustment (connected by a nonhorizontal double arrow).

Interim summary. For both standard levels, the threshold proportion commutativity axiom held for 6 of 7 participants, whereas probability reduction was violated for all the participants except O.L. These results correspond to the ones reported for magnitude production by Ellermeier and Faulhammer (2000). Taken together, this indicates that listeners can assess the loudness ratio of two sounds in loudness fractionation—that is, they are able to make adjustments on a ratio scale level. Contrary to the conventional, intuitive way of analyzing such data, however, the number words used in describing these loudness ratios do not match the (mathematical) number values they signify. Thus, Stevens's fundamental assumption that observers can directly assess the sensation magnitude a stimulus elicits does not hold.

In the following experiment, it was investigated whether it is possible to establish a well-defined transformation

Table 2
Experiment 1: Evaluating Threshold Proportion Commutativity

	72-dB Standard			82-dB Standard		
Listener	⅔* ¼* *M*	¼* ⅔* *M*	$z(U)$	⅔* ¼* *M*	¼* ⅔* *M*	$z(U)$
E.N.	24.2	20.57	**2.03**	32.13	23.77	**3.01**
H.O.	13.87	14.27	0.21	22.33	19.27	1.16
J.E.	26.17	25.33	0.15	36.6	35.73	0.12
J.U.	31.73	28.67	0.85	41.7	44.13	0.64
K.A.	39.2	39.8	0.10	46.03	46.63	0.04
O.L.	51.3	49.07	1.27	56.17	56.8	0.02
S.V.	18.0	24.17	1.14	27.87	32.3	0.98

Note—Given are the means of the consecutive adjustments specified (in dB [SPL]), based on 15 observations each, and the *z* scores of the test statistic [Mann–Whitney *U* test; $z_{crit}(U) = 1.68$]. Statistically significant axiom violations are given in boldface ($\alpha = .1$).

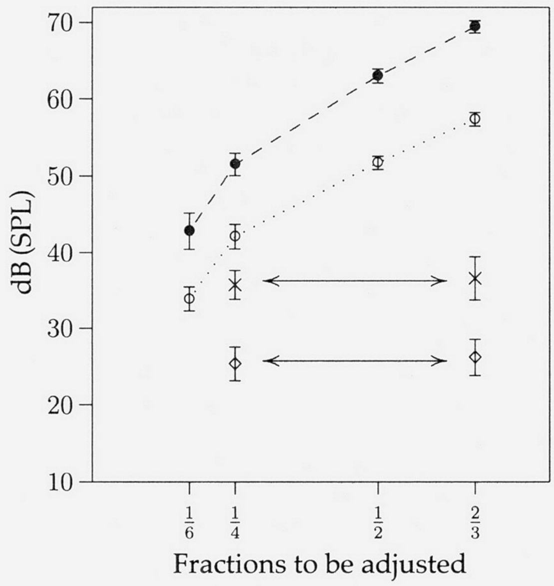

Figure 1. Experiment 1, Listener J.E.: test for threshold proportion commutativity. Given are the mean sound pressure levels and standard errors of single adjustments (denoted by • and ○, for the high and low standard levels, respectively) and of the successive adjustments (denoted by × and ◇) used in evaluating threshold proportion commutativity, $\frac{2}{3}* \frac{1}{4}*$ X versus $\frac{1}{4}* \frac{2}{3}*$ X. For the 82-dB (SPL) standard and 72-dB (SPL) standard, single adjustments are connected by a dashed and a dotted line, respectively. As can be seen from the horizontal double arrows, threshold proportion commutativity holds.

function that relates overt numerals to the latent mathematical numbers they reflect.

Experiment 2: Evaluating Probability Reduction and Reduction Invariance

Seven listeners took part in the experiment, each adjusting tones to be $\frac{1}{2}*$, $\frac{1}{3}*$, $\frac{1}{6}*$, and $\frac{1}{2}* \frac{1}{3}*$ as loud as the (82-dB) standard **X** in both experimental conditions—that is, when determining the premise of both the square-root and the cube-root conclusions. Thus, for each of these loudness fractions, a total of 2 × 28 = 56 adjustments were collected.[4] All other loudness fractions (which were used to evaluate the conclusions) were adjusted 28 times each. Ninety-six percent of all the adjustments reached the smallest step size available—that is, 0.5 dB—whereas 4% of the adjustments were made using a 1-dB step size. Standard errors were between 0.40 and 1.77 dB for single adjustments and between 0.54 and 2.44 dB for successive adjustments.

Table 3
Experiment 1: Evaluating Probability Reduction

	72-dB Standard			82-dB Standard		
Listener	$\frac{1}{2}* \frac{1}{3}*$ *M*	$\frac{1}{6}*$ *M*	$z(U)$	$\frac{1}{2}* \frac{1}{3}*$ *M*	$\frac{1}{6}*$ *M*	$z(U)$
E.N.	Commutativity violated			Commutativity violated		
H.O.	15.57	23.7	**2.74**	21.17	34.63	**4.02**
J.E.	25.6	33.87	**2.55**	34.57	42.8	**2.34**
J.U.	31.6	40.57	**2.39**	40.2	54.1	**3.76**
K.A.	37.8	41.7	**2.62**	47.0	50.3	**2.24**
O.L.	48.73	46.47	0.35	56.17	55.8	0.35
S.V.	13.27	23.57	**2.72**	23.33	33.0	**2.43**

Note—Mean adjustments of fractions (in dB [SPL]), based on 15 observations each, and the values of the test statistics are given for both standard levels. Statistically significant axiom violations are given in boldface [Mann–Whitney *U* test; $z_{crit}(U) = 1.68$, $\alpha = .1$].

In the following, tests of the probability reduction property, and of reduction invariance are given. As in the first experiment, the axioms were evaluated on an individual basis. In testing probability reduction, nonparametric Mann–Whitney *U* tests were again used. In order to evaluate reduction invariance, more than two sets of observations had to be compared—namely, the upper and the lower bound and the successive adjustments. Therefore, rank order Kruskal–Wallis analyses of variance (ANOVAs) were computed.

Probability reduction property. Differences between mean single adjustments of a tone $\frac{1}{6}$ as loud as the standard **X** and mean successive adjustments of $\frac{1}{2}* \frac{1}{3}*$ **X** ranged between 5.16 and 15.15 dB (see the two leftmost columns of Table 4).

As in the first experiment, the sound pressure levels of the successive adjustments were below those of the single adjustments for all the listeners (Mann–Whitney *U* test; $\alpha = .1$), leading to a unanimous rejection of the probability reduction property (see Table 4).

Reduction invariance. To evaluate this axiom, the premise has to be established in a first step. In the present investigation, on the basis of the results of a pilot study with a different subject sample (see the Experiment 2 section), it is said to hold if the sound pressure level of the successive adjustment of making a tone $\frac{1}{2}$ as loud as the standard **X** and its outcome $\frac{1}{3}$ as loud falls within (single) adjustments of $\frac{1}{20}*$ **X** and $\frac{1}{10}*$ **X**:

$$\tfrac{1}{20}* \mathbf{X} \precsim \tfrac{1}{2}* \tfrac{1}{3}* \mathbf{X} \precsim \tfrac{1}{10}* \mathbf{X}.$$

Given that premise, the data adhere to the reduction invariance property, if the successive adjustments in both conclusion conditions fall within an *acceptance region* given by the squared and cubed fractions, respectively—namely,

$$\tfrac{1}{6}* \mathbf{X} \precsim \tfrac{7}{10}* \tfrac{3}{5}* \mathbf{X} \precsim \tfrac{1}{3}* \mathbf{X} \text{ (square-root conclusion)},$$

and

$$\tfrac{1}{3}* \mathbf{X} \precsim \tfrac{4}{5}* \tfrac{7}{10}* \mathbf{X} \precsim \tfrac{1}{2}* \mathbf{X} \text{ (cube-root conclusion)}.$$

Figure 3 illustrates the results for 1 listener, P.A.: The mean successive adjustments for the premise condition $\frac{1}{2}* \frac{1}{3}*$ **X** (denoted by ○), the square-root Conclusion 1, $\frac{7}{10}* \frac{3}{5}*$ **X** (denoted by ◇), and the cube-root Conclusion 2 $\frac{4}{5}* \frac{7}{10}*$ **X** (denoted by ×) lie within the acceptance regions—that is, the respective dotted areas.

The results for all the participants are listed in Table 5. The first three columns give the sound pressure levels of the mean adjustments. From the next column, displaying the *H* statistic, it can be seen that the sound pressure level of adjustments differs for all the participants in all the conditions (Kruskal–Wallis ANOVA, $\alpha = .1$). Post hoc

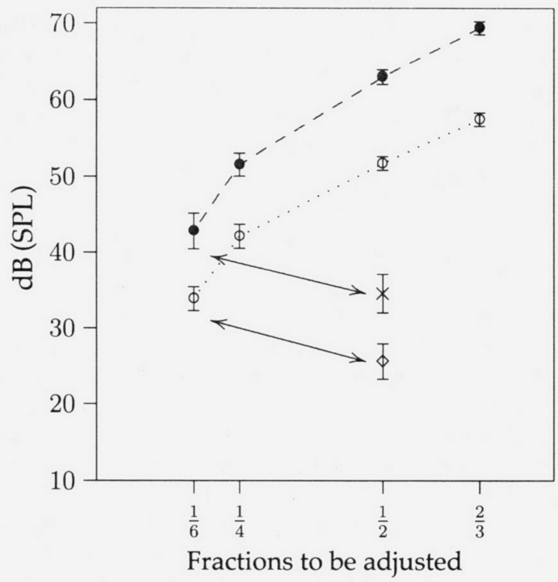

Figure 2. Experiment 1, Listener J.E.: evaluating probability reduction. Mean sound pressure levels and standard errors of the single adjustments (from Figure 1), and of the successive adjustments used in evaluating probability reduction (multiplicativity), $\frac{1}{2}$* $\frac{1}{3}$* X versus $\frac{1}{6}$* X, are given. Adjustments for the two standard levels are indicated in the same manner as in Figure 1. The nonhorizontal double arrows illustrate that, for J.E., the probability reduction property is violated.

tests (Wilcoxon & Wilcox, 1964) show that, in the premise condition, the successive adjustments are not outside the range of single adjustments (see the rightmost column of Table 5, top panel). The premise can, therefore, be concluded to hold for all the listeners.

In the square-root condition, the successive adjustment $\frac{7}{10}* \frac{3}{5}* \mathbf{X}$ falls below the acceptable range for 1 participant, B.E. (see the right column of Table 5, middle panel). In the cube-root condition, adjustments of $\frac{4}{5}* \frac{7}{10}* \mathbf{X}$ fall outside the acceptable range for B.E. again, as well as for 3 other listeners (see the rightmost column of Table 5, bottom panel). Thus, although the premise holds for all the listeners, at least one of the conclusion conditions is violated for 4 out of 7 participants. For these listeners, reduction invariance fails, and the weighting function proposed by Luce (2002) cannot be used to transfer the fraction numerals into mathematical numbers.

Alternative evaluation of reduction invariance. As was elaborated in the Method section, the acceptance regions used in testing reduction invariance are larger than the ones warranted from theory. They are also generally not centered around the successive adjustments, so that an axiom might be rejected more easily for one listener than for another. Making use of the assessments of upper and lower bounds in the premise condition ($\frac{1}{10}* \mathbf{X}$, and $\frac{1}{20}* \mathbf{X}$, respectively), Luce suggested a stronger test of axiom validity in a personal communication during the 2003 Fechner Day meeting of the International Society for Psychophysics.[5]

Given some assessment $p* \mathbf{X}$ with a corresponding lower bound $k \cdot \mathbf{X}$, Luce specifies the values of the lower bounds for the assessments $p^N* \mathbf{X}$. Given his theory holds,

$$p* \mathbf{X} > k \cdot \mathbf{X} \rightarrow p^N* \mathbf{X} > k^c \cdot \mathbf{X},$$

with $c = N^\mu$, $N = \frac{1}{2}, \frac{1}{3}$, in the present case, and μ being a positive constant in the weighting function $W(p)$. The derivation is given in the Appendix.

The parameter values for k and μ can be estimated from the premise condition as follows: With the adjustment of the lower bound (lb; see the second column in Table 5) and the standard stimulus (st; here, 82 dB [SPL]) given as sound pressure levels, k can be computed by $k = 10^{\frac{1}{20} \cdot (lb - st)}$, while μ can be estimated from the fraction instructions used in establishing the lower bound (here, $\frac{1}{20}$) and the consecutive adjustments in the premise ($\frac{1}{2}$ and $\frac{1}{3}$): $(-\ln\frac{1}{20})^\mu = (-\ln\frac{1}{2})^\mu + (-\ln\frac{1}{3})^\mu$ (see also Equation 46, p. 528, in Luce, 2002). The sound pressure level of the lower bound lb_{pred} in a conclusion condition is then predicted to be $lb_{\text{pred}} = 20 \cdot N^\mu \log_{10} k + st$.

In predicting values for the upper bound, the direction of the inequality is reversed, and the fraction instruction for the upper bound in the premise ($\frac{1}{10}$) and the corresponding adjustments are used in assessing the parameters k and μ.

Post hoc testing (Wilcoxon & Wilcox, 1964; see the rightmost column in Table 6) suggests that for 6 of 7 participants, the consecutive adjustments ($\frac{1}{2}* \frac{1}{3}* \mathbf{X}$, and their square roots and cube roots) tend toward the lower bound, whereas for Listener P.A., they approach the upper bound. For the axiom to hold, the predictions of the lower bound in the square-root and cube-root conditions, as computed from the lower bound found in the premise condition, must not undershoot the corresponding consecutive adjustments. Similarly, the predictions for the upper bound in the conclusion conditions must not surpass the consecutive adjustments in those conditions.

Table 6 gives the predicted sound pressure levels for the lower bound and the observed levels for the consecutive adjustments in the square-root and cube-root conditions for 6 of the 7 listeners. For Listener P.A., the upper bound is used in the computations. The values are

Table 4
Experiment 2: Results for Probability Reduction

Listener	$\frac{1}{2}$* $\frac{1}{3}$* *M*	$\frac{1}{6}$* *M*	$z(U)$
A.M.	59.58	64.74	**5.66**
B.E.	55.58	62.12	**6.88**
J.O.	31.37	41.27	**7.95**
K.A.	49.33	56.38	**5.97**
O.L.	21.41	30.95	**7.90**
P.A.	50.72	56.26	**4.09**
R.A.	17.34	32.49	**7.09**

Note—Mean adjustments of fractions (in dB [SPL]) and the values of the test statistics are given. Values are based on 56 observations for each adjustment and participant. Statistically significant axiom violations are given in boldface [Mann–Whitney U test; $z_{\text{crit}}(U) = 1.64$, $\alpha = .1$].

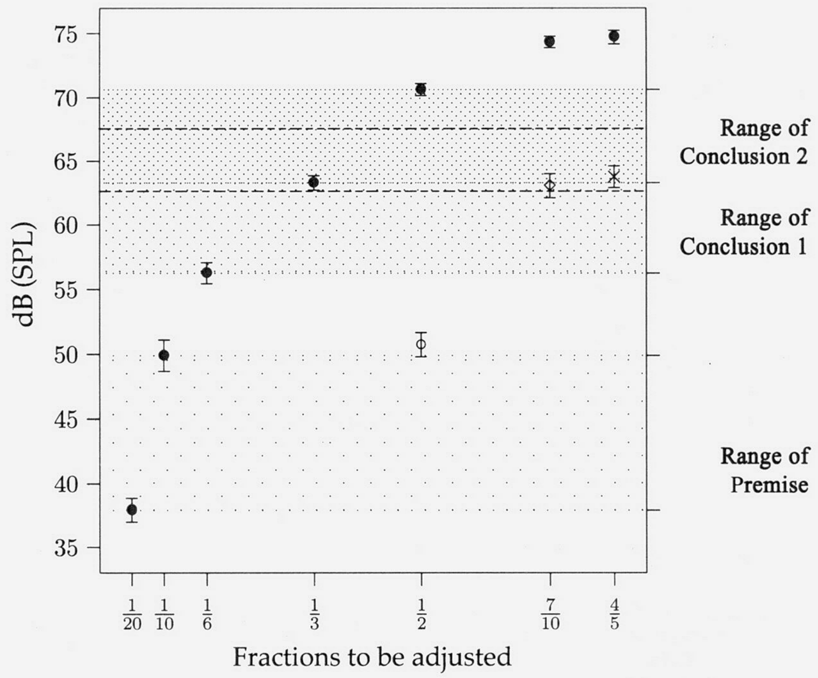

Figure 3. Experiment 2, Listener P.A.: test for reduction invariance. Given are the mean adjustments, with respective standard errors. Single adjustments are denoted by ●-symbols. The mean successive adjustments for the premise condition (½* ⅓* X; denoted by ○), the square-root Conclusion 1 (7/10* 5/6* X; denoted by ◇), and the cube-root Conclusion 2 (4/5* 7/10* X; denoted by ×) are depicted, as are the acceptance regions—that is, the respective dotted areas. The lower and upper horizontal dashed lines indicate the predicted sound pressure levels for the square-root and the cube-root conclusions, respectively. For details, see the text.

compared with the prediction by means of a one-sample t test, and statistically significant test statistics are given in boldface ($\alpha = .1$).

The axiom is found to be violated for 6 of the 7 listeners. For 5 of them, the predicted value of the lower bound is significantly higher than the corresponding consecutive adjustment in at least one of the conclusion conditions. As is illustrated in Figure 3, for Listener P.A., the consecutive adjustments lay at the upper bound in the premise (○). In the square-root condition, the successive adjustments (◇) were at the predicted upper bound (given by the lower horizontal dashed line in Figure 3), whereas the (observed) consecutive adjustment (×) fell below both the observed, and the predicted, upper bound in the cube-root condition (see the upper horizontal dashed line in Figure 3 and Tables 5 and 6), indicating that this condition is violated.

It must therefore be concluded that for only 1 in 7 participants, Listener A.M., reduction invariance can be assumed to hold when Luce's stricter test of axiom validity is applied. The family of strictly monotonic transformation functions suggested by Luce (2002) can therefore not be used, in general, to relate numerals and numbers in loudness fractionation.

DISCUSSION

In two experiments, using 1-kHz tones, the validity of loudness fractionations was investigated by testing basic conditions that are inherent in paradigms of direct scaling. In the following, the results of evaluating threshold proportion commutativity and probability reduction (multiplicativity) in the first experiment and probability reduction and reduction invariance in the second experiment will be discussed in turn.

Evaluating Stevens's Direct-Scaling Approach

With threshold proportion commutativity holding for 6 of 7 listeners, the first experiment showed that listeners are generally able to produce loudness fractions, which are meaningful on a ratio scale level of measurement. The fractions used in asking for the ratio adjustments, however, cannot be interpreted as "factual" mathematical numbers: Adjusting a tone to be half as loud as the reference and then making the outcome one third as loud resulted in tones with sound pressure levels that were (for the lower and higher standards, respectively) around 6.2 and 8.0 dB lower, on average, than the ones obtained by making a tone one sixth as loud as the refer-

Table 5
Experiment 2: Results for Reduction Invariance

Listener	$\frac{1}{20}*$ M	$\frac{1}{2}* \frac{1}{3}*$ M	$\frac{1}{10}*$ M	$H(2)$	t_3
	Premise Condition				
A.M.	58.29	59.58	62.61	**16.45**	n.s.o.
B.E.	54.96	55.58	59.25	**18.41**	n.s.o.
J.O.	32.71	31.37	39.21	**37.65**	n.s.o.
K.A.	47.66	49.33	52.11	**13.25**	n.s.o.
O.L.	17.39	21.41	24.55	**26.57**	n.s.o.
P.A.	37.93	50.72	49.89	**48.94**	n.s.o.
R.A.	17.25	17.34	25.39	**16.85**	n.s.o.
	Square-Root Conclusion				
Listener	$\frac{1}{6}*$ M	$\frac{7}{10}* \frac{3}{5}*$ M	$\frac{1}{3}*$ M	$H(2)$	t_2
A.M.	64.74	65.63	67.96	**9.24**	n.s.o.
B.E.	62.12	59.11	67.75	**43.45**	**2.24**
J.O.	41.27	41.95	46.03	**11.35**	n.s.o.
K.A.	56.38	56.50	62.89	**22.83**	n.s.o.
O.L.	30.95	33.16	37.02	**17.07**	n.s.o.
P.A.	56.26	63.05	63.29	**24.72**	n.s.o.
R.A.	32.49	27.05	41.70	**23.94**	n.s.o.
	Cube-Root Conclusion				
Listener	$\frac{1}{3}*$ M	$\frac{4}{5}* \frac{7}{10}*$ M	$\frac{1}{2}*$ M	$H(2)$	t_2
A.M.	67.96	68.18	71.82	**15.26**	n.s.o.
B.E.	67.75	59.77	70.16	**57.32**	**5.37**
J.O.	46.03	42.16	56.18	**49.59**	**2.38**
K.A.	62.89	58.98	67.79	**36.05**	**2.39**
O.L.	37.02	37.38	41.99	**16.62**	n.s.o.
P.A.	63.29	63.77	70.63	**34.85**	n.s.o.
R.A.	41.70	31.91	48.91	**24.56**	**2.59**

Note—Given are the mean sound pressure levels of adjustments in dB (SPL), based on 28 observations each, and the values of the Kruskal–Wallis test statistic H, all of which were statistically significant [$H_{\text{crit}}(2) = 4.61$; $\alpha = .1$], indicating that at least two of the three adjustments differed. To decide whether the successive adjustments ($\frac{1}{2}* \frac{1}{3}*$ **X**, $\frac{7}{10}* \frac{3}{5}*$ **X**, or $\frac{4}{5}* \frac{7}{10}*$ **X**) fell outside of the range given by the single adjustments, directional post hoc tests according to Wilcoxon and Wilcox (1964) were computed, with two tests per experimental condition and participant. If the post hoc tests turned out to be statistically significant, the test statistic t_2 is given (one-tailed criterion value $t_{2;\text{crit}} = 1.92$, testwise significance level $\alpha = .05$). The abbreviation n.s.o. indicates that the successive adjustment did not statistically lie outside of the range specified by the single adjustments.

ence sound right away. These results on loudness fractionations are in line with, and generalize, findings reported by Ellermeier and Faulhammer (2000) on magnitude productions of loudness. In their investigation, starting from levels at 40 and 55 dB (SPL), successive adjustments of doubling the loudness of a given tone and making the outcome three times as loud resulted in sound pressure levels that were about 6 dB higher in level than a single adjustment of "six times as loud" (see their Figure 2), thus showing a comparably large effect.

Functional Relationship Between Numerals and Numbers

With the probability reduction property violated in the first experiment, the second experiment served to explore whether a specific weighting function proposed by Luce (2002) was suited to transform the fraction numerals into scientific numbers reflecting sensation magnitude. Following Prelec (1998; see also Luce, 2001), Luce (2002) specified a qualitative—that is, parameter-free—behavioral property called *reduction invariance*, which is equivalent to the two-parameter Prelec function $W(p) = \exp\{-\lambda[-\ln(p)]^{\mu}\}$. Reduction invariance does not require multiplicativity to hold for the mathematical numbers corresponding to the fractions used in an experiment but, rather, uses the outcome from one experimental condition (the premise) to predict the outcomes of the conclusion conditions.

Prior to evaluating reduction invariance, the probability reduction hypothesis was again put to an empirical test, since a new sample of experimental participants was used. As before, probability reduction clearly failed in all listeners, in that successive adjustments of $\frac{1}{2}* \frac{1}{3}*$ **X** fell below single adjustments of $\frac{1}{6}*$ **X** by about 8.4 dB, on average.[6]

Furthermore, reduction invariance turned out to be violated in 6 of 7 listeners. Successive adjustments fell below predicted sound pressure levels by 4.3 and 8.0 dB in the first and second conclusions of the axiom, respectively. It must therefore be concluded that the attempt at establishing a transformation function proposed by Luce (2002) did not succeed.

It should be noted that the experimental evaluation of reduction invariance was imprecise, because, due to considerations of experimental design (outlined in the Method section), instead of the exact roots of the fractions as warranted by the reduction invariance axiom, their *smoother* (i.e., rational) approximations (given in Table 1) were used. These approximate values were slightly larger than the exact values in the second conclusion; in the first conclusion, one value was larger and one smaller than the proper fractions demanded by theory. If this *rounding error* alone had exerted an effect, then, at least in the second conclusion condition, it should have led to adjustments that tended to be louder than the predicted ones. As can be seen from Table 6, however, the observed adjustments were lower in level than predicted, in all the participants. Thus, using approximate fractions, instead

Table 6
Experiment 2: Alternative Evaluation of Reduction Invariance

	Square-Root Conclusion			Cube-Root Conclusion		
Listener	Pred.	M(Obs.)	t	Pred.	M(Obs.)	t
A.M.	66.01	65.63	−0.41	69.30	68.18	−1.30
B.E.	63.76	59.11	**−6.37**	67.51	59.77	**−14.41**
J.O.	48.75	41.95	**−6.28**	55.59	42.16	**−12.67**
K.A.	58.84	56.50	**−2.23**	63.60	58.98	**−4.83**
O.L.	38.42	33.16	**−3.88**	47.38	37.38	**−7.74**
P.A.	62.63	63.05	0.45	67.58	63.77	**−4.53**
R.A.	38.32	27.05	**−4.98**	47.31	31.91	**−6.32**

Note—Predicted (Pred.) sound pressure levels for the lower bound according to Luce's proposal (see the text) and means (in dB [SPL]) of the observed (Obs.) consecutive adjustments in the square-root and the cube-root conclusions, respectively, along with the outcome of the one-sample t tests. Significant deviations of the adjustments from the predicted bounding level are given in boldface ($t_{\text{crit}} = -1.70$, $\alpha = .1$). For Listener P.A., the adjustments were evaluated against the upper bound.

of the proper root values, did not produce a bias strong enough to (erroneously) retain or reject the axiom.

Furthermore, it is very unlikely that the results were caused by a general *floor effect*—that is, the inability of listeners to distinguish between the fractions $1/10$, $1/20$, and $1/2 * 1/3$, or their failure to produce different sound pressure levels (reflecting the respective sensation magnitudes) for the different fractions—because, for one, the Kruskal–Wallis test turns out to be statistically significant in all cases and, furthermore, the successive adjustments are at the lower bound for most, but not all, of the participants.

A functional relationship between numerals and numbers, which is not identity, the so-called *subjective number function*, has also been proposed by multistage models of psychophysical scaling (Attneave, 1962; Rule & Curtis, 1982). With the goal of disentangling sensory and judgmental aspects of the task, these models assume that a magnitude-scaling judgment is the result of two processes, the first relating the stimulus to sensation magnitude, and the second relating the sensation magnitude to a subjective number continuum. Each of these processes is supposed to be ruled by a power function.

Subsequent experimenters have tried to deduce the shape of the subjective number function—for example, by nonmetric conjoint scaling of judgments on which member of a stimulus pair consisting of a weight and a number was greater in magnitude (Rule & Curtis, 1973), by a nonmetric multidimensional scaling analysis based on similarity ratings of pairs of number words (Schneider, Parker, Ostrosky, Stein, & Kanow, 1974), or by rating the perceived randomness of a set of numbers (Banks & Coleman, 1981).

Although these experiments employed very different methods in order to estimate the functional form, they are unanimous in supporting a power function relationship between numerals—that is, subjective numbers and numbers proper—rather than a logarithmic or a linear transformation function. Note, however, that with $\mu = 1$, the transformation function $W(p)$ that Luce proposed for fractionation judgments (Luce, 2002, Equation 25, p. 527) includes power functions as a special case—$W(p) = \exp\{-\lambda[-\ln(p)]^{\mu}\} = p^{\lambda}$—for which, indeed, even the probability reduction property must hold. Thus, even though power functions seem to provide a good fit to the data collected, a fundamental condition underlying a power function representation (among other possible functional forms)—namely, reduction invariance—is violated. In principle, an axiom-testing approach such as the one taken here provides a much more severe, and valid, test of the appropriateness of any transformation function postulated than an approach based on fitting curves does. It must, therefore, be concluded that power functions cannot constitute the correct representation for relating numerals and numbers.

One might possibly argue that a given functional form may be inadequate for every single listener but can still

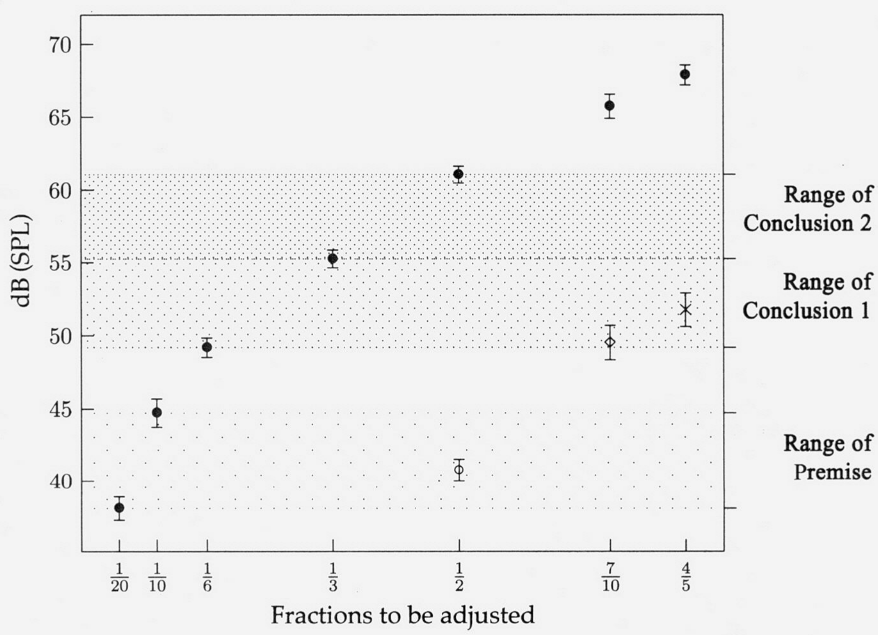

Figure 4. Results of Experiment 2: mean adjustments across all listeners. Given are the mean values of fractionation adjustments and the mean standard errors across all listeners. For the axiom to hold, the mean successive adjustments for the premise condition ($1/2 * 1/3 *$ X; denoted by ○), the square-root Conclusion 1 ($7/10 * 3/5 *$ X; denoted by ◇), and the cube-root Conclusion 2 ($4/5 * 7/10 *$ X; denoted by ×) must lie within the *acceptance regions*—that is, the respective dotted areas. Visual inspection suggests that although this may be the case for the premise and the square-root conclusion, it does not hold true for the cube-root conclusion.

constitute a valid representation for a group of listeners. This is not the case here, however. When instead of testing reduction invariance individually for every participant, the analysis is based on the pooled data set (as was the case in the experiments outlined above), the pattern of results is the same: Figure 4 gives the mean adjustments and mean standard error of adjustments across all listeners. As can be seen, the cube-root conclusion is violated, and thus reduction invariance must be rejected for the pooled data as well.

It must be noted that the present results do not refute Luce's general theory as such. As outlined in Luce (2002, 2004), the only general constraint put on the weighting function W and the psychophysical function ψ is that they be strictly increasing. On the basis of the results of the second experiment, a specific family of transformations (including the power function) can be rejected—namely, the Prelec functions proposed by Luce (2002) as adequate weighting functions relating numerals and numbers. Thus, in the present investigation, only one aspect of Luce's theory was addressed, and evaluations of other aspects are under way (e.g., Steingrimsson, 2002; Zimmer, Luce & Ellermeier, 2001).

In conclusion, the present experimental results seem to argue for a function with a *steeper slope* than can be modeled by the class of transformation functions Luce (2002) proposes. The future challenge, however, will *not* be to find some such function that provides a (more or less) satisfactory fit to the data but, rather, to define qualitative conditions that can be tested empirically and from which the functional relationship between the numerals, on the one hand, and the mathematical numbers—that is, the scale values of the sensation scale in question—on the other, can be rigorously (i.e., mathematically) derived.

REFERENCES

American National Standards Institute (1996). *Specification for audiometers* (ANSI S3.6-1996). New York: Author.

Attneave, F. (1962). Perception and related areas. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 4, pp. 619-659). New York: McGraw-Hill.

Banks, W. P., & Coleman, M. J. (1981). Two subjective scales of number. *Perception & Psychophysics*, **29**, 95-105.

Collins, A. A., & Gescheider, G. A. (1989). The measurement of loudness in children and adults by absolute magnitude estimation and cross-modality matching. *Journal of the Acoustical Society of America*, **85**, 2012-2021.

Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics*, **62**, 1505-1511.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, NJ: Erlbaum.

Hellman, R. P. (1976). The growth of loudness at 1000 and 3000 Hz. *Journal of the Acoustical Society of America*, **60**, 672-679.

Jesteadt, W., Wier, C. C., & Green, D. M. (1977). Intensity discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America*, **61**, 169-177.

Luce, R. D. (2001). Reduction invariance and Prelec's weighting functions. *Journal of Mathematical Psychology*, **45**, 167-179.

Luce, R. D. (2002). A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review*, **109**, 520-532.

Luce, R. D. (2004). Symmetric and asymmetric matching of joint presentations. *Psychological Review*, **111**, 446-454.

Merkel, J. (1888). Die Abhängigkeit zwischen Reiz und Empfindung. *Philosophische Studien*, **4**, 541-594.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, **40**, 109-129.

Prelec, D. (1998). The probability weighting function. *Econometrica*, **66**, 497-527.

Richardson, L. F. (1929). Imagery, conation, and cerebral conductance. *Journal of General Psychology*, **2**, 324-352.

Richardson, L. F., & Ross, J. S. (1930). Loudness and telephone current. *Journal of General Psychology*, **3**, 288-306.

Rule, S. J., & Curtis, D. W. (1973). Conjoint scaling of subjective number and weight. *Journal of General Psychology*, **97**, 305-309.

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

Schneider, B., Parker, S., Ostrosky, D., Stein, D., & Kanow, G. (1974). A scale for the psychological magnitude of number. *Perception & Psychophysics*, **16**, 43-46.

Steingrimsson, R. (2002). *Contributions to measuring three psychophysical attributes: Testing behavioral axioms for loudness, response time as an independent variable, and attentional intensity*. Doctoral dissertation, University of California, Irvine.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes—loudness. *American Journal of Psychology*, **69**, 1-25.

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

Teghtsoonian, M., & Teghtsoonian, R. (1971). How repeatable are Stevens's power law exponents for individual subjects? *Perception & Psychophysics*, **10**, 147-149.

Wilcoxon, F., & Wilcox, R. A. (1964). *Some rapid approximate statistical procedures*. Pearl River, NY: Lederle Laboratories.

Zimmer, K., Luce, R. D., & Ellermeier, W. (2001). Testing a new theory of psychophysical scaling: Temporal loudness integration. In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Fechner Day 2001: Proceedings of the seventeenth annual meeting of the International Society for Psychophysics* (pp. 683-688). Lengerich: Pabst.

NOTES

1. The notation used here follows neither Narens (1996) nor Luce (2002) but has been simplified with the goal of facilitating an immediate understanding of the empirically relevant axioms for readers who are not closely familiar with Narens's and Luce's theories. Note that Narens uses boldface letters for the number words, whereas in Luce's notation, $p*$ corresponds to $W(p)$, a strictly increasing weighting function relating numerals and numbers, with $W(0) = 0$, and $W(1) = 1$. The outcome of a trial in a magnitude production task—namely, a sound that is judged as being p times as loud as the standard **X**—is denoted by $p*\mathbf{X}$ in the present article, whereas it reads $(x, \mathbf{p}, t)$ in Narens's diction, with t signifying the standard and x the sound that is produced. In Luce (2002), the outcome of a magnitude production trial is given by $W(p)\psi(x)$, x being the standard and $\psi(x)$ the (strictly increasing) psychophysical function. In Luce's (2002) theory, all stimuli are expressed as intensities above sensation threshold, but as the author notes on pp. 521 and 522, except for stimuli very close to threshold, the distinction between stimuli measured as sound pressure levels and as sensation levels is not behaviorally relevant for the purposes of his axiomatization.

2. In order to avoid potential hearing damage, the maximal sound pressure level generated by the system was set to 94 dB SPL. It was never reached in the course of experimentation.

3. It is worth mentioning that choosing the more conventional significance level $\alpha = .05$ would not have influenced the statistical decision in any of the statistical tests performed.

4. For none of the participants did the sound pressure levels of the adjustments made for a given loudness fraction (e.g., $\frac{1}{6}*\mathbf{X}$) differ significantly between the square-root and the cube-root conditions (Mann–Whitney U test, $\alpha = .1$).

5. R. D. Luce, Larnaca (Cyprus), October 22, 2003.

6. Two participants, K.A. and O.L., took part in both experiments, the data collection for which was more than a year apart. Whereas K.A.'s adjustments did not differ much on these two occasions (i.e., by 2.33 and 6.08 dB, on average, for the successive and the single adjustments, respectively), O.L.'s adjustments were considerably lower when replicated a year later (by 34.76 and 24.85 dB, on average). Since audiometry did not reveal any changes in hearing threshold, and since a context effect (adjusting a fixed stimulus range to the range of fractionation instructions) would have resulted in higher sound pressure levels for the ½* ⅓* **X** and ⅙* **X** adjustments in the second experiment than in the first, there is no obvious explanation for this singular discrepancy.

APPENDIX

The following proof was outlined by R.D. Luce in a personal communication on October 22, 2003, at the annual Fechner Day Meeting of the International Society for Psychophysics, which was held at Larnaca, Cyprus.

Given a stimulus, x, and a fraction to be implemented, p, (x, p) denotes the stimulus that a listener judges to be p times as loud as x; ψ and W are the psychophysical function and the weighting function transforming the numeral p into a (mathematical) number, respectively, with associated function parameters α, β, λ, and μ (Luce, 2002).

In Luce's theory, the following equations hold:

(1) $\psi(x, p) = \psi(x) \cdot W(p)$ (Equation 20, p. 523, in Luce, 2002),

(2) $\psi(x) = \alpha x^{\beta}$ (Equation 39, p. 526),

and

(3) $W(p) = \exp\{-\lambda[-\ln(p)]^{\mu}\} \leftrightarrow W^{-1}(p) = \exp\{-[-\tfrac{1}{\lambda}\ln(p)]^{\frac{1}{\mu}}\}$ (Equation 45, p. 527).

Suppose a lower bound, so that $(x, p) > k \cdot x$ (with $0 < k < 1$ in the case of fractionation judgments). Then it is shown in the following that $(x, p^N) > k^{N^{\mu}} \cdot x$; $N \in \Re$.

Using (1) and (2) yields $\psi(x, p) = \psi(x) \cdot W(p) = \alpha x^{\beta} \cdot W(p)$ and $\psi(k \cdot x) = \alpha k^{\beta} x^{\beta}$.

Thus, $\alpha x^{\beta} \cdot W(p) > \alpha k^{\beta} x^{\beta} \leftrightarrow W(p) > k^{\beta} \leftrightarrow p > W^{-1}(k^{\beta}) \leftrightarrow p^N > [W^{-1}(k^{\beta})]^N$. Now, using (3),

$$
\begin{aligned}
W(p^N) &> W\{[W^{-1}(k^{\beta})]^N\} \\
&> \exp(-\lambda\{-\ln[W^{-1}(k^{\beta})]^N\}^{\mu}) \\
&> \exp(-\lambda \cdot N^{\mu}\{-\ln[W^{-1}(k^{\beta})]^{\mu}\}) \\
&> \exp(-\lambda \cdot N^{\mu}\{[-\tfrac{1}{\lambda}\ln(k^{\beta})]^{1/\mu}\}^{\mu}) \\
&> \exp[N^{\mu} \cdot \ln(k^{\beta})] \\
&> k^{\beta \cdot N^{\mu}}.
\end{aligned}
$$

In sum, $\psi(x, p^N) > \psi(k^{N^{\mu}} \cdot x) \leftrightarrow (x, p^N) > k^{N^{\mu}} \cdot x$.

(Manuscript received April 21, 2004;
revision accepted for publication August 4, 2004.)