AXIOMATIC APPROACHES TO STEVENS' MAGNITUDE SCALING: RECENT DEVELOPMENTS

Karin Zimmer[1,2] and Wolfgang Ellermeier[1]
[1]*Sound Quality Research Unit, Dept. of Acoustics, Aalborg University, Denmark*
[2]*Leibniz Institute for Science Education, Kiel University, Germany*
we@acoustics.aau.dk zimmer@ipn.uni-kiel.de

Abstract

In 1996, Narens showed that Stevens' methods of magnitude scaling are based on but a few qualitative assumptions that are straightforward to evaluate empirically. Two crucial assumptions are commutativity *(the outcome of a sequence of two assessments does not depend on their order) and* multiplicativity *(the outcome of a sequence of two assessments equals a single assessment if the number associated with the single assessment equals the product of the numbers associated with the sequenced assessments). In an initial test of these axioms in the loudness production of 1 kHz-tones (Ellermeier & Faulhammer, 2000), the authors found commutativity to hold and multiplicativity to fail in the majority of listeners, leading to the conclusion that, while respondents seem to be able to base their judgments on a ratio-scale of sensation strength, the numerals used in the assessments do not correspond to the scale values proper. This situation inspired research into the generalizability of both the empirical findings and Narens' (1996) theory. This paper will give an overview of these recent developments, focussing on empirical evaluations of commutativity and multiplicativity in different sensory modalities and experimental tasks on the one hand, and on theories that are relaxing these axioms which are inherent in Stevens' approach, on the other.*

The last decade has seen a resurgence of interest in methods of supra-threshold scaling. Current theoretical developments in the field of mathematical psychology (e.g. Narens, 1996; Luce, 2002; Steingrimsson & Luce, 2005; Augustin, 2006) defined qualitative and empirically testable assumptions which underlie the direct scaling procedures advocated by Stevens (1956; 1975). In the following, an overview of this recent theoretical and empirical work shall be presented.

Structural properties of Stevens' magnitude scaling

In a typical magnitude scaling task, the observer is asked to evaluate the sensation strength a stimulus elicits by assigning a number, or by expressing the ratio of the sensation it elicits compared to a reference sensation (which is elicited by a standard stimulus). This procedure is called *magnitude estimation*. Alternatively, the subject may be asked to adjust the stimulus level so that it generates a pre-specified sensation, or - more commonly - that it produces a certain ratio of sensation strengths when compared to a standard. That procedure is called *magnitude production*. It is assumed (1) that the numbers assigned, or matched to an appropriate stimulus level, are valid on a ratio scale, and (2) that the number words used may be treated as if they were mathematical numbers.

Magnitude-scaling procedures are widespread, with applications ranging from perception to domains as diverse as sociology ('social psychophysics'), medicine ('pain perception') and engineering ('sound quality', 'annoyance'), because they are easily implemented, fast to administer, and have proved to be generally consistent and reliable.

As Narens (1996) shows, however, such procedures are valid only if two behavioural assumptions, called *commutativity* and *multiplicativity*, are met. Characterized briefly, assessments are commutative, if the order in which two sequenced judgments (such as tripling, or doubling) are applied, is irrelevant to the outcome. Multiplicativity holds if the sensation magnitude of two sequenced assessments matches the magnitude of a single assessment (e.g. making the stimulus six times as intense) whenever the mathematical number associated with the single assessment is equal to the product of the mathematical numbers associated with the sequenced assessments.

Narens (1996) mathematically proves that judgments are valid on a ratio scale, if commutativity as well as a few assumptions that are either technical in nature, or un-critical in most stimulus domains (e.g. *monotonicity*), are met. If, moreover, judgments are multiplicative, the numerals used can be taken to denote the mathematical numbers they signify.

As predicted by Narens (1996), an empirical test of commutativity and multiplicativity led to results that surprised both advocates and opponents of Stevens' method. Ellermeier & Faulhammer (2000) had listeners produce 1 kHz-tones to be two times, three times and six times as loud as a reference tone of the same frequency. It turned out that for all listeners but one, commutativity held in these loudness productions: Doubling the loudness of a tone and then tripling the outcome produced the same sound pressure levels as did performing both operations in the reverse order. Following Narens' rationale, it could thus be concluded that the listeners operated on a ratio scale. Multiplicativity, however, was rejected: In all listeners, the levels produced by consecutively doubling and tripling the loudness of a reference tone far exceeded those produced when the instruction was to generate a tone six times as loud as the reference. It thus became clear that the overt numerals which elicited the participants' productions could not be taken at face value, i.e., they could not be interpreted as mathematical numbers representing sensation strength.

Consequently, Stevens' (1956; 1975) assumption that an observer can judge sensation magnitude on a *ratio-scale* level held, while the yet stronger assumption that he or she is also able to *directly quantify* the sensation magnitude, that is, provide the mathematical number corresponding to the sensation, failed. Clearly, this situation raises a number of issues regarding the generalizability of both the empirical findings and Narens' theory proper:

1. Is the pattern of results obtained by Ellermeier & Faulhammer (2000) specific to the domain investigated (loudness)?

2. Is it specific to the operation used, i.e. magnitude production?

3. Can a specific function be found which relates the numerals used by respondents to the mathematical numbers actually representing sensation magnitude?

4. Can we draw conclusions on whether the judgments made are based on sensation ratios, as postulated by Stevens, or on sensation differences, as claimed by others?

5. What structural possibilities remain if commutativity fails?

In the remainder of the article, these issues shall be addressed, by reviewing recent developments inspired by Narens' (1996) work. This will be done in a non-formal fashion, i.e. devoid of mathematical notation. For details, please refer to the original publications.

Magnitude scaling in other sensory modalities

In a parallel investigation with that of Ellermeier & Faulhammer (2000), Peißner (1999) investigated whether Narens' axioms held in describing the perceived brightness of test circles on a computer screen. Requiring a doubling, tripling, or six-tupling of brightness, Peißner (1999) reported results quite similar to those of Ellermeier & Faulhammer (2000): While commutativity held for 8 of 11 observers, multiplicativity failed in 9 of the 10 respondents considered. Thus, there is no evidence that the rejection of the multiplicative property is modality-specific.

A recent study investigating magnitude production in visual area perception (Augustin & Sunzenauer, 2006), however, found both multiplicativity and commutativity to be violated for the majority of participants, thus calling the existence of a ratio scale into question for that specific task.

Results from fractionation

In his comprehensive psychophysical theory, Luce (2002) extends Narens' behavioural axiomatization to allow for the evaluation of fractionation tasks. Investigating the loudness of 1 kHz-tones, Zimmer (2005) used the fractions $\frac{2}{3}$ and $\frac{1}{4}$ to test commutativity, and the fractions $\frac{1}{6}$, $\frac{1}{3}$, and $\frac{1}{2}$ to evaluate multiplicativity. Consistent with previous experiments, she found commutativity to hold in 6 of 7 listeners, and multiplicativity to be violated in 5 of 6 listeners. This pattern of results is very similar to the one obtained by Ellermeier & Faulhammer (2000) and Peißner (1999).

The search for a transformation function relating numerals and numbers

As part of his theory, Luce (2002) proposes an axiom, based on Prelec (1998), which he terms the *reduction invariance property*. This axiom implies a possible form of a transformation function connecting the numerical assessments used in the task with the mathematical numbers representing the actual sensation magnitude for those cases, in which multiplicativity is violated.

Zimmer (2005) conducted an empirical test of the reduction invariance property, using fractionations of the loudness of 1 kHz-tones. She found the axiom to be violated in six of seven listeners and concluded that, in this domain, the family of transformation functions proposed by Luce (2002) must be rejected.

A more favourable result is announced by Steingrimsson & Luce (2006). Their empirical test is based on Luce's (2002) theory, which encompasses Stevens' magnitude production as a special case of an operation called *generalized ratio production* (see Steingrimsson & Luce, 2005). That procedure relies on assessments of the magnitude of *intervals*. Observers are asked to produce an interval between two stimuli which is p times bigger than a pre-specified interval between another stimulus pair. Using this task, Steingrimsson & Luce (2006) report a general form of reduction invariance to hold in the binaural loudness of 1 kHz-tones for all of 11 listeners. As Steingrimsson & Luce (2006) show, it is possible to specify a transformation function which relates the numerals used in the experimental task to mathematical numbers reflecting the sensation magnitude on a ratio-scale, if the assumption is dropped that the numeral "1", i.e. asking participants to produce a stimulus that elicits the same sensation strength as the reference, can be taken at face value, i.e., to denote the mathematical number 1.

Do ratio and and difference judgments relate to the same sensation scale?

Narens (1996, 1997, 2006) shows that if judgments of differences and ratios are made on the same ratio scale, then Torgerson's (1961) conjecture holds, which states that, in direct scaling, subjects only judge but one single quantitative relation between stimuli. Narens (1997) uses the commutativity axiom to propose an empirical test of Torgerson's conjecture, which Ellermeier et al. (2003) carried out. Listeners were asked to double, or triple the loudness of 1 kHz-tones, and to add a small, or large, loudness difference to a standard. Productions obtained by using these two types of instructions turned out to be commutative in four of six listeners. Thus, for the majority of the subjects, the instruction to produce ratios, or differences, produced results which are compatible, and - by implication - ratio-scalable on the common sensation scale. Note, however, that a different interpretation has recently been proposed by Masin (2006).

Extension to other scale types

The recent finding by Augustin & Sunzenauer (2006) that ratio productions of area were not even commutative in many cases, necessitates inquiring whether the respondent's scaling behaviour may be represented by a lower scale type than a ratio scale. Therefore, Augustin (2006) extended Narens' axiomatization to ordinal, interval, and log-interval scales. In particular, he showed that relaxing the multiplicative property leads to the formulation of an *independence condition* which is necessary for representing the respondent's scaling behavior by interval, log-interval, or ordinal scales. The independence axiom states that for any starting values used, generating a sensation p times as strong, and making the outcome of that operation q times as strong in turn, will result in a sensation that is $h(p,q)$ times as strong as the starting level, with $h(p,q)$ being a unique number, but not being required to be the product of p and q (as with multiplicativity). Further axioms that specify the log-interval or interval-scale type are provided by Augustin (2006), and an initial empirical test (Augustin & Sunzenauer, 2006) on area ratio production showed 9 out of 12 participants to produce at least one violation of the axiom required for an interval-scale representation.

Conclusion

In sum, Narens' (1996) axiomatization has rekindled interest in the foundations of Stevens' magnitude scaling, and produced a number of theoretical and empirical papers from different research groups, all firmly rooted in the tradition of axiomatic measurement theory. The empirical evaluation of the assumptions underlying magnitude scaling has demonstrated the soundness of the (axiomatic) method in different sensory modalities and for different task varieties such as ratio production and fractionation. Nevertheless, more theoretical work is needed to define structural properties which will enable the identification of a function relating the numerals used in the magnitude-scaling task to the mathematical numbers reflecting the sensation magnitude.

It is to be hoped that the recent interest in axiomatizing the quantitative judgments Stevens trusted his subjects with will put his influential methodology on a firmer theoretical foundation.

Acknowledgements

This work was prepared as part of the "Centercontract on Sound Quality" which establishes participation in and funding of the "Sound Quality Research Unit" (SQRU) at Aalborg University. The participating companies are Bang & Olufsen, Brüel & Kjær, and Delta Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

References

Augustin, T. (2006). Stevens' direct scaling methods and the uniqueness problem. *Psychometrika*, *70*, in press.

Augustin, T. & Sunzenauer, H. (2006). Stevens' direct scaling methods and the uniqueness problem: Empirical evaluation of axioms fundamental to ratio and interval scale level. *Acta Psychologica*, under revision.

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

Ellermeier, W., Narens, L., & Dielmann, B. (2003). Perceptual ratios, differences, and the underlying scale. In B. Berglund & E. Borg (Eds.), *Fechner Day 2003. Proceedings of the 19th annual meeting of the International Society for Psychophysics* (pp. 71–76). Stockholm, Sweden: International Society for Psychophysics.

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

Masin, S. C. (2006). The judgment operation of ratio production. *Perception & Psychophysics*, in press.

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. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 289–205). Mahwah, N.J.: Erlbaum.

Narens, L. (2006). Symmetry, direct measurement, and Torgerson's conjecture. *Journal of Mathematical Psychology*, *50*, in press.

Peißner, M. (1999). *Experimente zur direkten Skalierbarkeit von gesehenen Helligkeiten.* Universität Regensburg: Unpublished masters thesis.

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

Steingrimsson, R. & Luce, R. D. (2005). Empirical evaluation of a model of global psychophysical judgments: I. Behavioural properties of summations and productions. *Journal of Mathematical Psychology*, *49*, 290–307.

Steingrimsson, R. & Luce, R. D. (2006). Empirical evaluation of a model of global psychophysical judgments: IV. Forms for the weighting function, in preparation.

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

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

Torgerson, W. (1961). Distances and ratios in psychophysical scaling. *Acta Psychologica*, *19*, 201–205.

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