Atten Percept Psychophys (2015) 77:1767–1780
DOI 10.3758/s13414-015-0846-0

Octuplicate this interval! Axiomatic examination of the ratio properties of duration perception

Jana Birkenbusch · Wolfgang Ellermeier · Florian Kattner

Published online: 27 March 2015

Abstract The relationship between the physical intensity of a stimulus and its perceived magnitude can be described by Stevens' power law (Stevens, *American Journal of Psychology, 69*(1), 1–15, 1956), i.e., a power function with an exponent depending on the sensory modality studied. Direct scaling methods used to determine the power function exponent are based on the assumption that subjects are capable of processing ratios of magnitudes. The present experiments investigate whether this assumption holds for duration perception by empirically testing (Narens, *Journal of Mathematical Psychology, 40*(2), 109–129, 1996) fundamental axioms of monotonicity, commutativity, and multiplicativity. To determine whether the exponent can be interpreted in a meaningful way, i.e., whether it is invariant under changes of the reference stimulus, two further axioms, invertibility and weak multiplicativity (Augustin, *Acta Psychologica*, *128* (1), 176–185, 2008) are evaluated. $N = 25$ participants were required to adjust the duration of a comparison tone to specific ratios of different standard durations in two experiments. In accordance with previous findings for other sensory continua, monotonicity held for the duration adjustments of most participants. Significant violations of the commutativity axiom were found in 12.5 % of all pertinent tests, whereas multiplicativity was violated in 32 % of such tests. The axioms of weak multiplicativity and invertibility, however, were violated in over 50 % of the tests. These results indicate that even though a ratio scale for perceived duration exists, the numbers as used by the participants cannot always be taken at face value and that even though power functions fit the data quite well, the exponent depends on the size of the standard and therefore cannot always be interpreted in a meaningful way.

Keywords Direct scaling · Axiomatic evaluation · Ratio production · Duration perception

J. Birkenbusch (✉) · W. Ellermeier · F. Kattner
Institut für Psychologie, Alexanderstrasse 10,
Technische Universität Darmstadt, 64283 Darmstadt Germany
e-mail: birkenbusch@psychologie.tu-darmstadt.de

Introduction

Specifying the relationship between physical time and perceived duration has been explored in many facets in psychophysics. Particularly when duration perception is compared with other sensory modalities, Stevens' power law is invoked. Employing it implies two related, and fundamental questions: First, whether perceived duration satisfies the condition of ratio scalability and second, whether the power law parameters obtained in duration scaling experiments remain unaffected by certain characteristics of the task. This study examines these questions by testing the validity of a number of pertinent axioms from representational measurement theory.

The relationship between the physical intensity of a stimulus and its perceived magnitude can be described by Stevens' power law (1946, 1956), which is formulated as:

$$\varphi(t) = \alpha t^{\beta}, t > 0. \quad (1)$$

That is, the perceived magnitude of a stimulus t is described by a power function αt^{β}. Whereas the parameter α is a proportionality factor depending on the units used, the exponent β depends on the sensory modality. If the value of β is > 1, the perceived magnitude of the stimulus grows faster than the intensity of the physical stimulus. If β is < 1, the increments in perceived stimulus magnitude become smaller with increasing physical stimulus intensity. In the

case of $\beta = 1$, there is a directly proportional relationship between physical and perceived stimuli, i.e., the relationship can be described by a simple linear function.

Physical time and its perceived duration were also found to be related by a power function (Stevens and Galanter, 1957; Allan, 1979). The power function was fitted in several experiments applying different scaling methods (Eisler, 1975), among them ratings and magnitude estimation with and without a standard (Bobko et al., 1977). These approaches yielded exponents ranging from 0.44 to 1.87 (Kornbrot et al., 2013), with an average exponent of 0.90 most suitably describing the relationship between physical and perceived duration (Eisler, 1976).

Established methods to determine the exponent of Stevens' psychophysical function are scaling procedures, in which participants are asked to produce correspondences between the perceived intensity of stimuli and numerical values consistent with the instruction. Stevens (1956) described two direct scaling methods, which are called magnitude estimation and magnitude production.

Though Stevens, in his later writings (e.g., Stevens, 1975) expressed a preference for using these methods without any constraints such as fixed standards or pre-assigned numerical values, their earliest applications were implemented in a similar manner as the classical methods to measure sensory thresholds, that is they used a fixed stimulus, the *standard*, and a variable stimulus called the *comparison*. These versions of magnitude estimation and production have later been termed 'ratio estimation' and 'ratio production', respectively (Gescheider, 1997).

There are two implicit assumptions fundamental to these direct scaling procedures: It is assumed that the participants are able to estimate or to produce perceived intensities on a ratio-scale level and, furthermore, that the numerals the participants use to describe their sensations may be treated like rational numbers in mathematics and therefore can be taken at face value.

Narens (1996) may be credited with making these assumptions explicit—never actually tested by Stevens or his followers—and formulated mathematical axioms providing a possibility to validate them. He distinguishes between behavioral and cognitive axioms: The untestable cognitive axioms describe the relationship between the participant's unobservable sensation of a stimulus' intensity and its numerical representation. The behavioral axioms characterize the participant's behavior in a scaling experiment and relate their numerical representation to the number words used to describe the stimulus' intensity. In contrast to the cognitive axioms, the behavioral axioms are empirically testable.

The behavioral axioms crucial for the assumption that participants are capable of estimating or producing ratios of stimulus intensities are monotonicity, commutativity, and multiplicativity. Their validity can be evaluated by analyzing data collected in magnitude or ratio production experiments (Luce, 2002). In the latter, when applied to the psychophysics of duration, the participant is instructed to adjust the duration of a comparison stimulus (such as w, x, y, z in the following), of the ratio of $\mathbf{p}$, $\mathbf{q}$ or $\mathbf{r}$ of the perceived duration of the standard stimulus t: The notation $(x, \mathbf{p}, t)$ represents a participant's adjustment x, which is perceived to last $\mathbf{p}$ times as long as the standard interval t, with the boldface letter referring to the number word used in the magnitude production instructions.

First of all, besides a number of technical axioms concerning the continuity of the physical stimulus values, the axiom of monotonicity (Augustin, 2008; Axiom 3.1 in Narens, 1996), also known as *ordering*, has to be tested. It is formulated as:

$$\text{If } (x, \mathbf{p}, t) \in E \text{ and } (y, \mathbf{q}, t) \in E, \text{ then } p > q \Leftrightarrow x \succ y. \tag{2}$$

This means, if x has been adjusted to appear $\mathbf{p}$ times as long ($\times\mathbf{p}$, in the following) as the standard t and another adjustment y is $\mathbf{q}$ times as long ($\times\mathbf{q}$, in the following) as the standard, and $\mathbf{p}$ is greater than $\mathbf{q}$, then the adjusted duration x must be longer than the duration y. According to Narens' (1996) theory, if the axiom of monotonicity holds, it can be assumed that the perception of stimuli of the investigated modality occurs on a sensory continuum. It is a necessary condition not only for the subsequently elaborated axioms of commutativity and multiplicativity, but also fundamental to any scaling at all, because even the categories of an ordinal scale can be arranged in an ascending or descending (and therefore monotonic) order. Furthermore, the axiom of commutativity can be evaluated, which is formulated as:

$$\text{If } (x, \mathbf{p}, t) \in E, (z, \mathbf{q}, x) \in E, (y, \mathbf{q}, t) \in E, \text{ and } (w, \mathbf{p}, y) \in E, \text{ then } z = w. \tag{3}$$

In other words, commutativity holds, if the stimulus duration resulting from a successive production sequence $\times\mathbf{p}\times\mathbf{q}$ is equal to the stimulus duration resulting from successive adjustments with interchanged ratio production factors $\times\mathbf{q} \times \mathbf{p}$. For example, doubling the duration of a standard tone and then tripling the outcome should result in the same final duration as tripling the standard duration first and then doubling the result. Narens showed that if the axiom of commutativity holds, it can be assumed that the participant perceives stimulus magnitudes of the investigated modality on ratio scale level. But even if a ratio scale of perception does exist, there is no evidence that the scale values used by the participants can be interpreted as scientific numbers.

To show the latter, the axiom of multiplicativity has to be evaluated, which is formulated as:

$$\text{If } (x, \mathbf{p}, t) \in E, (z, \mathbf{q}, x) \text{ and } r = qp, \text{ then } (z, \mathbf{r}, t) \in E. \quad (4)$$

In other words, the multiplicativity property holds, if the stimulus duration resulting from the successive adjustments $\times\mathbf{p}\times\mathbf{q}$ is equal to the stimulus duration resulting from a single adjustment $\times\mathbf{r}$ with r being the mathematical product of p and q. For example, doubling the duration of a standard tone and then tripling this adjustment should result in the same final duration as making the standard six times as long in a single adjustment. If the axiom of multiplicativity holds, the numerals as used by the participants to describe the perceived stimulus magnitudes can be taken at face value.

During the last decade, the axiomatic approach to magnitude scaling pioneered by Narens (1996) has been extended by Luce and colleagues (Luce, 2002, 2008; Luce et al., 2010). One recent interpretation concerning the axiom of multiplicativity argues, that a veridical interpretation of numbers and thus the validity of multiplicativity is not mandatory for direct ratio scaling: If the axiom of commutativity is satisfied, thus implying ratio scalability for the modality studied, it may be said that the participants interpret the numbers as some ratio, though not the exact ratio stated in the instructions.

The axiomatic framework has been applied to a number of psychophysical dimensions such as loudness (Ellermeier and Faulhammer, 2000; Steingrimsson and Luce, 2005a, b; Zimmer, 2005), area (Augustin and Maier 2008), brightness (Steingrimsson, 2011; Steingrimsson et al., 2012), and, most recently, pitch (Kattner and Ellermeier, 2014). Duration perception, however, has not been studied in this axiomatic manner.

Therefore, the aim of the first experiment was to investigate whether the fundamental axioms of Narens' theory hold for duration perception, i.e., whether participants are capable of processing durations on a ratio scale. This was tested in a ratio production experiment in which participants were required to adjust the duration of a comparison tone to specific positive integer ratios of two different standard durations ($t_1 = 100$ ms, $t_2 = 400$ ms).

The experiment employed a method that is typical for axiomatic testing requiring the participant to adjust the duration of the comparison interval in an iterative fashion until it subjectively matches with the desired ratio. In contrast to one-shot estimations (e.g., "Turn the sound off as soon as it is p times as long."), which seem to be less cumbersome, this procedure does not introduce a bias due to motor latency. Furthermore, the initial duration of the comparison was randomly chosen to fall above and below the estimated 'target duration' for the purpose of counterbalancing trials in which the participants had to shorten or lengthen the comparison tone.

In the second experiment, two further axioms, weak multiplicativity and invertibility (Augustin, 2008), were tested to provide evidence for the psychological meaningfulness of scaling perceived duration, i.e., whether the size of the power law exponent for duration perception remains unaffected by the size of the standard used in ratio production. Again, participants had to adjust the duration of auditory intervals to a certain ratio with respect to a standard tone ($t_3 = 600$ ms). This time, fractions as well as integers were used as ratio production factors.

Experiment 1

Method

Participants

Ten participants took part in the experiment. The sample consisted of four female and six male participants with a median age of 24 years ranging from 21 to 56 years. They did not have any prior knowledge of the hypotheses being tested. The experiment was conducted individually in a double-walled sound-attenuated listening chamber (IAC).

Stimuli and apparatus

All stimuli were sine waves of the same frequency of 440 Hz (A4 standard pitch) converted with a sampling rate of 44.1 kHz, and with 16-bit resolution. Their duration varied as a result of the protocol and contained 10-ms cosine-shaped rise-and-decay ramps to avoid unwanted switching transients. The standards were of fixed durations of 100 and 400 ms or of individual duration generated according to the adjustments of the participants. The comparison stimuli varied accordingly; their initial length was randomly chosen between one and ten times the duration of the corresponding standard. The tones were preset to a comfortable sound pressure level of 65 dB SPL. After passing through a headphone amplifier (Behringer HA 800 Powerplay PRO 8), the tones were presented diotically via headphones (Beyerdynamics DT 990 PRO). The experiment was programmed in MATLAB using the PsychToolbox-package by Brainard (1997) and Pelli (1997).

Procedure

In the first time-production experiment, the participants had to complete 264 trials altogether. They were divided into four identical test sessions taking place on different days. Each session was composed of three blocks of 22 trials,

resulting in a total of 66 trials, respectively. After the completion of a block, the participants were allowed to take a short break. The recording of the data started after the participants had become familiar with the task during three practice trials at the beginning of each session.

Each trial consisted of two duration intervals marked by continuous tones, which were presented successively. The first tone, or standard, was of fixed duration, either 100 or 400 ms, while the second tone, or comparison, was of variable starting duration and could be adjusted by the participants. The tones were separated by a fixed silent inter-stimulus interval of 500 ms. During the presentation of both tones, a yellow numeral $\mathbf{p}$ ($\mathbf{p} = 1, 2, 3, 4, 6, 8$) was displayed in the upper part of the screen, which was the instruction for the participant to adjust the duration of the second tone so that it was perceived to be $\mathbf{p}$-times as long as the first tone. The adjustments could be made by pressing either the left cursor key for decreasing or the right cursor key for increasing the duration of the comparison tone. The steps for incrementing/decrementing duration were $\frac{1}{20}$ of the standard interval, that is 5 ms for the standard of 100 ms and 20 ms for the standard of 400 ms. To increase step size, participants could press the shift key together with the cursor key resulting in steps being ten times as long as the original steps, that is 50 ms or 200 ms, respectively.

The participants were asked to adjust the duration of the comparison tone step by step, i.e., after each key press response, the current standard and the altered comparison were replayed and the instruction was presented again. The participants were instructed to adjust the comparison until they were satisfied with the result and to eventually press the enter key to register the final value. The next trial started after an inter-trial interval of 2,000 ms. There was no time restriction to performing the task.

In each of the blocks, the standards of $t_1 = 100$ and $t_2 = 400$ ms were combined with the ratio production factors $\mathbf{p} = 1, 2, 3, 4, 6$ and 8. These trials are called *basic trials* and their outcomes are primarily used to test monotonicity. The testing of commutativity and multiplicativity is based on the outcomes of so-called *successive trials*, in which the individual adjustments produced by the participants in the basic trials were used as standards. They were combined with the ratio production factors $\mathbf{q} = 2, 3$, and 4. Each type of adjustment was made 12 times, $i = 12$. In the following, the basic adjustments are indicated by $(x_i, \mathbf{p}, t)$. As an example, $(x_3, \mathbf{2}, 100)$ is the third ($i = 3$) adjustment of a trial with a ratio production factor $\mathbf{p} = 2$ and a standard stimulus $t = 100$ ms.

In the successive trials, for each participant, the individual basic adjustments of each $(x_i, \mathbf{p}, 100)$ and $(x_i, \mathbf{p}, 400)$ were used as standard stimuli. More precisely, the new standards $(x_i, \mathbf{2}, 100)$ and $(x_i, \mathbf{2}, 400)$, derived from a basic doubling trial, had to be made $\mathbf{q} = 2, 3$, and four times as long. Likewise, the standards $(x_i, \mathbf{3}, 100)$, $(x_i, \mathbf{4}, 100)$, $(x_i, \mathbf{3}, 400)$ and $(x_i, \mathbf{4}, 400)$ were subsequently doubled ($\mathbf{q} = 2$). The procedure might become more obvious by inspecting Fig. 1: The arrows starting from the x-axis depict the basic adjustments, whereas the arrows starting from the arrowheads depict the successive adjustments.

On the whole, there were 22 different types of adjustments: Each of the two standard stimuli was paired with each of the six ratio production factors $\mathbf{p} = 1, 2, 3, 4, 6$, and 8, resulting in 12 types of basic $\times\mathbf{p}$ adjustments. In addition, each standard was combined with each of the five pairs $(p, q) = (2, 2), (2, 3), (2, 4), (3, 2)$, and $(4, 2)$, resulting in ten different types of successive $\times\mathbf{p} \times \mathbf{q}$ adjustments. Each type of adjustment was made 12 times, resulting in 264 trials per participant.

Results and discussion

Overall results

Overall mean adjustments for ($N = 10$) participants are depicted in Fig. 1 in the upper panel for the shorter standard duration of $t_1 = 100$ ms and in the lower panel for the longer standard of $t_2 = 400$ ms. The mean number of adjustments made in one trial was $M = 13.3$. In 66 % of all trials, participants made fine-step adjustments of duration. In further analyses, after a brief descriptive overview, the data sets of each participant are treated separately.

Monotonicity

The axiom of monotonicity was tested to confirm that duration perception of short intervals (100 to 4000 ms) occurs on a sensory continuum, i.e., that unequal temporal intervals are perceived as such and can be discriminated, respectively. From a descriptive point of view, the axiom of monotonicity seems to hold, because, as Fig. 1 shows, the mean outcome durations increase for increasing ratio production factors.

For the inferential statistics, two one-factor, repeated-measures analyses of variance (ANOVAs) tested the effect of the ratio production factor on the mean individual duration adjustments produced in basic trials only, separately for the two standards. For the standard $t_1 = 100$ ms, the ANOVA yielded significant differences among the different ratio production factors, $F(5, 45) = 306.9$, $p < .001$, $\eta^2 = .97$. A post hoc Tukey HSD test was conducted to check whether the mean adjustments of a pair of two adjacent ratio production factors are similar ($\sim$). The results showed that all pairs of ratio production factors ($(x, \mathbf{1}, 100) \sim (x, \mathbf{2}, 100)$, $(x, \mathbf{2}, 100) \sim$

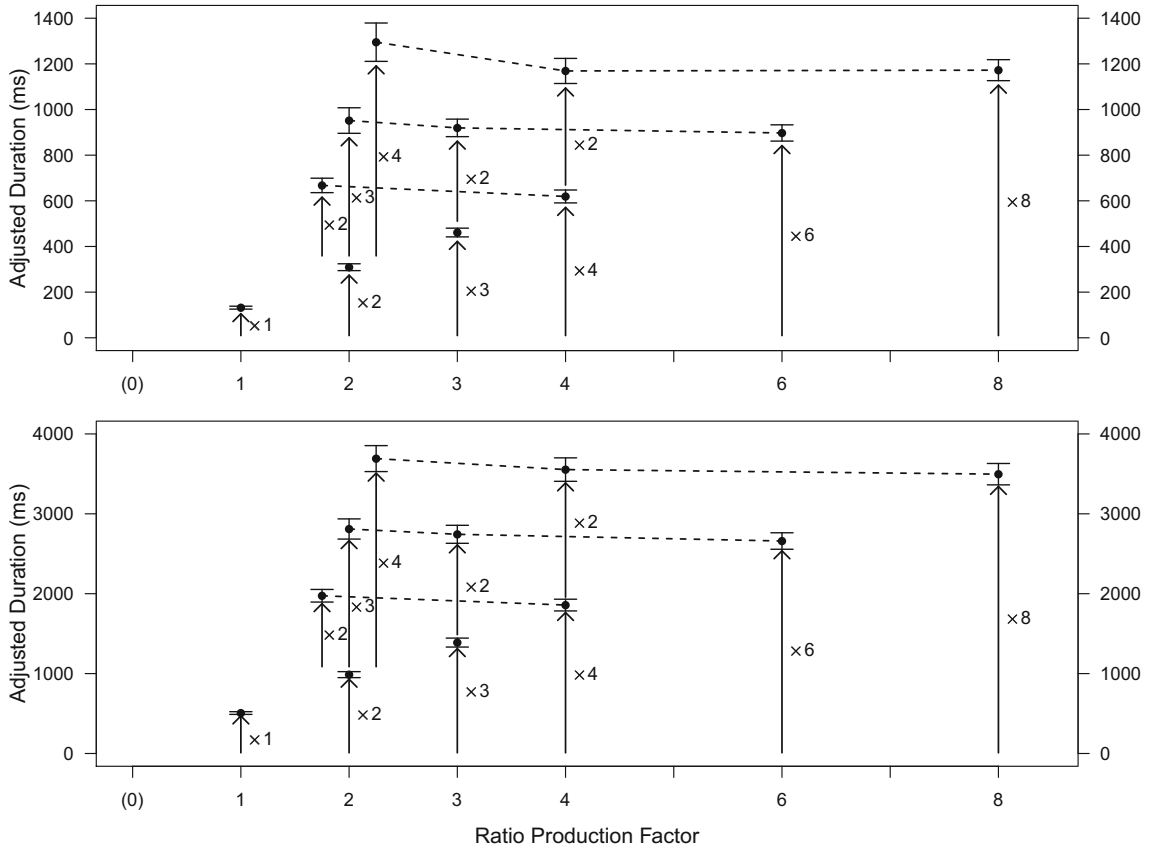

Fig. 1 Ratio productions made by ($N = 10$) participants in Experiment 1: Arithmetic means and standard deviations of basic and successive trials for $t_1 = 100$ ms top and $t_2 = 400$ ms bottom. Adjustments connected by dashed lines should coincide, if commutativity and multiplicativity hold

$(x, \mathbf{3}, 100), (x, \mathbf{3}, 100) \sim (x, \mathbf{4}, 100), (x, \mathbf{4}, 100) \sim (x, \mathbf{6}, 100)$ and $(x, \mathbf{6}, 100) \sim (x, \mathbf{8}, 100)$) differ significantly at $p < .001$. For the standard $t_2 = 400$ ms, a comparable ANOVA also yielded significant variations among the ratio production factors, $F(5, 45) = 140.7$, $p < .001$, $\eta^2 = .94$. Post hoc Tukey HSD comparisons revealed significant differences for all pairs of ratio production factors, $p < .001$ for $(x, \mathbf{4}, 400) \sim (x, \mathbf{6}, 400)$ and $(x, \mathbf{6}, 400) \sim (x, \mathbf{8}, 400)$, $p < .01$ for $(x, \mathbf{1}, 400) \sim (x, \mathbf{2}, 400)$ and $(x, \mathbf{3}, 400) \sim (x, \mathbf{4}, 400)$, and $p < .05$ for $(x, \mathbf{2}, 400) \sim (x, \mathbf{3}, 400)$. Further analyses of variance containing the factors block and session revealed no main effects for them, thus any practice effects can be ruled out.

Furthermore, a graphical analysis based on cumulative sums of the adjustments made, as proposed by Augustin and Maier (2008), was conducted for each participant. The axiom of monotonicity requires, that, for a fixed standard stimulus t, a ratio production factor $\mathbf{p}$ and a fixed number of repetitions i, the inequality $S(x_i, \mathbf{p}, t) < S(x_i, \mathbf{q}, t)$ holds, with $p < q$ and S representing the sum of duration adjustments x made up to the i-th trial. That is, the axiom of monotonicity holds, if for each standard t and each number i of repetitions (adjustments), the cumulative sums can be ordered by the ratio production factors used: $S(x_i, \mathbf{1}, t) < S(x_i, \mathbf{2}, t) < S(x_i, \mathbf{3}, t) < S(x_i, \mathbf{4}, t) < S(x_i, \mathbf{6}, t) < S(x_i, \mathbf{8}, t)$. Thus, for each participant and both standards t_1 and t_2, the $n = 12$ outcome durations of each type of $\times\mathbf{p}$ adjustments were summed up successively across trials. The cumulative sums, $S(x, \mathbf{1}, t)$, $S(x, \mathbf{2}, t)$, $S(x, \mathbf{3}, t)$, $S(x, \mathbf{4}, t)$, $S(x, \mathbf{6}, t)$ and $S(x, \mathbf{8}, t)$, of participant mg12, who is representative for the sample, are depicted in Fig. 2, the left panel shows the shorter and the middle panel shows the longer standard duration. Although the outcome durations of all trials $n = 1$ to 12 were summed up successively, only the cumulative sums in the range of trials $n = 7$ to 12 are plotted, in order to avoid inspecting the effects resulting from random influences for a small number of observations. Both graphs show that the curves for different ratio production factors never cross, e.g., that for the standard duration t_1, each cumulated outcome duration for $\mathbf{p} = 2$ is shorter than the corresponding cumulated outcome duration for $\mathbf{p} = 3$, meaning that at no point in the sequence of trials is monotonicity violated, thereby providing a more rigorous test than a comparison of overall condition means would.

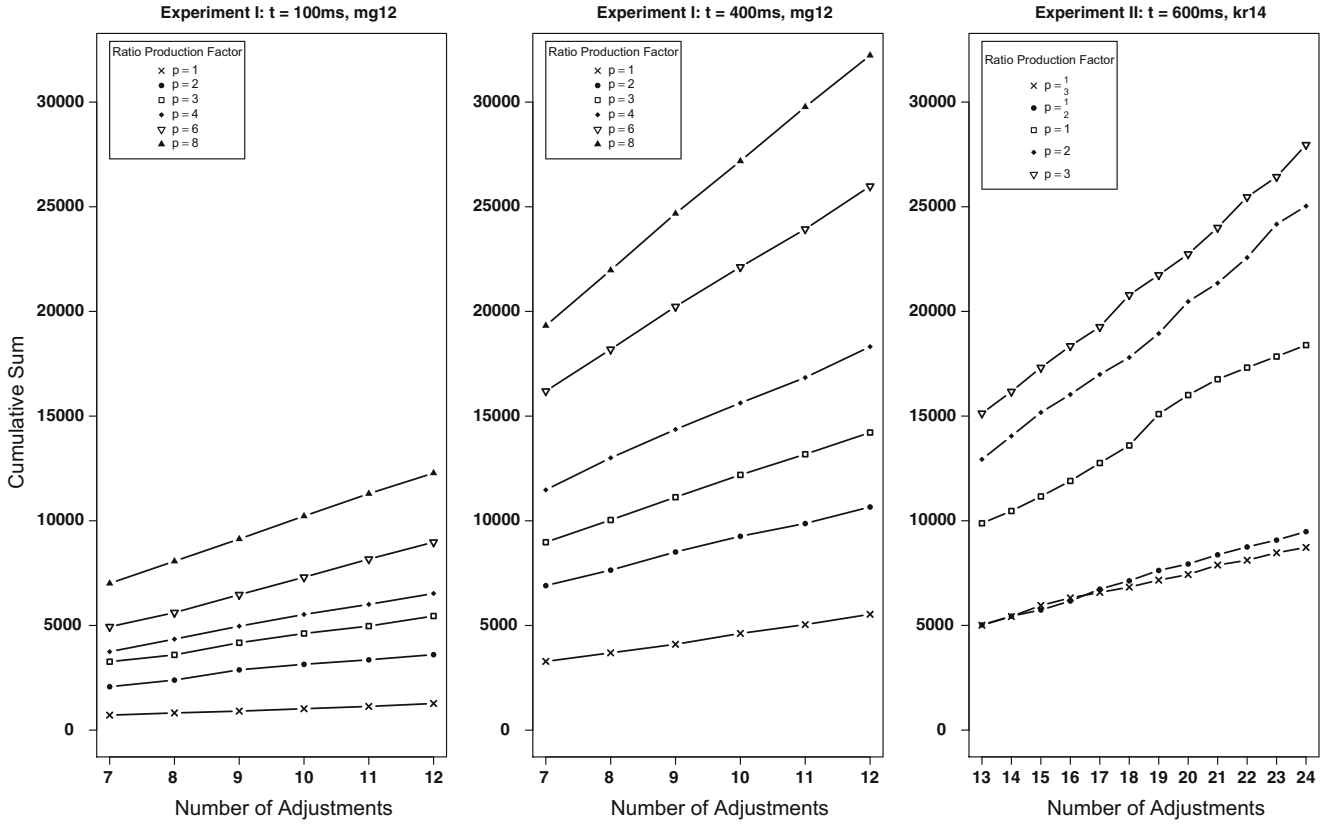

Fig. 2 Cumulative sums of the ratio productions made in Experiments 1 and 2. Each *curve* depicts the cumulative sums for a particular ratio production factor **p** as a function of the trial number (7 to 12, or 13 to 24, respectively, to minimize the effect of random influences for 'small' number of repetitions). The *left graph* refers to the shorter standard duration ($t_1 = 100$ ms), the middle one refers to the longer standard duration($t_2 = 400$ ms), both produced by participant mg12 and showing no violations of monotonicity, representative for the outcome of Experiment 1. The *right graph* refers to Experiment 2 and a standard of $t_3 = 600$ ms, showing magnitude productions by participant kr14 and violations of monotonicity for basic trials with $\mathbf{p} = \frac{1}{3}$ and $\mathbf{p} = \frac{1}{2}$

Commutativity

The axiom of commutativity provides evidence for the assumption that duration perception is based on a ratio scale. For testing commutativity, adjustments produced in successive trials are analyzed: Commutativity is taken to be satisfied, if a successive $\times\mathbf{p} \times \mathbf{q}$ adjustment is statistically indistinguishable from a successive $\times\mathbf{q} \times \mathbf{p}$ adjustment, i.e., if both types of raw adjustments emanate from the same distribution. For a descriptive analysis, Fig. 1 shows that most of the corresponding pairs of successive adjustments $\times\mathbf{p}\times\mathbf{q}$ and $\times\mathbf{q} \times \mathbf{p}$ which are connected by dashed lines coincide, indicating that the axiom holds for the overall means.

For individual inferential testing, nonparametric Mann–Whitney U tests (two-tailed, $\alpha = .1$) for both pairs $(p, q) = (2, 3)$ and $(2, 4)$ and both standards were conducted resulting in four tests per participant and a total of 40 tests for the entire sample.

A standard significance level of $\alpha = .1$ was used, because the aim of the analysis was to accept a statistical null hypothesis, thus making it harder to assume that an axiom holds for a particular comparison. A correction for multiple comparisons was not applied for the same reason.

For the entire sample, five violations in the 40 tests of the axiom of commutativity were observed (compare Table 1). Four of the five violations were produced by two participants (ml06, mn21), both for the standard of 100 ms. For seven of ten participants, the axiom of commutativity held in all cases.

Multiplicativity

The axiom of multiplicativity was tested to check whether the numerals as used by the participants can be taken at face value, i.e., whether there is a veridical transformation between perceived and mathematical numbers. For testing multiplicativity, the adjusted durations resulting from successive trials are compared with durations adjusted in basic trials: The axiom holds, if the duration resulting from the successive $\times\mathbf{p} \times \mathbf{q}$ ($\times\mathbf{q} \times \mathbf{p}$, respectively) adjustment is statistically indistinguishable from the basic $\times\mathbf{r}$ adjustment, with $r = pq$. In a descriptive manner, Fig. 1 also shows

Table 1 Experiment 1: Empirical evaluation of the commutative property for both standard stimuli with $t_1 = 100$ ms and $t_2 = 400$ ms for each ($N = 10$) participant

Participant	$100_{p,q} = 100_{q,p}$		$400_{p,q} = 400_{q,p}$	
	(p,q)			
	(2,3)	(2,4)	(2,3)	(2,4)
as11	−0.64	1.27	0.14	1.21
jb13	−0.06	1.39	0.69	0.29
mg12	0.29	1.39	1.62	0.81
mh15	−1.04	0.32	1.44	0.64
ml06	**2.02****	**2.71****	0.92	0.01
ml16	0.75	0.01	−0.23	−0.06
mn21	**−2.14***	**−1.85***	−0.06	0.55
mw28	−0.98	−1.27	0.40	0.75
tb01	0.90	0.92	1.33	1.50
we28	0.52	0.46	**−2.37***	−1.27

The table entries are $z(U)$-values of the computed Mann–Whitney U tests (two-tailed, $\alpha = 0.1$, $z_{(crit)} = 1.68$). Violations of commutativity are printed in boldface

Levels of significance: .1*, .01**, .001†

Table 2 Experiment 1: Empirical evaluation of the multiplicative property for both standard stimuli with $t_1 = 100$ ms and $t_2 = 400$ ms for each ($N = 10$) participant

Participant	$100_{p,q} = 100_r$			$400_{p,q} = 400_r$		
	(p,q)					
	(2,2)	(2,3)	(2,4)	(2,2)	(2,3)	(2,4)
as11	0.40	−0.10	1.34	**3.41**†	**2.10***	**2.15***
jb13	**2.02***	1.17	**−1.91***	0.06	−0.39	−0.94
mg12	1.47	**1.88***	−0.52	0.64	0.13	1.41
mh15	**2.83****	**2.65****	0.97	**2.71****	**3.12****	**3.39**†
ml06	**3.23**†	**−2.89****	**1.75***	−0.55	−0.13	1.38
ml16	0.17	−0.97	0.37	−0.17	0.84	0.54
mn21	−0.09	0.64	−1.02	1.04	−0.40	−0.44
mw28	−1.13	−0.87	0.44	1.56	**2.58****	−0.07
tb01	**−2.02***	**−2.05***	−1.17	0.17	1.07	0.50
we28	−1.67	**−3.39****	−1.44	−0.23	**−1.78***	−0.71

The table entries are $z(U)$-values of the computed Mann–Whitney U tests (two-tailed, $\alpha = 0.1$, $z_{(crit)} = 1.68$). Violations of multiplicativity are printed in boldface

Levels of significance: .1*, .01**, .001†

that most of the pairs of successive adjustments $\times\mathbf{p} \times \mathbf{q}$ and $\times\mathbf{q} \times \mathbf{p}$ are commensurate with the corresponding adjustments of $\times\mathbf{r}$ (with which they are connected by dashed lines), thus indicating multiplicativity to hold for the entire sample.

The individual inferential statistics tested multiplicativity by conducting Mann–Whitney U tests (two-tailed, $\alpha = .1$) for the three pairs $(p, q) = (2, 2)$, $(2, 3)$ and $(2, 4)$ and both standards, which results in six tests for each participant and a total of 60 tests for the entire sample. Altogether, 19 violations of 60 comparisons for the axiom of multiplicativity were observed (compare Table 2). For only two participants did the axiom of multiplicativity hold in all cases, whereas the other participants showed violations in one to five of six tests.

Model fitting procedure

Furthermore, linear regressions were computed for all participants and both standards to estimate the parameters α and β for the power law ($\varphi(t) = \alpha t^{\beta}$) as well as the parameters a and b for a simple linear function ($\varphi(t) = a + bt$). It was assumed that the individually adjusted durations of $(x, \mathbf{p}, 100)$ and $(x, \mathbf{p}, 400)$ are perceived to be $\mathbf{p}$ times as long as the standards, respectively. Thus, for the linear model, a linear regression of the ratio production factors $\mathbf{p}$ constituting the dependent variable on the individual adjustments constituting the independent variable was computed. For the power function, a linear regression was computed as well, with the logarithmically transformed ratio production factor $\mathbf{p}$ as the dependent variable and the logarithmically transformed individual adjustments serving as the independent variable.

The estimated parameters and squared correlation coefficients R^2 for both linear model and power function and for both standards are shown in Table 3. The comparison between linear and power-function model shows, that for the short standard, the power-function model results in a slightly better fit ($t(13.15) = 1.885$, $p = .082$) explaining 4.7 % more of the variance. For the longer standard, the linear model seems to fit the data as well as the power-function model ($t(15.96) = 0.735$, $p = .47$), the latter explaining only 2.3 % more of the variance. Furthermore, the power function exponents estimated for the two standards significantly differ in size, $t(11.58) = 3.67$, $p = .003$. The exponent β of the power function yielded an average of $\beta(t_1) = 0.87$ ($\beta < 1$ in all cases) for the shorter standard and $\beta(t_2) = 1.02$ ($\beta > 1$ in 6 of 10 cases) for the longer standard duration. Both the linear and the power function indicate a reasonable fit to the data with R^2 ranging from 0.71 to 0.98 for the raw-data adjustments.

Summary

The analyses showed that the axiom of monotonicity was not violated, i.e., the participants were able to produce monotonically ordered adjustments according to the different ratio production factors. The axiom of commutativity

Table 3 Experiment 1: Estimated parameters and squared correlation coefficients for linear model and power function for both standard stimuli with $t_1 = 100$ ms and $t_2 = 400$ ms and each ($N = 10$) participant

Participant		Linear Model			Power Function		
	t	a	b	R^2	$ln(\alpha)$	β	R^2
as11	t_1	0.76	5.23	0.71	0.78	0.90	0.84
	t_2	−0.34	2.41	0.91	0.29	1.15	0.92
jb13	t_1	0.48	7.32	0.88	0.90	0.95	0.91
	t_2	−0.11	2.49	0.97	0.36	1.07	0.97
mg12	t_1	0.02	7.53	0.93	0.84	0.89	0.93
	t_2	−0.32	2.91	0.92	0.40	1.12	0.94
mh15	t_1	0.41	7.20	0.91	0.88	0.91	0.95
	t_2	−0.58	2.72	0.95	0.32	1.19	0.96
ml06	t_1	0.67	5.46	0.84	0.78	0.80	0.90
	t_2	0.29	1.41	0.98	0.24	0.87	0.98
ml16	t_1	0.91	4.09	0.76	0.71	0.85	0.85
	t_2	0.52	1.77	0.77	0.34	0.87	0.82
mn21	t_1	0.44	6.30	0.90	0.82	0.85	0.94
	t_2	−0.18	2.62	0.96	0.37	1.10	0.97
mw28	t_1	0.25	6.55	0.89	0.81	0.88	0.90
	t_2	0.47	2.13	0.80	0.36	1.05	0.85
tb01	t_1	0.46	5.51	0.89	0.77	0.88	0.91
	t_2	0.26	1.94	0.95	0.35	0.89	0.97
we28	t_1	0.57	4.86	0.89	0.73	0.79	0.93
	t_2	0.56	1.95	0.90	0.38	0.91	0.94

was violated in 12.5 % of all tests, while multiplicativity was violated in 32 % of all tests. The estimated power function exponents for the two standards clearly differ in value, that is, the estimation of the parameters of the power law seems to depend on the duration of the standard, and, for the longer standard, seems to be close to 1 resulting in a simple linear function.

Experiment 2

The previous experiment investigated the axioms of monotonicity, commutativity, and multiplicativity for the perception of duration to test the validity of assumptions basic to Stevens' direct scaling methods. Since the axiom of commutativity was found to be valid in 87.5 % of all cases, it can be assumed that participants' processing of short duration in a ratio production experiment is based on a ratio scale. However, it might be difficult to describe the relationship between the mathematical numbers provided in the experimental instruction and the numbers as interpreted by the participants, because the axiom of multiplicativity held in only 68 % of the tests, i.e., roughly a third of the participants do not appear to process the numbers at their face value. Comparisons of the estimated exponents of the power functions describing the relationship between physical and perceived duration yielded significantly different exponents for the two standard durations employed.

The observation that the two different standard durations used in Experiment 1 result in diverging exponents has traditionally been classified as a context effect. In the domain of psychophysical scaling, several types of context effects have been described: Besides the stimulus range used in the experiment (Garner, 1954; Ward et al., 1996), the numerical examples given in the experimental instruction (Robinson, 1976) and the number values assigned to the standard stimuli (Beck and Shaw, 1965), or even the entire experimental context might have an influence on the size of the exponent. Therefore, the psychological meaningfulness of the exponent has been called into question (Lockhead, 1992). In contrast to this point of view, other investigators have argued that finding the 'true' exponent is still possible (Teghtsoonian and Teghtsoonian, 2003; Teghtsoonian, 2012).

However, in the axiomatic-measurement literature, this problem has been framed as a more fundamental issue of *meaningfulness* (Stevens, 1946; Luce, 1978; Narens, 1981). For each power function describing the relationship between

the physical intensity of a stimulus and its perceived magnitude, one might ask whether the parameters of this function are psychologically meaningful, i.e., invariant under certain transformations. Note that the exponent of the power function depends on the sensory continuum, the participant's individual perception—which does not exert a very strong influence (Teghtsoonian and Teghtsoonian, 1983)—and potential contextual influences as mentioned above. Furthermore, it might also vary under changes of the physical measurement scale f (Narens and Mausfeld, 1992) and the size of the standard (Augustin, 2008) used in an experiment. If, for example, the measurement scale f is transformed to another scale g measuring the same physical intensity as f and if these scales are neither log-interval nor ratio scales, then it must be assumed that the choice of the scale has an influence on the exponent of the power function. Thus, the obtained exponent has no psychological relevance, or is not meaningful.

But even if the exponent of the power function is invariant under changes of the physical stimulus scale applied in the experiment, it has to be investigated, whether the exponent is invariant under changes of the standard stimulus t being the basis for the estimates or adjustments made by the participants. Augustin (2008) suggests a mathematical method to examine the dependency on the standard by postulating two further axioms that can be evaluated empirically that is weak multiplicativity and invertibility. The axiom of *weak multiplicativity* is formulated as:

$$\text{For } t, y, z \in X \text{ and a real number } p > 0,$$
$$(y, \mathbf{p}, t) \in E, (z, \mathbf{1/p}, y) \in E \Rightarrow (z, \mathbf{1}, t) \in E. \quad (5)$$

That means, weak multiplicativity holds, if the stimulus intensity resulting from successive adjustments $\times\mathbf{p} \times \frac{\mathbf{1}}{\mathbf{p}}$ is equal to the stimulus intensity resulting from the basic adjustment with $\mathbf{p} = 1$. For example, doubling the duration of the standard and then halving this adjustment should result in the same final duration as matching the duration of the comparison interval to that of the standard. Weak multiplicativity is very similar to Narens' axiom of multiplicativity. But while multiplicativity has to hold for all cases $\mathbf{p} > 0$ and $\mathbf{q} > 0$, weak multiplicativity is a special case of multiplicativity with $\mathbf{q} = \frac{\mathbf{1}}{\mathbf{p}}$, i.e., even if the axiom of multiplicativity is violated, the axiom of weak multiplicativity might hold.

The axiom of *invertibility* is formulated as:

$$\text{For } t, y \in X \text{ and } \mathbf{p} > 0, (y, \mathbf{p}, t) \in E \Leftrightarrow (t, \mathbf{1/p}, y) \in E. \quad (6)$$

In other words, invertibility holds, if the intensity of a stimulus resulting from successive adjustments $\times\mathbf{p} \times \frac{\mathbf{1}}{\mathbf{p}}$ is equal to the stimulus intensity of the standard t or, put simply, if it is possible to undo a $\times\mathbf{p}$ adjustment by requiring to produce its reciprocal $\times\frac{\mathbf{1}}{\mathbf{p}}$. So weak multiplicativity and invertibility differ in whether the successive adjustment resulting from $\times\mathbf{p} \times \frac{\mathbf{1}}{\mathbf{p}}$ is equal to the adjustment of $\times\mathbf{1}$ in the first case and the actual duration of the standard in the second case. As Augustin (2008) stated, both axioms are necessary and sufficient conditions for the exponent of Stevens' power law to be invariant under changes of the standard t.

However, previous magnitude production experiments using ratio production factors $p < 1 < q$ assume fractions and integers to be processed differently: A study by Luce, Steingrimsson and Narens (2010) showed the axiom of commutativity to be violated for the $N = 2$ participants tested when fractions and integer ratios were mixed. Steingrimsson and Luce (2007) found comparable discrepancies for the axiom of multiplicativity for $N = 3$ participants in an experiment on loudness production. Augustin (2008) explicitly tested the two crucial axioms of weak multiplicativity and invertibility and found them to be violated for all $N = 10$ participants who performed ratio productions of the area of visually presented circles.

For the perception of duration, numerous experiments to determine the exponent of Stevens' power law were conducted using standard durations ranging from 50 ms to 300 s (Eisler, 1976). Although the exponents derived from these experiments vary between $\beta = 0.23$ and 1.36, it has not been sufficiently investigated whether these differences may be caused by the use of different standards. A study by Kane and Lown (1986) used standard durations of 30 and 180 s and did not find the length of standard duration to affect the size of the power law exponent. Eisler's (1976) review of 111 studies on duration perception, however, reported lower exponents obtained from experiments using standard durations shorter than 500 ms, but they did not specify this observation in more detail.

Because, in contrast, even the exponents derived from Experiment 1, using standards of $t_1 = 100$ and $t_2 = 400$ ms, significantly differ in size, $\beta(t_1) = 0.87, \beta(t_2) = 1.02$, it is plausible to investigate the meaningfulness of the power law exponent for the perception of duration by means of Augustin's (2008) additional axioms.

Methods

Participants

Fifteen participants were tested in the experiment. The sample consisted of 14 female participants and one

male with a median age of 23 years, ranging from 23 to 45 years. They were all students of psychology, but did not have any prior knowledge of the current hypotheses. Again, testing was conducted individually in a double-walled sound-attenuated listening chamber (IAC).

Stimuli and apparatus

Stimuli were generated using the same apparatus and signal parameters as in Experiment 1. The fixed standard, however, had a duration of 600 ms, while comparison stimuli varied in duration; their initial length was randomly chosen between 200 and 1,800 ms.

Procedure

In Experiment 2, the participants had to complete 216 trials altogether. The trials were divided into two identical test sessions taking place on two different days. Each session was composed of 12 blocks of nine trials, each. After three practice trials, data were recorded. After having completed three blocks, the participants could take a short break.

As in Experiment 1, participants had to adjust the comparison interval, separated from the standard[1] by an interstimulus interval of 500 ms, according to a certain ratio production factor $\mathbf{p}$ ($\mathbf{p} = \frac{1}{3}, \frac{1}{2}, 1, 2, 3$) presented on the screen. To increase or decrease the duration of the comparison interval, participants had to press the appropriate cursor key, either in small (20 ms) or in large steps (200 ms). Again, both tones were replayed after each keystroke, with the comparison tone having changed in duration.

In each of the 24 blocks, the standard of $t_3 = 600$ ms was combined with the ratio production factors $\frac{1}{3}, \frac{1}{2}, 1, 2$, and 3 resulting in five types of $\times\mathbf{p}$ adjustments and 120 basic trials altogether. In the successive trials, the individual basic adjustments $(x_i, \mathbf{p}, 600)$ were used as standard stimuli, i.e., the new standard $(x_i, \frac{1}{3}, 600)$ had to be adjusted using the ratio production factor $\mathbf{q} = 3$, the standard $(x_i, \frac{1}{2}, 600)$ was combined with the ratio production factor $\mathbf{q} = 2$, the standard $(x_i, 2, 600)$ was combined with the ratio production factor $\mathbf{q} = \frac{1}{2}$ and the standard $(x_i, 3, 600)$ had to be adjusted with the ratio production factor $\mathbf{q} = \frac{1}{3}$. The four types of $\times\mathbf{p} \times \mathbf{q}$ adjustments resulted in 96 successive trials altogether.

[1]The order of standard and comparison was counterbalanced in this experiment, i.e., in 12 of 24 repetitions of each trial type, the standard was presented *after* the comparison tone. However, the analyses did not reveal any effects of the order of standard and comparison.

Results and discussion

Overall results

The overall means based on all ($N = 15$) participants are depicted in Fig. 3. The mean number of adjustments made in one trial was $M = 7.4$. In 35 % of the adjustments, participants were using large steps to reach their final decision. In further analyses, after a brief descriptive overview, the data sets of each participant are treated separately.

Monotonicity

An ANOVA on the duration adjustments yielded significant differences between the different ratio production factors, $F(4, 56) = 200.6$, $p < .001$, $\eta^2 = .93$. Post hoc Tukey HSD comparisons revealed significant differences ($p < .001$) for all but one pair of ratio production factors, i.e., $(x, \frac{1}{3}, 600) \sim (x, \frac{1}{2}, 600)$, $p = .55$.

Furthermore, a graphical analysis based on cumulative sums was applied. As suspected from the results of the Tukey test, 4 of 15 participants exhibited violations of monotonicity in their adjustments of $\times\frac{1}{3}$ and $\times\frac{1}{2}$. An example is shown in the right panel of Fig. 2 for participant kr14, whose lines for $\times\frac{1}{3}$ and $\times\frac{1}{2}$ are at the same level or even cross. These four participants were excluded from further analyses.

Weak multiplicativity

The axiom of weak multiplicativity is satisfied, when the outcome of the $\times\mathbf{p} \times \frac{\mathbf{1}}{\mathbf{p}}$ adjustment is statistically indistinguishable from the duration of the $\times\mathbf{1}$ adjustment. This axiom was tested by performing nonparametric Mann–Whitney U tests (two-tailed, $\alpha = .1$) comparing the outcome of the four combinations $(p, q) = (\frac{1}{3}, 3)$, $(\frac{1}{2}, 2)$, $(2, \frac{1}{2})$, and $(3, \frac{1}{3})$ with that of the $\times 1$ adjustment. That was done individually for each of ($N = 11$) participants, resulting in a total of 44 tests. For the entire sample, 24 violations in 44 tests of the axiom of weak multiplicativity were observed. 18 of 22 violations were produced in trials with $(p, q) = (\frac{1}{3}, 3)$ and $(p, q) = (\frac{1}{2}, 2)$, while in trials with $(p, q) = (2, \frac{1}{2})$ and $(p, q) = (3, \frac{1}{3})$, only six violations of 22 tests were found; compare Table 4, left column.

Invertibility

The axiom of invertibility is satisfied, when the final outcome of the successive $\times\mathbf{p} \times \frac{\mathbf{1}}{\mathbf{p}}$ adjustments is statistically indistinguishable from the duration of the standard, $t_3 = 600$ ms. By conducting nonparametric Mann–Whitney U tests (two-tailed, $\alpha = .1$), it was tested whether the duration

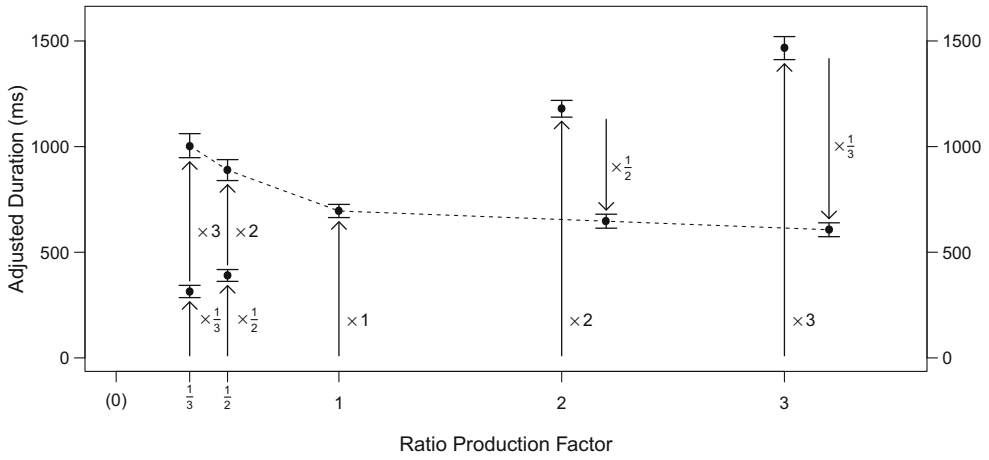

Fig. 3 Ratio productions obtained in Experiment 2: Arithmetic means and standard deviations of basic and successive trials for $t_3 = 600$ ms and $(N = 11)$ participants. Adjustments connected by dashed lines should coincide, if weak multiplicativity holds

adjustments of the successive trials with $(p, q) = (\frac{1}{3}, 3)$, $(\frac{1}{2}, 2)$, $(2, \frac{1}{2})$, and $(3, \frac{1}{3})$ may be produced by distributions with $\mu = 600$ ms. Tests were performed separately for the four combinations and individually for each of $(N = 11)$ participants, resulting in a total of 44 tests. For the entire sample, 25 violations of 44 tests of the axiom of invertibility were observed. 20 violations of 22 tests were produced in trials with $(p, q) = (\frac{1}{3}, 3)$ and $(p, q) = (\frac{1}{2}, 2)$, while in trials with $(p, q) = (2, \frac{1}{2})$ and $(p, q) = (3, \frac{1}{3})$, only five violations of 22 tests were found; compare Table 4, right column.

Model fitting procedure

Furthermore, regressions were computed for all participants to estimate the parameters for a linear psychophysical function as well as for a power function. The estimated parameters and squared correlation coefficients R^2 for both linear model and power function are shown in Table 5. The estimation of the exponent β of the power function revealed a $\beta > 1$ in 9 of 11 cases with an average of $\beta = 1.16$. The comparison between the two models shows no significant difference in their goodness of fit ($t(18.62) = 0.058$,

Table 4 Experiment 2: Empirical evaluation of weak multiplicativity and invertibility for the standard of $t_3 = 600$ ms duration for each $(N = 11)$ participant

Participant	$600_{p,q} = 600_{(1)}$				$600_{p,q} = 600$			
	(p,q)							
	$(\frac{1}{3}, 3)$	$(\frac{1}{2}, 2)$	$(2, \frac{1}{2})$	$(3, \frac{1}{3})$	$(\frac{1}{3}, 3)$	$(\frac{1}{2}, 2)$	$(2, \frac{1}{2})$	$(3, \frac{1}{3})$
ah06	**3.90**†	**2.06***	−0.98	−1.00	**4.95**†	**1.98***	−0.99	−1.48
ar14	**4.75**†	**3.28****	−0.77	−1.63	**5.44**†	**4.45**†	0.99	0.01
ar18	**4.61**†	**2.24***	−0.15	**−2.03***	**4.95**†	**2.47***	0.99	**−2.47***
cb22	**2.86****	1.05	−1.24	**−2.21***	**2.97****	1.48	−1.48	**−3.46**†
cg26	**1.93***	0.33	−0.03	**−2.06***	**2.47***	0.49	1.48	−0.99
dy02	**4.76**†	**2.97****	−0.654	0.02	**5.94**†	**3.96**†	0.49	0.99
ek23	**4.93**†	**4.31**†	0.88	1.18	**5.94**†	**5.44**†	1.48	**1.98***
hs29	**4.33**†	**5.53**†	−0.86	−0.94	**4.95**†	**5.94**†	−0.49	−0.49
ji08	**4.68**†	**4.50**†	**−1.68***	**−2.12***	**5.44**†	**5.44**†	−0.99	−1.48
kw22	**5.44**†	**3.81**†	−0.96	0.06	**5.94**†	**5.44**†	1.48	**2.47***
lm15	0.25	0.48	−0.95	**−3.14****	**3.96**†	**4.45**†	**4.45**†	0.99

The table entries are $z(U)$-values of the computed Mann–Whitney U tests (two-tailed, $\alpha = 0.1$, $z_{(crit)} = 1.68$). Violations of weak multiplicativity and invertibility are printed in boldface

Levels of significance: .1*, .01**, .001†

Table 5 Experiment 2: Estimated parameters and squared correlation coefficients for linear model and power function for the standard stimulus of $t_3 = 600$ ms and each ($N = 11$) participant

Participant	Linear Model			Power Function		
	a	b	R^2	$ln(\alpha)$	β	R^2
ah06	−0.06	1.76	0.89	0.21	1.16	0.88
ar14	−0.27	2.07	0.78	0.22	1.31	0.82
ar18	−0.06	1.47	0.81	0.16	0.93	0.81
cb22	0.03	1.77	0.73	0.23	1.08	0.78
cg26	−0.11	1.92	0.83	0.22	1.07	0.82
dy02	−0.32	2.17	0.78	0.22	1.36	0.81
ek23	−0.59	2.69	0.79	0.29	1.60	0.79
hs29	−0.02	1.63	0.92	0.19	1.00	0.91
ji08	−0.06	1.82	0.78	0.20	0.97	0.76
kw22	−0.30	2.16	0.79	0.22	1.24	0.80
lm15	−0.12	1.77	0.75	0.18	1.05	0.82

$p = .53$), but they both explain considerably less variance, 78 %, than the models fitted in Experiment 1.

Summary

The analyses showed that the axiom of monotonicity was violated by four participants, i.e., these participants were not able to produce monotonically ordered adjustments for the ratio production factors $\mathbf{p} = \frac{1}{2}$ and $\mathbf{p} = \frac{1}{3}$. The axiom of weak multiplicativity was violated in 55 % of all tests. The axiom of invertibility showed comparable violation rates of 57 %. For $\times\frac{1}{\mathbf{p}} \times \mathbf{p}$ adjustments, both axioms were violated more often (82 %, 91 %) than for $\times\mathbf{p} \times \frac{1}{\mathbf{p}}$ adjustments (27 %, 23 %).

General discussion

In two experiments, the present study examined the validity of a number of axioms from representational measurement theory for the ratio production of time intervals. These axioms are fundamental for determining whether subjective duration may be assumed to constitute a ratio scale, and how the numerical scale values obtained may be interpreted. Furthermore, they can confirm the psychological meaningfulness of the function describing the relationship between physical and subjective duration.

Axiomatic evaluation and model fitting

In Experiment 1, multiple analyses revealed that, with all ratio production factors $\mathbf{p} \geq 1$, the axiom of monotonicity was corroborated, indicating that all participants were able to produce monotonically increasing durations in response to appropriate ratio instructions, thus satisfying a basic ordinal requirement for a scale.

The individual evaluation of commutativity and multiplicativity revealed large differences between participants: For some participants, such as mg12, ml16, and mw28, we found almost no axiom violations, whereas others (mh15, ml06) showed as many as five violations in ten tests. This finding implies that some participants were able to deal with the instructions of a ratio production experiment, i.e., they use the numbers presented in the experiment as they are requested to, whereas others were not.

The overall axiomatic evaluation showed the commutative property to hold for most participants (12.5 % violations) implying that, generalized, they are capable of processing duration on a ratio scale. However, the multiplicative property was violated in 32 % of all tests showing that the numerals as used by the participants or in the instructions to describe perceived duration cannot always be taken at face value. Thus, Narens' (1996) axioms which are fundamental to Stevens' direct scaling approach could be validated, in that a ratio scale of duration can be assumed, but there is no obvious way to derive the actual scale values.

The results for commutativity and multiplicativity of the present experiment are comparable with findings for other sensory continua. For the perception of area, Augustin and Maier (2008) reported violation rates of 12 % for the axiom of commutativity and 61 % for multiplicativity. Ellermeier and Faulhammer (2000) found commutativity to be violated in 11 % of all cases and violations of multiplicativity in 94 % of the tests, while Zimmer (2005) reported violations rates of 14 % and 89 %, with both studies examining the perception of loudness. For the perception of pitch, Kattner and Ellermeier (2014) found the axioms to be violated in 22 % and 33 % of all tests, respectively.

Furthermore, power function exponents fitted to the ratio productions made relative to the two different standards used in Experiment 1 turned out to differ significantly. Therefore, it was tested whether the dependency of the standard can be confirmed by axiomatic testing. The axioms of weak mutiplicativity and invertibility, necessary and sufficient conditions for the invariance of the exponent of the power function under changes of the standard, were evaluated in Experiment 2.

The results show the crucial axioms of weak multiplicativity and invertibility to be violated in 55 % and 57 % of all cases, respectively, suggesting, as already assumed in Experiment 1, the power function exponent to depend on the size of the standard. From a scaling perspective, this might be construed as a context effect due to the use of a fallible method: ratio production. It might be argued that an unconstrained method using 'no designated standard, no assigned modulus' disposes of the influence of the standard

simply by omitting it. But since there is no axiomatic framework to test this (one stimulus - one response) methodology for internal consistency, we appear to be stuck with ratio production (or estimation) for the time being.

The results of the present axiomatic evaluation are comparable with findings made in the perception of area, where violation rates of 70 % for the axiom of weak multiplicativity and 72.5 % for the axiom of invertibility were reported (Augustin, 2008). For the perception of loudness and pitch, weak multiplicativity and invertibility were not evaluated yet.

Comparisons between a linear model and a psychophysical power function reveal both types of models to fit the data quite well, with comparably high proportions of variance explained. However, since Experiment 2 has shown that their estimated parameters depend on the size of the standard, the psychophysical functions fitted do not appear to be meaningful.

Implications

The results of the present experiments can be helpful to draw conclusions on the conception of further studies of duration scaling.

An interesting question—suggested by one of the reviewers—might be, whether the participants who performed ratio productions of duration without any axiom violations do so for other sensory modalities, as well. That might clarify whether full compliance with the axioms is due to a superior way of handling numbers in general or whether it is specific to a given modality studied.

For successive adjustments, a systematic bias as reported in other studies was found: The final adjustments reached in successive trials, e.g., $\times 2 \times 3$, often exceeded the adjustments made in corresponding basic trials, e.g., $\times 6$. This patterns seems to be systematic, since it was found for other sensory modalities as well, e.g., Augustin (2008) reported a similar bias for area adjustments. Ellermeier and Faulhammer (2000) found $\times 2 \times 3$ loudness adjustments to be systematically higher in level than $\times 6$ adjustments, and Zimmer (2005) found the same pattern for loudness fractionation, i.e., the outcome of a $\times \frac{1}{6}$ adjustment produced less of a level reduction than the outcome of successive $\times \frac{1}{2} \times \frac{1}{3}$ adjustments. Steingrimsson and Luce (2007) investigated this bias and explained it by referring to a 'numerical distortion'. They stated that the relationship between scientific numbers and numbers used by the participants is not linear but can be described by another function, e.g., by a power function with an exponent < 1 causing successive adjustments to be greater in size than basic adjustments. So, if multiplicativity as tested in this experiment fails, so-called *k*-multiplicativity can be tested to examine whether the relationship between scientific numbers and numbers used by the participants follows a power function with a constant exponent.

Furthermore, in Experiment 2 , the very basic axiom of monotonicity was found to be violated for four of 15 participants. These participants did not produce distinguishable duration adjustments for ratio production factors $\mathbf{p} < 1$, although their adjustments for $\mathbf{p} \geq 1$ clearly follow a monotonic order. It can be ruled out that this finding might be due to a kind of floor effect, because in Experiment 1, even shorter durations were adjusted without any difficulty.

Furthermore, an unpublished experiment conducted in our laboratory investigated whether monotonicity, commutativity and multiplicativity can reliably be shown to hold for the fractionation of time intervals and revealed violation rates comparable to the results of Experiment 1. Thus, it might be assumed that the participants who violated monotonicity in Experiment 2 did not necessarily have difficulties in processing fractions, but might have misconceptions regarding the instructions of the mixed condition itself.

Additionally, a noteworthy order effect was observed when comparing the adjustments of $\times \frac{1}{3} \times 3$ and $\times \frac{1}{2} \times 2$ with the adjustments of $\times 3 \times \frac{1}{3}$ and $\times 2 \times \frac{1}{2}$: All successive adjustments $\times \mathbf{p} \times \mathbf{q}$ with $\mathbf{p} < 1$ preceding $\mathbf{q} > 1$ resulted in considerably longer outcome durations than successive adjustments with $\mathbf{p} > 1$ followed by $\mathbf{q} < 1$. Augustin (2008) did not report this pattern for the perception of area, so this finding may be assumed to be specific for the perception of duration, but will have to be further investigated.

Furthermore, it might be investigated, how exactly the exponent of the power function varies under changes of the standard stimulus. It might be plausible, as the results of the present experiments assume, that the exponent increases with increasing standard duration.

Conclusions

In conclusion, the present experiments show that if using ratio production of temporal intervals, the measurement is based on a ratio scale, although a 'numerical distortion' impedes an unequivocal interpretation of the scale values. Thus, before the shape of the transformation function relating perceived and mathematical numbers is determined, power law fitting using ratio production should be taken with a grain of salt.

Furthermore, the fitting of curves describing the relationship between physical and perceived time, regardless of power function or linear relationship, is difficult: Even if both kinds of models seem to describe the relationship quite well, the estimated parameters depend on the size of the reference stimulus used in the experiment and thus can hardly be interpreted in a psychologically meaningful way.

References

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, *26*(5), 340–354.

Augustin, T. (2008). Stevens' power law and the problem of meaningfulness. *Acta Psychologica*, *128*(1), 176–185.

Augustin, T., & Maier, K. (2008). Empirical evaluation of the axioms of multiplicativity, commutativity, and monotonicity in ratio production of area. *Acta Psychologica*, *129*(1), 208–216.

Beck, J., & Shaw, W. A. (1965). Magnitude of the standard, numerical value of the standard, and stimulus spacing in the estimation of loudness. *Perceptual and Motor Skills*, *21*(1), 151–156.

Bobko, D. J., Thompson, J. G., & Schiffman, H. R. (1977). The perception of brief temporal intervals: Power functions for auditory and visual stimulus intervals. *Perception*, *6*(6), 703–709.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Eisler, H. (1975). Subjective duration and psychophysics. *Psychological Review*, *82*(6), 429–450.

Eisler, H. (1976). Experiments on subjective duration 1868–1975: A collection of power function exponents. *Psychological Bulletin*, *83*(6), 1154–1171.

Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics*, *62* (8), 1505–1511.

Garner, W. R. (1954). Context effects and the validity of loudness scales. *Journal of Experimental Psychology*, *48*(3), 218–224.

Gescheider, G. A. (1997). *Psychophysics the fundamentals*. London: LEA.

Kane, L. S., & Lown, B. A. (1986). Stevens' power law and time perception: Effect of filled intervals, duration of the standard, and number of presentations of the standard. *Perceptual and Motor Skills*, *62*(1), 35–38.

Kattner, F., & Ellermeier, W. (2014). Fractionation of pitch intervals: An axiomatic study testing monotonicity, commutativity, and multiplicativity in musicians and non-musicians. *Attention, Perception, & Psychophysics*, *76*(8), 2508–2521.

Kornbrot, D. E., Msetfi, R. M., & Grimwood, M. J. (2013). Time perception and depressive realism: Judgment type, psychophysical functions and bias. *PloS One*, *8*(8).

Lockhead, G. R. (1992). Psychophysical scaling: Judgments of attributes or objects. *Behavioral and Brain Sciences*, *15* (3), 543–558.

Luce, R. D. (1978). Dimensionally invariant numerical laws correspond to meaningful qualitative relations, Philosophy of Science.

Luce, R. D. (2002). A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review*, *109*(3), 520–532.

Luce, R. D. (2008). Symmetric and asymmetric matching of joint presentations: Correction to Luce (2004). *Psychological Review*, *115*(3), 601.

Luce, R. D., Steingrimsson, R., & Narens, L. (2010). Are psychophysical scales of intensities the same or different when stimuli vary on other dimensions? Theory with experiments varying loudness and pitch. *Psychological Review*, *117*(4), 1247–1258.

Narens, L. (1981). A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision*, *13*(1), 1–70.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, *40*(2), 109–129.

Narens, L., & Mausfeld, R. (1992). On the relationship of the psychological and the physical in psychophysics. *Psychological Review*, *99*(3), 467–479.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Robinson, G. H. (1976). Biasing power law exponents by magnitude estimation instructions. *Perception & Psychophysics*, *19* (1), 80–84.

Steingrimsson, R. (2011). Evaluating a model of global psychophysical judgments for brightness: II. Behavioral properties linking summations and productions. *Attention, Perception, & Psychophysics*, *73*(3), 872–885.

Steingrimsson, R., & Luce, R. D. (2005a). Evaluating a model of global psychophysical judgements - I: Behavioral properties of summations and productions. *Journal of Mathematical Psychology*, *49*(4), 290–307.

Steingrimsson, R., & Luce, R. D. (2005b). Evaluating a model of global psychophysical judgements - II: Behavioral properties linking summations and productions. *Journal of Mathematical Psychology*, *49*(4), 308–319.

Steingrimsson, R., & Luce, R. D. (2007). Empirical evaluation of a model of global psychophysical judgments - IV: Forms for the weighting function. *Journal of Mathematical Psychology*, *51*(1), 29–44.

Steingrimsson, R., Luce, R. D., & Narens, L. (2012). Brightness of different hues is a single psychophysical ratio scale of intensity. *The American Journal of Psychology*, *125*(3), 321–333.

Stevens, S. S. (1946). On the theory of scales of measurement.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *American Journal of Psychology*, *69*(1), 1–15.

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

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(6), 377–411.

Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception & Psychophysics*, *33*(3), 203–214.

Teghtsoonian, M., & Teghtsoonian, R. (2003). Putting context effects into context. In *Proceedings of the Nineteenth Annual Meeting of the International Society for Psychophysics*.

Teghtsoonian, R. (2012). The standard model for perceived magnitude: A framework for (almost) everything known about it. *The American Journal of Psychology*, *125*(2), 165–174.

Ward, L. M., Armstrong, J., & Golestani, N. (1996). Intensity resolution and subjective magnitude in psychophysical scaling. *Perception & Psychophysics*, *58*(5), 793–801.

Zimmer, K. (2005). Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics*, *67* (4), 569–579.