Perception & Psychophysics
2000, 62 (8), 1505-1511

Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production

WOLFGANG ELLERMEIER and GÜNTHER FAULHAMMER
University of Regensburg, Regensburg, Germany

Stevens's direct scaling methods rest on the assumption that subjects are capable of reporting or producing *ratios* of sensation magnitudes. Only recently, however, did an axiomatization proposed by Narens (1996) specify necessary conditions for this assumption that may be put to an empirical test. In the present investigation, Narens's central axioms of *commutativity* and *multiplicativity* were evaluated by having subjects produce loudness ratios. It turned out that the adjustments were consistent with the commutativity condition; multiplicativity (the fact that consecutive doubling and tripling of loudness should be equivalent to making the starting intensity six times as loud), however, was violated in a significant number of cases. According to Narens's (1996) axiomatization, this outcome implies that although in principle a ratio scale of loudness exists, the numbers used by subjects to describe sensation ratios may not be taken at face value.

In their comprehensive review of psychophysical methods, Luce and Krumhansl (1988) distinguished two largely unrelated schools of thinking about psychological measurement: that of the *axiomatizers* and that of the *scalers*. The axiomatizers adhere to the traditions of measurement theory (first layed out by Krantz, Luce, Suppes, & Tversky, 1971; recently summarized by Iverson & Luce, 1998, and Narens & Luce, 1986). Their fundamental objective is to formulate qualitative conditions (called *axioms*) that, when satisfied, justify psychological measurement for a given domain and methodology (the *representational problem*) and to specify the scale type attained by such measurement (the *uniqueness problem*). *Scaling*, the actual assignment of numbers to the objects under study, is seen as secondary to solving these two problems, and scale values are never directly assigned by the subjects but, rather, are derived by the investigator, often from simpler qualitative judgments obtained when testing the validity of the axioms.

The *scalers*, by contrast, focus on the problem of obtaining numerical assignments, devise methods to solicit these from observers directly, and are primarily concerned with studying biases or context effects that might jeopardize the validity of the endeavor.

S. S. Stevens's (1956, 1957, 1975) direct scaling paradigm, rooted in earlier attempts to have subjects make statements about ratios of sensation magnitudes (Merkel, 1888; Richardson & Ross, 1930), clearly belongs to the latter category. The pragmatic stance, taken together with the outlook of obtaining metric data of the ratio-scaling type in a highly economic fashion, has made the *new psychophysics* the most successful research program in the realm of scaling (Gescheider, 1997; Marks & Algom, 1998; S. S. Stevens, 1975). The assumption, however, that subjects are capable of reporting (or producing) *ratios* of sensation magnitudes has largely remained untested. Thus, the concept of ratio scaling may hardly be said to have more than face validity (via instructions given to subjects), and attempts to substantiate it within the concept of magnitude scaling have largely been confined to demonstrations of reliability (e.g., Logue, 1976; Teghtsoonian & Teghtsoonian, 1971) and of the transitivity of cross-modality matches (e.g., Collins & Gescheider, 1989; J. C. Stevens, Mack, & S. S. Stevens, 1960). That is ironic, especially since S. S. Stevens (1946, 1951) introduced the concept of scale types (such as ordinal, interval, or ratio), which plays a central role in axiomatic measurement theory.

Nevertheless, efforts to develop an axiomatic foundation for Stevens's ratio-scaling approach have been scarce (Krantz, 1972; Luce, 1959, 1990; Shepard, 1981) and often restricted in applicability (e.g., to cross-modality matching). What has been lacking so far is a rigorous mathematical formulation of the assumptions inherent in S. S. Stevens's approach and a specification of the conditions (axioms) necessary for its success.

Portions of the data were presented at the 14th Annual Meeting of the International Society for Psychophysics, Quebec, August 1998 (supported by DFG Travel Grant EL 129/3-1). This work was partially supported by a research grant from Deutsche Forschungsgemeinschaft (EL 129/2-1), which enabled the first author to spend 3 months in 1996 at the University of California, Irvine, where the project was initiated. Thanks are due Louis Narens, who made valuable suggestions on how to test his axiomatization during a laboratory visit at Regensburg in the summer of 1997. The authors also gratefully acknowledge the comments of Bruce Schneider, Scott Parker, and Lawrence Ward during the review process. Correspondence concerning this article should be addressed to W. Ellermeier, Institut für Psychologie, Universität Regensburg, D-93040 Regensburg, Germany (e-mail: wolfgang.ellermeier@psychologie.uni-regensburg.de).

Such a proposal has recently—50 years after the inception of the new psychophysics—been made by Narens (1996). It is fundamentally different from earlier proposals and distinguishes a *behavioral axiomatization*, relating the observer's use of *numerals* to their numerical representations, from a *cognitive axiomatization*, relating the numerical representation to the unobservable *sensations*. The present application of Narens's (1996) axiom system exclusively refers to the behavioral axiomatization, which spells out the assumptions inherent in Stevens's approach, and takes care to treat the *numerals* uttered by the subject in a magnitude estimation task as distinct from (scientific) *numbers*, of which the subject may or may not have a sound understanding.

Of the axioms formulated by Narens (1996), two (commutativity and multiplicativity) are empirically testable and crucial to the interpretation of subjects' scaling behavior. Since these axioms state certain equivalences, they are most conveniently operationalized in ratio production tasks (Gescheider, 1997, chap. 11), but Narens postulates that they are equally well suited for magnitude estimation or cross-modality matching situations. In Narens's terminology, the axioms read as follows:

> AXIOM 4. (*Commutative property*) If $(x,\mathbf{p},t) \in E$, $(z,\mathbf{q},x) \in E$, $(y,\mathbf{q},t) \in E$ and $(w,\mathbf{p},y) \in E$, then $z = w$,

where the triple $(x,\mathbf{p},t)$ refers to a trial in a ratio production task (taken from the set of all possible ratio productions E)—that is, a subject's making an adjustment x that appears $\mathbf{p}$ times as intense as a standard t—with t,w,x,y, and z referring to physical stimulus intensities and the boldface letters **p** and **q** referring to the numerals used by or given in instructions to the subject. Specifically, if the subjective dimension considered is loudness, Axiom 4 states that doubling the loudness of a reference sound ($\mathbf{p} = \mathbf{2}$), for example, and then tripling the result ($\mathbf{q} = \mathbf{3}$) should generate the same final adjustment as first making the reference sound three times as loud and then doubling the outcome.

If this commutative property holds, along with a number of technical axioms (1–3) making statements about the physical continuum, for example, and about the monotonicity of magnitude productions, a *ratio scale* may be said to exist; the numerals involved, however, may not be interpreted at face value (that is, as scientific numbers). That is the case only, if the following axiom is shown to hold:

> AXIOM 9. (*Multiplicative property*) If $(x,\mathbf{p},t) \in E$, $(z,\mathbf{q},x) \in E$ and $r = qp$, then $(z,\mathbf{r},t) \in E$,

with the same conventions as those specified for Axiom 4, and r, p, and q referring to scientific numbers. Applied to loudness, this means that making a reference six times as loud ($\mathbf{r} = \mathbf{6}$) should produce the same stimulus level as making it three times as loud first ($\mathbf{p} = \mathbf{3}$) and then doubling the resultant ($\mathbf{q} = \mathbf{2}$).

If the multiplicative property holds, in addition to Axiom 4 and a number of assumptions about the "inner psychological measurement structure" that is used to generate responses, the numerals used by the subject may be interpreted as numbers—that is, they may be taken "at face value," as they typically are in a magnitude estimation task.

The goal of the present experiment was to investigate whether the axioms of *commutativity* and *multiplicativity* hold for ratio productions of loudness, the most extensively studied subjective dimension in the magnitude scaling literature. That was done by using the smallest integers applicable—that is, for $\mathbf{p} = \mathbf{2}$ and $\mathbf{q} = \mathbf{3}$, as in the examples previously given. In order to increase the generality of the conclusions, the validity of the axioms was investigated by starting from two absolute sound pressure levels (40 and 55 dB SPL). Furthermore, upon Narens's suggestion, additional adjustments of five times and seven times as loud were included in the experimental design, in order to take a closer look at potential nonmonotonicities in subjects' magnitude productions.

METHOD

Subjects

The authors and 8 students at the University of Regensburg participated in the experiment. This sample had a median age of 25 years (range, 23–41 years) and consisted of 4 male and 6 female subjects. All the subjects had normal hearing at the standard audiometric frequencies, as determined by Békésy tracking. None (except for the authors) had prior knowledge of the hypotheses being investigated. Originally, a total of 13 subjects had been recruited, 3 of which had to be excluded from the experiment after a practice session, since their adjustments consistently tended to exceed the ceiling, set at 95-dB sound pressure level to prevent damage to their hearing.

Stimuli and Apparatus

The stimuli were 1-kHz sinusoids of 500-msec duration, including 10-msec rise and decay ramps. They were computed via a Tucker–Davis Technologies (TDT) signal processor card (model AP2) and were played from a 16-bit digital-to-analog converter (TDT model DD1) at a sampling rate of 50 kHz. After passing through a low-pass filter set at 10 kHz (TDT model FT5), the signal was set to the proper level by means of a programmable attenuator (TDT model PA4), after which it was diotically delivered to the subject seated in a double-walled sound-attenuated chamber via Beyerdynamic DT 48 headphones. The equipment was calibrated by measuring sound pressure levels at the headphones, using an artificial ear (Bruel & Kjær type 4153) and a sound-level meter (Bruel & Kjær type 2610).

Procedure

As is required by the most basic implementation of Narens's (1996) axiomatization conceivable, the subjects had to produce stimuli resulting in two, three, and six times the loudness of a reference tone, as well as to triple the loudness of a prior doubling, and to double the loudness of a prior tripling (see Narens's, 1996, Axioms 4 and 9). In addition, in order to bracket the crucial *six-times* adjustment, instructions to make the tone five times and seven times as loud were given as well. Combined with two different absolute starting levels (40 and 55 dB SPL), this resulted in a total of 14 dif-

ferent types of adjustments, which were randomly intermixed in a block of trials, thus generating sufficient variability and reducing predictability for the subject.

Each trial proceeded as follows: A light-emitting diode mounted on a hand-held response unit signaled to the subject which loudness ratio to produce (e.g., *five times as loud*). After the subject pressed a "ready" button, the standard stimulus kept alternating with the (variable) comparison, which the subject was asked to adjust. The interstimulus interval between standard and comparison was 500 msec, and the subject could adjust the level of the comparison tone during the 2 sec that elapsed before the next pair was presented. That was done by pressing a "−" or a "+" button, which decreased or increased the level of the comparison, respectively. This sequence continued until the subject indicated a satisfactory match by pressing an "o.k." button. A method of adjustment with features borrowed from adaptive procedures was implemented to control the changes in stimulus level: In order to reduce potential biases, the starting intensity of the comparison tone was randomly chosen from a 10-dB interval starting 5 dB above the level of the standard tone. The step size by which the intensity was changed was halved after each *reversal* (the subject's shifting responses from "+" to "−" or vice versa), starting from 4 dB, until the minimal step size of 0.5 dB was reached. The final adjustment was accepted only if the subject had listened to it at least once without altering its intensity.

In addition to a practice session consisting of 2 blocks in which each of the 14 trial types occurred, each subject completed 15 blocks of trials in five sessions, thus producing a total of 15 adjustments of each type.

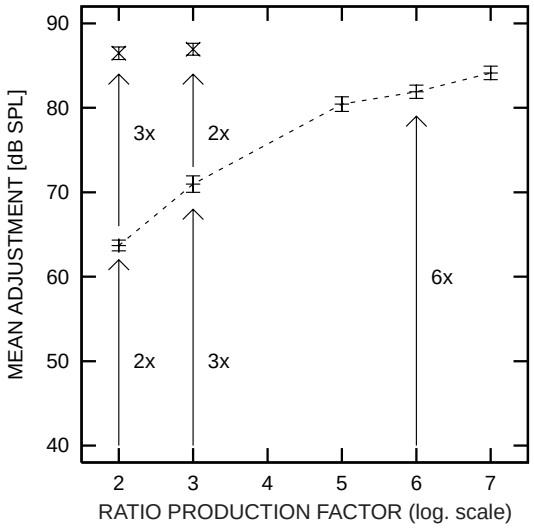

Figure 1. Ratio productions generated by a single subject (E.M.) starting from a standard level of 40 dB SPL. Means plus/minus one standard error of the mean are given on the basis of 15 adjustments per data point. The upward arrows highlight those adjustment sequences crucial to testing the axioms of commutativity and multiplicativity (see the text).

RESULTS

Quality of Measurements Made

A few descriptive statistics might serve as evidence for the quality of the measurements made: The subjects took an average time of 37 sec to produce an individual adjustment, making a median of eight level changes, and arriving at the final stepsize of 0.5 dB in the vast majority of cases. The mean standard error of a set of 15 adjustments was 0.88 dB (see Figure 1), which is on the order of a just-noticeable difference.

The most troublesome problem in conducting the experiment was the potential presence of ceiling effects. Particularly, successively doubling and tripling the loudness of the 55-dB tone occasionally led subjects to reach the 95-dB ceiling implemented for their protection. Note, however, that this is to be expected, given that the *sone* function predicts a loudness ratio of only 1:16 (a doubling with every 10-dB increase in level) for the range from 55 to 95 dB SPL. The fact that individual loudness functions deviate considerably from this average value (see, e.g., Algom & Marks, 1984) and that magnitude production slopes tend to be steeper than those obtained in magnitude estimation (*regression bias*; S. S. Stevens, 1975, chap. 9; S. S. Stevens & Greenbaum, 1966) makes it even more likely that subjects may not be able to make the desired adjustments within the stimulus range chosen. Fortunately, that was the case for only 1 subject (M.H.), whose data collected with the higher starting level (55 dB SPL) were discarded. The remaining data were carefully checked with respect to two criteria: (1) more than three excursions of the adjustment track's hitting the ceiling of 95 dB SPL, and (2) a final value within 2 dB of that ceiling. Only 22 out of 1,995 adjustments—that is, 1.1%—met these criteria. Therefore, no further data selection was made, particularly since the nonparametric tests chosen to evaluate the axioms are well suited to handle this residual indeterminacy.

Testing Narens's (1996) Axioms

An illustrative example. The axioms of *commutativity* and *multiplicativity* were tested individually for each subject and separately for the two starting levels. Figure 1 shows the adjustments made by a subject (E.M.) typical of the present sample. First of all, it is evident that the instructions to make the comparison tone two, three, five, six, or seven times as loud as the 40-dB standard have distinguishable and monotonic effects. Furthermore, commutativity of successive doublings and triplings in loudness appears to hold, as is evident in the two sets of arrows in the left of the figure converging onto nearly identical sound pressure levels (86.47 dB SPL for 2×3 vs. 86.93 dB for 3×2×). A Mann–Whitney U test shows the two sets of 15 adjustments each that generate these two means to be statistically indistinguishable [$z(U) = 0.10$, n.s.].

Multiplicativity, on the other hand, does not seem to hold: The average adjustment made in response to a *six times as loud* instruction is 81.9 dB SPL (rightmost upward arrow in Figure 1), falling short of successively doubling and tripling by almost 5 dB. The difference between the 6 × adjustments and the pooled 2×3× and 3×2× adjustments is statistically significant [$z(U) = 4.11$, $p < .05$].

Table 1
Empirical Validity of Narens's (1996) Axioms

	Commutativity		Multiplicativity	
Subject	40 dB SPL	55 dB SPL	40 dB SPL	55 dB SPL
E.M.	−0.10	−1.62	**+4.11**	**+4.56**
G.F.*	+1.19	+0.67	**+4.15**	**+3.27**
I.K.	−0.71	−1.14	**+5.10**	**+5.13**
J.G.	+1.56	−0.54	**+3.50**	**+2.40**
M.B.	+1.79	+0.13	**+4.46**	**+4.80**
M.H.	**−2.47**	—	**p.a.v.**	—
P.S.	+0.67	+0.40	**+4.37**	**+3.62**
S.G.	−0.79	+0.25	+1.67	**+2.45**
T.P.	−0.81	+0.13	**+4.11**	**+4.11**
W.E.	+0.56	**−2.10**	**−3.07**	**p.a.v.**

Note—The entries are *z* scores for the outcome of Mann–Whitney *U* tests computed separately for each subject and for each of the two starting levels (40 and 55 dB SPL). Values of $|z| > 1.96$ indicate violations of a given axiom (two-tailed test, $\alpha = .05$) and are printed in boldface. The sign of the *z* score indicates the direction of the effect (see the text). The table entry **p.a.v.** means the prerequisite axiom (of commutativity) is violated. *Subject G.F. was run with standard levels of 50 and 60 dB SPL.

Validity of Narens's (1996) axioms in the sample. Table 1 shows that the behavior of this particular subject (E.M.) is typical of the sample at large. The table reports the outcome of individual nonparametric tests (Mann–Whitney *U* tests) for axiom violations—that is, for deviations from the equivalences required by Narens's (1996) Axioms 4 and 9. Nonparametric tests were chosen in order to deal with potential problems arising from (1) indeterminate values near the ceiling of the adjustment range (see the Method section) or (2) inhomogeneity of variance, which might be expected when successive productions (e.g., 2×3×) bearing the risk of error propagation are compared with direct productions (e.g., 6×). Both problems turned out to be of minor importance, however (see the location and precision of average adjustments given in Figure 2). The table entries are *z* scores computed for the *U* statistic employed. Those *z* scores having absolute values greater than 1.96 (two-tailed test, $\alpha = 0.05$) reflect systematic discrepancies between the sets of adjustments that should coincide according to Narens's (1996) axioms. The sign of a given *z* score indicates the direction of the effect; positive *z* scores in the last two data columns, for example, identify those cases, in which consecutive doubling and tripling *exceeds* the adjustment of *six times as loud*.

The picture emerging is unequivocal: Where commutativity holds in 17 out of 19 cases, multiplicativity is violated in 16 of the 17 cases that had passed the prerequisite test of commutativity.

This pattern of outcomes may be transformed into an overall statistical statement by consulting the binomial distribution. For any given subject, the probability of violating multiplicativity at both of the two standard levels by chance alone (given $\alpha = .05$) is $p = .0025$. Encountering seven or more of these double violations in the 8 subjects for which data at both levels are available (see Table 1) is highly unlikely ($p \approx 6.10 \cdot 10^{-19}$), justifying the conclusion that we are dealing with a systematic effect. For commutativity, on the other hand, the probability of obtaining a violation at only one of the two standard levels (given $\alpha = .05$) by chance alone is $p = .095$, or roughly 1 in 10. Since only 1 of the 9 subjects for whom data were available at both levels showed such a violation and the rest showed none, the outcome is as would be expected if commutativity held.

Violations of rank order. Since the picture emerging is so clear-cut and consistent across individuals, it is feasible to inspect mean adjustments as well. When subjects G.F. (for whom different decibel levels were used), M.H., and W.E. (both of whom showed violations of commutativity) are excluded, the remaining 7 subjects' deviations from multiplicativity all point in the same direction. That is, the successive loudness doublings and triplings (plotted next to the 6× adjustment in Figure 2) overshoot the single adjustment of six times as loud and, on the average, do so by approximately 5 dB.

Furthermore, as becomes evident when the data are plotted in double-logarithmic coordinates, as in Figure 2, the present ratio productions are well described by psychophysical power functions with exponents of 0.76 (40-dB standard; lower curve) and 1.35 (55-dB standard; upper curve in Figure 2). To researchers familiar with loudness scaling, these exponents may seem unusually large; note, however, that S. S. Stevens and Greenbaum (1966) found production exponents to exceed those obtained in magnitude estimation by factors reaching two or more. The fact that the two curves plotted in Figure 2 have different slopes argues against a unitary power function for loud-

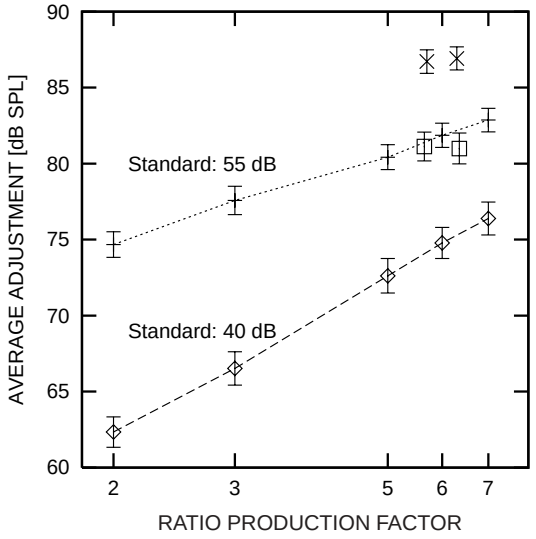

Figure 2. Mean magnitude productions generated by 7 subjects from two starting levels in response to the ratio instructions given on the abscissa. Final adjustments when the subjects were asked to first double and then triple the standard intensity (or vice versa) are plotted next to the six-times adjustment. Squares denote final adjustments starting from 40 dB, crosses final adjustments starting from 55 dB SPL.

ness, although it is consistent with the recent finding of a steepening[1] of loudness functions at higher sound pressure levels (Buus, Florentine, & Poulsen, 1997).

The most remarkable feature of the data depicted in Figure 2 is, however, that the average 2×3× adjustments exceed the 7× adjustments by 3.5–4.0 dB, thus producing a reversal of rank order in the outcomes of ratio instructions. This reversal is significant in 13 out of the 17 tests in which multiplicativity was analyzed (*U* tests, one-tailed, $p < .05$).

DISCUSSION

The present investigation of two axioms fundamental to S.S. Stevens's ratio-scaling approach found *commutativity* to hold for magnitude productions of loudness. That is, subjects generate different increments in intensity when asked to double or triple loudness, respectively, and they can concatenate these operations to converge on the same final outcome, no matter in which order the operations are applied. This type of consistency, according to Narens's (1996) axiomatization, implies that, in principle, the subject uses a ratio scale. The failure of finding *multiplicativity* to hold, however, precludes interpreting the numerals used by the subject as scientific numbers.

Before interpreting the fact, however, that 2×3× and 6× adjustments diverge so dramatically, procedural artifacts inherent in the ratio production sequence suggested by Narens's (1996) axiomatization have to be ruled out. We were particularly concerned about the fact that a single adjustment (of six times as loud) is compared with two consecutive adjustments (first three times, then twice as loud, or vice versa). If the adjustments were subject to an additive bias (such as a time-order error; see Hellström, 1978), that might affect the 2×3× adjustments twice, thus explaining the fact that they tend to overshoot the single 6× adjustment. In order to investigate this possibility, 2 subjects (E.M. and M.B., who had passed the commutativity test) were asked to perform an additional experiment in which both of the crucial conditions involved in testing commutativity were consecutive: That is, a 1×6× condition, in which subjects first had to produce a loudness match (Factor **1**) between standard and comparison and the resulting adjustments were later made six times as loud, was compared with a 2×3× obtained as in the main experiment. In addition, starting levels for the comparison tone were chosen according to two strategies: (1) between 5 and 15 dB above the 55-dB standard, and (2) from a 10-dB interval surrounding the standard level. The results are given in Table 2: There is no evidence that the choice of starting level produces an additive bias: Both sets of productions (designated as "elevated" and "equal" comparison levels in Table 2) generated nearly indistinguishable results, differing only in unsystematic ways. What is more important, however, is that concatenating adjustments did not seem to matter for either subject: In no case were the single 6× adjustments significantly different from the 1×6× adjustments, and the violations of multiplicativity remained significant with the new consecutive procedure (see the last column of Table 2).

Clearly, a match (a factor of **1.0**) is a special case of ratio production, and it might be more revealing to look at 2×6× vs. 3×4× adjustments, for example. For the present choice of stimuli, however, a factor of 12 would have exceeded an acceptable loudness limit for most of our subjects and might have generated artifactual ceiling effects.

Given that the control experiment rules out procedural artifacts, such as the propagation of errors, to account for the result, one may turn to another criticism often raised against axiomatic evaluations of conventional methodology—that is, that they exploit small lapses of precision revealed by overly powerful adaptive methods that may be ignored for practical purposes. That criticism does not apply to the present investigation for three reasons: First of all, note that we show one axiom to be violated (multiplicativity), whereas another important one holds (commutativity), applying identical statistical power in both cases. Second, we are dealing with big and systematic effects: All but one violation point in the same direction, and their magnitude is large both in terms of the physical dimension studied (averaging 5 dB; see Figure 2) and in terms of discriminability (roughly five just-noticeable differences). Finally, detecting reversals of order, such as the 2×3× adjustments exceeding those in response to the 7× instruction, reveals qualitative, rather than mere quantitative, inconsistencies with respect to multiplicativity that may not be brushed aside.

Since we report a first investigation employing a new axiomatic approach to ratio scaling, two questions may

Table 2
Control Experiment on Single Versus Consecutive Adjustments

					Mann–Whitney *U* Tests			
		Adjustments [dB SPL]			2×3× > 6×		2×3× > 1×6×	
Condition	Subject	2×3×	6×	1×6×	$z(U)$	p	$z(U)$	p
Elevated comp.	E.M.	87.20	84.37	84.67	**+2.63**	.004	**+2.49**	.006
	M.B.	88.80	83.60	85.00	**+2.17**	.015	**+1.75**	.040
Equal-level comp.	E.M.	86.43	83.00	83.60	**+2.84**	.007	**+2.58**	.005
	M.B.	90.30	86.13	87.73	**+2.41**	.008	**+1.95**	.025

Note—Each adjustment represents a mean of 15 magnitude productions. Based on the outcome of the main experiment, statistical tests are unidirectional, as is indicated in the column headings. Thus, values of $z > 1.64$ indicate a violation of multiplicativity and are printed in boldface.

be asked with regard to the generality of the present result. They concern (1) generality across perceptual dimensions and (2) generality across different varieties of magnitude scaling methodology available. Regarding the latter point, there are no empirical studies as yet. Narens's (1996) article, however, contains a section on "generalized ratio magnitude estimation" interpreting his axiomatization as extending to conventional magnitude estimation and cross-modality matching tasks. In our view, however, it is difficult to see how the axiomatization applies to *one-stimulus*, *one-response situations*, such as magnitude estimation without standard or modulus, or to free cross-modality matches. Narens's conceptualization seems, by the trial structure $(x,\mathbf{p},t)$, basically relational in nature, as was Krantz's (1972) and Shepard's (1981). It remains to be seen whether investigators will come up with operationalizations of *absolute magnitude estimation* or *cross-modality matching* in terms of Narens's axiomatization.

Evidence for the generality of the present findings across sensory modalities has been produced in a parallel investigation by Peißner (1999), using brightness productions of test circles on a computer screen. Since the brightness and loudness studies were planned collaboratively, they agreed in basic rationale, except for methodological modifications necessitated by the modality in question. The only significant difference was the data analysis strategy employed: Rather than using standard nonparametric tests in order to detect violations of the equivalences required by Narens's axiomatization, Peißner employed a variety of discriminant analysis to check how well the pooled adjustments may be "blindly" classified as belonging to the ascending set of ratio instructions (*monotonicity*), whether the 2×3× adjustments are interchangeable with the 3×2× responses (*commutativity*), and whether they may be substituted for the 6× adjustments (*multiplicativity*). Despite these differences in methodology, the pattern emerging is the same as that in the present investigation: Whereas commutativity is satisfied in the majority of subjects (8 out of 11), multiplicativity fails for most of them (9 of the 10 considered).

The empirical outcome of the present auditory and of the parallel visual study (Peißner, 1999) is the one Narens predicted in his theoretical article: "I suspect that the multiplicative property would fail empirically for most of the kinds of situations where magnitude estimation is employed" (Narens, 1996, p. 110). How may that outcome be stated in Narens' terminology? The situation we are left with is characterized by the following relationship (Narens, 1996, Equation 1):

$$\varphi(x) = f(\mathbf{p}) \cdot \varphi(t). \quad (1)$$

While numerical representations on a ratio scale φ exist both for the standard t and for the ensuing production x, the numeral $\mathbf{p}$ used in the ratio instructions may not be taken as the factor linking these two representations. Rather, a transformation function $f(\mathbf{p})$ must be found to convert numerals into numbers—that is, to reveal the underlying scale values. Given the failure of multiplicativity, this function is not the identity function $f(\mathbf{p}) = p$ that is typically assumed when analyzing magnitude scaling data.

How may that function be characterized, given the outcome of the present experiment? First of all, it cannot be a power function of the type $f(\mathbf{p}) = p^{\beta}$, since such a function would satisfy multiplicativity [$f(\mathbf{p}) \cdot f(\mathbf{q}) = f(\mathbf{p} \cdot \mathbf{q})$]. A simple multiplicative constant [$f(\mathbf{p}) = \alpha \cdot p$] is conceivable, and in fact, assuming $\alpha = 1.5$ would make the 2×3× adjustments roughly coincide with a ratio of 9×, as is suggested by extrapolating from the bottom curve in Figure 2; however, such a function would conflict with Narens's (1996) Axiom 2.3 that requires the identity function[2] for loudness matches $(t,\mathbf{1},t)$ and, hence, $f(\mathbf{1}) = 1$.

Clearly, a far larger number of different ratio instructions and separate analyses for individual subjects would be needed to determine the function $f(\mathbf{p})$ with some degree of confidence, but it is worth noting that the type of function qualifying is already quite constrained by the outcome of the present experiment.

It should be noted that the conclusions from the present axiomatic study mirror the scepticism toward subjects' use of numbers expressed in the scaling literature itself, which is reflected in the distinction of *input* and *output* functions in multistage models of psychophysical judgment (Attneave, 1962; Rule & Curtis, 1978) or in attempts to corroborate results from magnitude scaling experiments by validating them against *nonmetric* scaling techniques derived from paired comparisons (e.g., Parker & Schneider, 1994; Schneider, 1980; Schneider & Cohen, 1997; Schneider, Parker, & Stein, 1974). The encouraging conclusion from the present axiomatic treatment of magnitude scaling is that overt magnitude productions are consistent with the existence of an underlying ratio scale of loudness. The practice, however, of deriving that scale from numerical estimates directly—as is implicit in the *sone scale* of loudness, for example—must be seen as highly problematic in the light of the present findings.

Further studies based on Narens's (1996) axiomatization will have to determine (1) the function converting *numerals* into *numbers* or (2) whether subjects are capable of distinguishing between ratios and differences of sensations at all. The latter issue has been investigated extensively (for reviews, see Birnbaum, 1982, 1990), but with mixed results. Whereas for most continua (such as loudness or heaviness), observers seem to employ a single operation whether instructed to judge differences or ratios (e.g., Mellers, Davis, & Birnbaum, 1984; Schneider, Parker, Farrell, & Kanow, 1976), there are occasional reports of different operations' being employed under these instructions (e.g., Birnbaum, Anderson, & Hynan, 1989; Popper, Parker, & Galanter, 1986). Narens's conceptualization, especially the inherent distinction between numerals and numbers, may shed new light on the issue of comparing direct ratio estimation with difference estima-

tion both by providing a thorough theoretical basis (the foundations of which are developed in Narens, 1997) and by suggesting new methodologies to tackle the problem.

REFERENCES

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

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.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychological measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1990). Scale convergence and psychophysical laws. In H.-G. Geissler (Ed.), *Psychophysical exploration of mental structures* (pp. 49-57). Toronto: Hogrefe & Huber.

Birnbaum, M. H., Anderson, C., & Hynan, L. (1989). Two operations for "ratios" and "differences" of distances on the mental map. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 785-796.

Buus, S., Florentine, M., & Poulsen, T. (1997). Temporal integration of loudness, loudness discrimination, and the form of the loudness function. *Journal of the Acoustical Society of America*, **101**, 669-680.

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.

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

Hellström, Å. (1978). Factors producing and factors not producing time errors: An experiment with loudness comparisons. *Perception & Psychophysics*, **23**, 433-444.

Iverson, G., & Luce, R. D. (1998). The representational measurement approach to psychophysical and judgmental problems. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 1-79). San Diego: Academic Press.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

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

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, **66**, 81-95.

Luce, R. D. (1990). "On the possible psychophysical laws" revisited: Remarks on cross-modal matching. *Psychological Review*, **97**, 66-77.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, L. Gardner, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology: Vol. 1. Perception and motivation* (2nd ed., pp. 9-74). New York: Wiley.

Marks, L. E., & Algom, D. (1998). Psychophysical scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 81-178). San Diego: Academic Press.

Mellers, B. A., Davis, D. M., & Birnbaum, M. A. (1984). Weight of evidence supports one operation for "ratios" and "differences" of heaviness. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 216-230.

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.

Narens, L. (1997). On subjective intensity and its measurement. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 189-205). Mahwah, NJ: Erlbaum.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, **99**, 166-180.

Parker, S., & Schneider, B. (1994). The stimulus range effect: Evidence for top-down control of sensory intensity in audition. *Perception & Psychophysics*, **56**, 1-11.

Peißner, M. (1999). *Experimente zur direkten Skalierbarkeit von gesehenen Helligkeiten* [Experiments on the direct scalability of perceived brightness]. Unpublished master's thesis, Universität Regensburg.

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

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

Rule, S. J., & Curtis, D. W. (1978). 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. A. (1980). A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika*, **45**, 357-372.

Schneider, B. A., & Cohen, A. J. (1997). Binaural additivity of loudness in children and adults. *Perception & Psychophysics*, **59**, 655-664.

Schneider, B. [A.], Parker, S., Farrell, G., & Kanow, G. (1976). The perceptual basis of loudness ratio judgments. *Perception & Psychophysics*, **19**, 309-320.

Schneider, B. [A.], Parker, S., & Stein, D. (1974). The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology*, **11**, 259-273.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology*, **24**, 21-57.

Stevens, J. C., Mack, J. D., & Stevens, S. S. (1960). Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology*, **59**, 60-67.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677-680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Stevens, S. S. (1956). The direct estimation of sensory magnitude-loudness. *American Journal of Psychology*, **69**, 1-15.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153-181.

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

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics*, **1**, 439-446.

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

NOTES

1. Note that in Figure 2, the axes are reversed with respect to a conventional magnitude estimation plot: Sound pressure levels are plotted on the ordinate, so the (upper) 55-dB curve is in fact *steeper* over dB SPL.

2. Interestingly, the loudness matches made by subjects E.M. and M.B. in the control experiment tended to systematically overshoot the standard level by 2–3 dB, indicating a violation of Narens's Axiom 2.3. We tend to attribute this effect to a response bias, however, owing to interleaving matches (1×) with trials (e.g., 6×) that require turning up the intensity considerably.

(Manuscript received December 10, 1998;
revision accepted for publication March 6, 2000.)