Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 49 (2005) 290–307
Journal of Mathematical Psychology
www.elsevier.com/locate/jmp

Evaluating a model of global psychophysical judgments—I: Behavioral properties of summations and productions

Ragnar Steingrimsson[a,*,1], R. Duncan Luce[b]

[a]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA
[b]Department of Cognitive Science, University of California Irvine, Social Science Plaza, Irvine, CA 92697-5100, USA

Received 19 May 2004; received in revised form 16 February 2005
Available online 22 April 2005

Abstract

The research presented is a partial empirical evaluation of the second author's proposed psychophysical theory [Luce (2002). *Psychological Review*, *109*, 520–532; Luce (2004). *Psychological Review*, *111*, 446–454]. The theory deals with the global percept of subjective intensity, in which there is a psychophysical function Ψ that maps pairs of physical intensities onto the positive real numbers and represents, in an explicit mathematical way, subjective summation and a form of ratio production. A number of behavioral properties have been shown to follow from these specific representations, and in the presence of certain plausible background assumptions these properties are also sufficient for the representations. In four auditory experiments, key behavioral properties of summation over the two ears and a form of generalized ratio production are evaluated empirically. Considerable support is reported for particular forms of Ψ for summations and ratio productions separately. A second article, Steingrimsson and Luce (*Journal of Mathematical Psychology*, in press), explores the behavioral properties that link summations and productions.

Keywords: Auditory summation; Ratio production; Magnitude production; Psychophysics; Matching; Production commutativity; Thomsen condition; Double cancellation; Auditory bias; Magnitude estimation

Psychophysicists have long been interested in how subjective attributes associated with physical intensity grow with intensity. This includes the literature on how a subjective measure arises using locally defined measures of discriminability or using more global methods such as magnitude estimation or production (Dzhafarov, 2002; Fechner, 1860; Stevens, 1975). Also studied is how intensity summates when, e.g., signals are administered independently to the two ears (Falmagne, 1976; Falmagne, Iverson, & Marcovici, 1979; Levelt, Riemersma, & Bunt, 1972; Gigerenzer & Strube, 1983; Schneider, 1988). Global psychophysics deals with understanding the nature of the perception of intensive dimensions throughout their full range of intensities. In contrast, local psychophysics attempts to understand very local relations of signals at or near discrimination threshold (Luce & Krumhansl, 1988, pp. 39–40).

Some relevant examples of global methods are, first, the ratio estimation and production methods of Stevens (1975) that led him to argue that for intensive dimensions the growth is as a power function and, second, the various discussions (usually involving some form of empirical matching) of how subjective intensities "add" when there is a natural form of joint presentation, such as to the two ears or two eyes. Specific examples of the latter that are concerned with additivity in the conjoint measurement sense are described below. Recently, the second author has developed a class of theories whose main conceptual novelty is to relate ratio production to summation

*Corresponding author.
E-mail addresses: ragnar@nyu.edu (R. Steingrimsson), rdluce@uci.edu (R. Duncan Luce).
[1]This article is based, in part, on the first author's Ph. D. dissertation (Steingrimsson, 2002).

0022-2496/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2005.03.003

(Luce, 2002, 2004). Moreover, parameter-free conditions are developed that are necessary and, under certain natural background conditions, sufficient for particular representations that are described in detail in Section 1. The goal of this, the first of several empirical articles, is to test some of these predictions in the context of loudness judgments manipulated by changes in physical intensity.

An important aspect of the approach taken is that one studies the adequacy of a representation in which there are both free functions and free parameters without estimating either the unspecified functions or the parameters followed by some form of goodness-of-fit estimation. The latter approach is commonly used, often with considerable difficulty, in cognitive modeling. The behavioral properties that we test here are all parameter free.

Two further points should be stressed. First, the theory is not domain specific in the sense that it can, in principle, apply to any intensive dimensions (e.g., loudness, brightness, or weight). Second, although neuronal activity ultimately underlies perception, the approach taken here is entirely behavioral. The important abstraction is that the results we obtain are valid answers to our questions regardless of what the neural machinery may be, only the behavior matters. In effect, we could use the very same approach to an alien life form or a robot. Consequently, we do not make any attempt to draw conclusions about the biological workings of the perceptual system from our results.

The specific domain for testing is binaural loudness of pure tones. In particular, the present article reports four experiments that test behavioral aspects of these theories for summation and ratio production separately, and Steingrimsson and Luce (in press) explores linking relations between the two.

Section 1 describes in detail the representations of signal summation and a generalized operation of ratio production (Luce, 2002, 2004). Then two necessary behavioral conditions that follow from these representations are stated, one for summation and the other for ratio production. Section 2 describes experimental methods that are common to all experiments. Section 3 examines the possibility of bias in the ears with both one- and two-ear matching. Section 4 reports two experiments that test the behavioral properties for summation and for ratio production separately. Finally, Section 5 summarizes the experimental findings of this article and suggests further work.

1. Underlying features of the theory

The properties are constructed using two psychological "operations". One is a form of summation of which loudness summation over the two ears is one example. The second is a production operation that is a generalized form of ratio production.

The original impetus for Luce's (2002) model of global psychophysical judgments were results originally developed within the context of utility theory (Luce, 2000). They were based on an assumption of no bias, or asymmetry, in the summation. Data from Experiments 1 and 2 here, led him to generalize those results so as to incorporate biased summation. This formulation was subsequently improved and generalized further in Luce (2004). Previous theoretical work of a similar nature to that of Luce (2002) has employed operations analogous to either the summation or production operations (e.g., Levelt et al., 1972; Narens, 1996). However, Luce takes the approach, typical of physics, of linking these two operations together—a feature that proves critical for establishing that a common psychophysical function can be used to represent both empirical manipulations, i.e., summation and production (see Steingrimsson & Luce, in press).

1.1. Primitives

1.1.1. Joint presentations

The first primitive is the set of ordered pairs (x,u), which here corresponds to the intensities to the two ears. We mean physical intensity, not the corresponding dB measure. Alternative interpretations include the perception of line-lengths, visual brightness to the eyes, weight-lifting in the two hands (e.g., de Weert & Levelt, 1974), cross-modal cases such that x and u belong to different modalities (Ward, 1990), etc. Each domain will, of course, require a separate experimental analysis of the theory. Even within the domain of audition, several interpretations are possible. For example, Schneider (1988) used signal intensities of frequencies separated by more than a critical band, so that the stimulus (x,u) has intensity x at frequency f and intensity u at frequency f'. Second, temporal, rather than a joint presentation, summation may be used in which (x,u) is interpreted as the presentation of x for a brief duration followed immediately by u presented for the same duration. Experiments based on this interpretation using white noise stimuli were reported by Zimmer, Luce, and Ellermeier (2001). Our interpretation here of the pair (x,u) is that of pure tones x and u of the same frequency and phase presented simultaneously to the left and right ears, respectively. In this context, let ε_{l} and ε_{r} denote thresholds for the left and the right ear respectively and let x' and u' be intensities actually presented in the left and the right ear respectively; then our notation is $x = x' - \varepsilon_{\mathrm{l}}$ and $u = u' - \varepsilon_{\mathrm{r}}$. Thus, $x = 0$ denotes the threshold intensity (or less) in the left ear and $u = 0$, denotes the same in the right ear.

For the behavioral task, respondents produce a tone z that (in some to-be-specified sense) is perceived as equal in loudness to the stimulus (x,u).

1.1.2. Ordering

The second primitive, $\succsim$, is the ordering of stimuli by, in our case, loudness: $(x,u)\succsim(y,v)$ means that the stimulus (x,u) is judged to be at least as loud as (y,v). The indifference relation of matching $\sim$ is defined by: $(x,u)\sim(y,v)$ if and only if both $(x,u)\succsim(y,v)$ and $(y,v)\succsim(x,u)$ hold. It is not difficult to show from the assumptions formulated by Luce (2002, 2004) that $\succsim$ is a weak order, i.e., transitive and connected, on the stimulus conjoint structure (Proposition 1 of Luce, 2002). Thus $\succsim$ behaves similarly to the ordering $\geqslant$ of the real numbers. Moreover, we assume that it agrees with physical intensity in the sense that if the intensity is held constant in one ear, the loudness varies monotonically with intensity changes in the other ear.

One natural question we may ask about loudness matching is the *symmetry* of joint presentations:

$$(x,u)\sim(u,x), \tag{1}$$

which we abbreviate as *jp-symmetry*. Whether or not this holds matters considerably for the nature of the theory. Therefore, our first two experiments focus on this property.

Each joint presentation (x,u) can be matched in any of three ways: by an intensity z_{l} in the left ear with 0 in the right ear, by 0 in the left ear and z_{r} in the right, and by z_{s} in both. Formally,

$$(x,u)\sim(z_{\mathrm{l}},0),\quad (x,u)\sim(0,z_{\mathrm{r}}),\quad (x,u)\sim(z_{\mathrm{s}},z_{\mathrm{s}}). \tag{2}$$

We refer to z_{l} and z_{r} as *asymmetric* matches and z_{s} as a *symmetric* match.

1.1.3. Generalized ratio production

The third primitive is a generalized form of ratio production. Suppose that $x>y\geqslant 0$ and let $p>0$ be a positive number. Let (z,z) denote a signal pair that the respondent says makes the subjective "interval"[2] from (y,y) to (z,z) stand in the ratio p to the subjective interval from (y,y) to (x,x). Clearly z is a function of x,y,p. It is convenient to write this function as a mathematical operator of the following form

$$(x,x)\circ_p(y,y) := (z,z), \tag{3}$$

where $A := B$ means A is defined by B.

Using the several background assumptions (listed in Luce, 2002, 2004), one can replace the symmetric pairs of (3) by non-symmetric pairs to get a more general expression

$$(x,u)\circ_p(y,v) = (z,w) \tag{4}$$

and vice versa. Thus, there is no loss of generality in studying $\circ_p$ in the symmetric case.

This approach generalizes ordinary *ratio production*, where no reference (y,y) is stated and is, implicitly, assumed to be $(0,0)$.[3] It is an empirical question whether the result is continuous as $y\to 0$.

Some years ago a debate erupted in the literature (Schneider, 1980; Birnbaum, 1982) about the question of whether people are actually able to produce reliable ratios and differences. We do not address that directly in the present work, but merely point out that if the behavioral properties we have isolated are sustained empirically, then the behavior can be represented *as if* subjective differences and ratios are involved; see (6) below.

1.2. Representations of (x,u) and $\circ_p$

We next state the numerical representations that derive from four behavioral properties (Luce, 2002, 2004), which are detailed in Sections 1.1.2, 1.3.1, and 1.3.2 of the present paper.[4]

Using asymmetric matching, z_{l} or z_{r} in (2), it has been shown that the following exist:

- A psychophysical function Ψ mapping joint presentations into the non-negative real numbers that is strictly increasing in each of its arguments with $\Psi(0,0)=0$;
- A strictly increasing distortion W of real numbers to the real numbers with $W(0)=0$;
- Constants $\delta\geqslant 0$, $\gamma>0$; and
- The following three properties hold:

$$\Psi(x,u) = \Psi(x,0)+\Psi(0,u)+\delta\Psi(x,0)\Psi(0,u) \quad (\delta\geqslant 0), \tag{5}$$

$$W(p) = \frac{\Psi[(x,x)\circ_p(y,y)]-\Psi(y,y)}{\Psi(x,x)-\Psi(y,y)} \quad (x>y\geqslant 0), \tag{6}$$

$$\Psi(x,0) = \gamma\Psi(0,x) \quad (\gamma>0). \tag{7}$$

Several comments:

First, as remarked earlier, we may replace any joint presentation by another one that is indifferent to it, so

[2] The term "interval" is being used figuratively to refer to the difference in loudness that respondents experience between two intensity pairs.

[3] A variant on this method is Stevens' (1975) method of magnitude production in which $y=0$ and no reference x is specified. When x is specified, it is usually called magnitude production with a standard, which involves the special case $(y,y)=(0,0)$, i.e., $(x,x)\circ_p(0,0)=(z,z)$.

[4] Luce (2002) presented all results in terms of psychophysical functions on each signal dimension, likewise in the first submitted version of Luce (2004). Ehtibar Dzhafarov, as a reviewer, saw that they could be neatly brought together as a psychophysical function over the signal pairs, a formulation that Luce subsequently adopted.

we may write (6) in the more general form

$$\frac{\Psi[(x,u)\circ_p(y,v)]-\Psi(y,v)}{\Psi(x,u)-\Psi(y,v)}=W(p)$$
$$[(x,u)\succ(y,v)\gtrsim(0,0)].$$

Second, this pair of functions, $\langle\Psi, W\rangle$, has a great deal of freedom for individual differences, namely in the two unspecified functions Ψ and W plus the two constants. Yet, these functions and the representation are guaranteed to exist, provided that four parameter-free behavioral conditions are satisfied. This situation may seem contradictory to some, but it is typical of axiomatic derivations of representations: no free parameters in the axioms and considerable freedom in the representation. The behavioral freedom lies entirely in individual differences in the ordering $\gtrsim$, not in parameters. Although, many methodological issues surround experimental tests of such behavioral properties, as we shall see below, doing so seems more definitive than trying to fit to the data representations that have many degrees of freedom.

Third, if (7) is satisfied, we speak of a left bias when $\gamma>1$, no bias when $\gamma=1$, and a right bias when $\gamma<1$. Constant bias is a very strong property which we do not study here but it is a focus of on-going research.

Fourth, using symmetric matching [z_s in (2)] one can prove that the summation property (5) and the production representation (6) both hold, but with (5) restricted to the case where $\delta=0$. Furthermore, the constant-bias property, (7), need not hold under symmetric matching.

Fifth, two important special cases of (5) are: the symmetric case where $(x,u)\sim(u,x)$ and the case where $\delta=0$. We will show in Experiments 1 and 2 that such symmetry rarely appears to be satisfied.

1.3. Two behavioral properties of the representations

The following subsections state behavioral properties that are implied individually by the representations (5) and (6). These properties are tested in Sections 3 and 4. Linking of the representations via a common function Ψ is the topic of Steingrimsson and Luce (in press).

1.3.1. Subjective summation

Note that although expression (5) is symmetric in the roles of $\Psi(x,0)$ and $\Psi(0,u)$, it is not really symmetric because, in general, we do not have that $\Psi(x,0)=\Psi(0,x)$. Thus, it does not imply joint presentation symmetry, (1), unless of course $\gamma=1$ in (7).

Second, representation (5) can always be transformed into a binary additive conjoint representation of summation of the form:

$$\Phi(x,u)=\Phi_l(x)+\Phi_r(u). \tag{8}$$

This is obviously true when $\delta=0$, using the identifications

$$\Phi(x,u)\coloneqq\Psi(x,u),\quad \Phi_l(x)\coloneqq\Psi(x,0),\quad \Phi_r(u)\coloneqq\Psi(0,u).$$

It is less obvious when $\delta>0$, but it is nonetheless true using the logarithmic transformation

$$\Phi(x,u)=\ln[1+\delta\Psi(x,u)]\quad(\delta>0).$$

This fact means that the key necessary condition of binary additive conjoint measurement, the *Thomsen condition*

$$\left.\begin{array}{l}(x,t)\sim(z,v)\\(z,u)\sim(y,t)\end{array}\right\}\Longrightarrow(x,u)\sim(y,v), \tag{9}$$

must hold (Krantz, Luce, Suppes, & Tversky, 1971). In a qualitative sense, the Thomsen condition describes the "additive cancellation" of t and z.

In the presence of monotonicity, solvability, and Archimedeanness (a way of stating that subjectively measured intensities are commensurable), which are a part of the general background assumptions, then the Thomsen condition, (9), implies the additive representation, (8). This in turn implies a stronger property called *double cancellation*, which is the same as (9) but with each $\sim$ replaced by $\gtrsim$. Obviously, double cancellation implies the Thomsen condition, but not conversely except when we have solvability and monotonicity (see Krantz et al., 1971, for details).

Some theoretical work has been carried out in the area of binaural loudness summation. For instance, Levelt et al. (1972) employed aspects of measurement theory to formulate an experiment that led them to conclude that loudness summation is additive in the sense of additive conjoint measurement and resulted in psychophysical functions that are approximate power functions for each ear. Luce (1977) established three conditions, including additivity, that lead to the results of Levelt et al. (1972). Luce's recent theories are in this tradition, but somewhat more elaborate. Falmagne (1976) developed a probabilistic version of additive conjoint measurement that he argued would facilitate empirical testing.

The published empirical studies concerning conjoint additivity have all looked at double cancellation whereas we shall study the weaker Thomsen condition.

The results of the existing studies are inconsistent. Supporting double cancellation are Falmagne et al. (1979), Levelt et al. (1972), and Schneider (1988), where the latter differed from the other studies in having frequencies varying by more than a critical band in the two ears. Rejecting it are Falmagne (1976), with but one respondent, and Gigerenzer and Strube (1983) with 12 respondents. Because of this inconsistency, we re-examine conjoint additivity by testing the Thomsen condition in Experiment 3 (Section 4).

1.3.2. Production commutativity

The basic idea embodied in the representation of generalized ratio production, (6), is that the respondents perform the task as they are told to, using the distortion Ψ of intensities and the distortion W of numbers. An important, easily demonstrated consequence of (6) is the behavioral property called (*subjective*) *production commutativity*: For $p>0, q>0$,

$$[(x,x)\circ_p(y,y)]\circ_q(y,y)\sim[(x,x)\circ_q(y,y)]\circ_p(y,y). \tag{10}$$

Observe that the two sides differ only in the order of applying p,q, which is the reason for the term "commutativity". Proportion commutativity with $y=0$ also arose in Narens' (1996) theory. That hypothesis was sustained in an auditory intensity study by Ellermeier and Faulhammer (2000) and Zimmer (in press). Our work here will focus only on the general case of (10) for which $y>0$.

1.4. The experimental program

Both for reasons of length and because our experimental program evolved with experience and as the theory evolved, we have divided the work into two articles with an additional two nearing completion. This one focuses on the properties that underlie the two representations separately, namely, the Thomsen condition and production commutativity (Experiments 3 and 4). We begin with an experimental test of joint presentation symmetry in order to guide subsequent testing. Note that from the present data alone we cannot conclude that the same function Ψ applies both to summations and productions, but rather there are really two functions $\Psi_{\oplus}$ and $\Psi_{\circ_p}$.

The goal of the second article in the series (Steingrimsson & Luce, in press) is to examine parameter-free properties that permit us to conclude that $\Psi_{\oplus}=\Psi_{\circ_p}=\Psi$. But first in that article we examine a behavioral property called bisymmetry which, if sustained, is equivalent to $\delta=0$ in (5). Its holding renders the rest of the experimental program of that article appreciably simpler than it would be otherwise. We then study two properties (not defined in this article) called simple joint-presentation decomposition and segregation, that link joint presentations and production.

Work on two additional articles is currently underway. The first of these examines several issues including an account of attentional bias in asymmetric matching and the possible form for the function Ψ and properties that are equivalent to these forms. Using some of that we arrive at a test of constant bias, (6), for some respondents.

The second focuses on the distortion function W. We arrive at two possible forms for it, and behavioral properties that characterize each, which we then test.

2. Experimental methods common to all experiments

The experiments reported have a number of testing strategies in common that are now outlined. Other aspects are described later as relevant. In Appendix A, we list some details on suggested methodological improvements that may be of interest to some readers.

2.1. Signal presentations and notation

The experiments were carried out in the auditory domain using 1000 Hz sinusoidal tones in phase presented for 100 ms, which included 10 ms on and off ramps.

The theory is expressed in terms of intensities at or above threshold. That is, for the left ear $x=x'-\varepsilon_{\mathrm{l}}$, with the intensity for the right ear being analogous. However, in our description of methods we report sound pressure levels in decibels, i.e., $x'=x+\varepsilon_{\mathrm{l}}$ in dB SPL. For signals well above threshold and for respondents with little hearing loss, for which they were selected, the error $x_{\mathrm{dB}}-x'_{\mathrm{dB}}$ is negligible.

A safety limit of 85 dB was imposed in all experiments.

2.2. Respondents

A total of 33 students—graduate[5] and undergraduate—from the University of California, Irvine, participated in the four experiments of this article. Of these 33, three respondents stopped for personal reasons before sufficient data had been collected for analysis; three individuals participated in piloting sessions only or in experiments whose data are not reported in full. This leaves 27 individuals, 6 male and 21 female, whose data are reported in full. All respondents were within 20 dB of normal hearing thresholds (ANSI, 1996) in the range 250–8000 Hz, assessed by an audiometric test (Micro Audiometric EarScan ES-AM).

All respondents, except the first author, received compensation of $10 per session. Each person provided written consent and was treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992). Consent forms and procedures were approved by the U.C. Irvine's Institutional Review Board.

2.3. Estimating one- and two-ear matches

The three types of matches used are listed in (2). Let $\langle A,B\rangle$ denote a presentation of A followed by a temporally displaced presentation of B. We used a

[5]Included was the first author (R22). We judged this acceptable because the behavioral measures of matching and ratio production are not determined by the experimental design.

temporal delay between A and B of 450 ms. Three trial types, corresponding to (2), were used.

$$\langle (x,u),(z_{\mathrm{l}},0)\rangle, \tag{11}$$

$$\langle (x,u),(0,z_{\mathrm{r}})\rangle, \tag{12}$$

$$\langle (x,u),(z_{\mathrm{s}},z_{\mathrm{s}})\rangle. \tag{13}$$

That is, respondents heard a tone followed 450 ms later by another tone in the left, right, or both ears. Following the tone presentation, respondents used key presses either to adjust the sound pressure level of z_i, $i = \mathrm{l}, \mathrm{r}, \mathrm{s}$, to repeat the previous trial, or to indicate satisfaction with the loudness match. Respondents could chose between the four sound pressure level adjustments of 0.5, 1, 2 or 4 dB. These were named and presented to respondents as extra-small, small, medium, and large steps. These increments were tied to the keyboard keys "a", "s", "d", and "f" for increasing and ";", "l", "k", and "j" for decreasing of sound pressure. After a sound pressure adjustment, the altered tone sequence was played. The previous trial could be repeated by pressing the "r" key. This process was repeated until respondents were satisfied with the match, indicated by pressing "b" at which time the process ended and the value of z_i was recorded as the response. Decision time was not limited, but a minimum of one second separated all trials.

Information about the current block number and the function of each of the keyboard keys used was displayed on the computer monitor.

In verbal instructions to respondents, the task was explained as that of making the second stimulus equal in loudness to the first. The instructions stressed the importance of paying attention solely to the loudness of the stimuli and ignoring the subjective sense of tone location.

2.3.1. Tone localization

Although solvability (2) is very plausible within the psychophysical context, the actual perception of single ear matches (z_{l} and z_{r}) using headphones is that the tone percept moves from a somewhat head-centered localization towards the matching ear. Although we asked observers to ignore this attribute while judging loudness, we cannot be certain as to how successful they were at doing this. As is true for auditory theory in general, it would be desirable to have a theory that establishes how localization and loudness interact. Although we are not aware of any such theory, we note that if tone localization changes cause substantial distortions in loudness judgments, our tests, hence the theory, would likely fail. Therefore, we prefer symmetric matches over asymmetric ones, as they considerably reduce localization differences.

2.4. Procedure

Experiments were conducted in sessions lasting no more than 1 h. The initial session was devoted to obtaining written consent, explaining the practice task—a matching task as outlined in Experiment 2—and running the practice blocks. Respondents typically ran two to four sessions per week and, with rare exceptions, no more than one session per day. All respondents completed one session of training using a matching task; if they participated in an experiment with ratio production, they also completed one session of practice with that operator. Because some observers participated in multiple experiments, the total practice that individuals had prior to any one experiment varied substantially. The average number of sessions run by respondents from whom data are reported was between 18 and 19, which includes practice, piloting, and experimental sessions. Depending on the experiment, practiced respondents typically completed 60–64 estimates per session, organized into blocks of six or eight estimates. Rest periods were encouraged but their frequency and duration were entirely under the respondents' control.

2.5. Equipment

Stimuli were generated digitally using a personal computer and played through a 16-bit digital-to-analog converter (Quikki; Tucker-Davis Technology) at a conversion rate of 40μ's per sample. Presentation level was controlled by manual and programmable attenuators, and stimuli were presented over Sennheiser HD265L headphones to listeners seated in an individual, single-walled, IAC sound booth.

2.6. Statistical method and presentation of results

The properties to be tested are each a null hypothesis of the form $A = B$, but of course our estimates of A and B are variable. The theory will be judged (tentatively) as supported if these null hypotheses are not rejected. Accepting the null hypothesis is a fairly common problem in testing explicitly formulated mathematical models. It has at least two dangers. One is that we do not deal with a sufficiently large sample of estimates of A and B or of respondents. The other is that experimental artifacts may easily lead one to reject the property being tested. In the course of running these and similar experiments, we did encounter such artifacts and worked out remedies to overcome them. Since, as experimenters, we were attempting to decide on the adequacy of a theory whose behavioral properties are all assertions of indifferences, we wanted to avoid rejecting these "null hypotheses" for irrelevant reasons. We were,

therefore, particularly motivated to root out any such problems, which guided our experimental designs.

Because we do not have a theory that predicts the distributions for these expressions, a nonparametric statistical analysis is appropriate. We chose the Mann–Whitney U-test, with a significance level of 0.05. The Mann–Whitney tests equality of medians by asking whether the two distributions seem to be two samples drawn from the same unknown distribution. To do so, a ranking procedure is followed. The data compared are assumed to be from continuous variables but, in fact, the sizes of the sound pressure level adjustments permitted were discrete. Thus, although intensity itself is a continuous variable, the data involve only a discrete subset of intensities. This requires correction for ties in the ranking procedure as well as using averages as estimates for the median.

To indicate variability in adjustments, the standard deviations are reported except when the results are presented graphically, where error-bars represent the standard error of the estimates.

If the hypothesis $A = B$ is correct, we are asserting that both A and B are drawn from the same distribution. A particular concern is whether the sample sizes for A and for B are sufficiently large so that a failure of the null hypothesis can be distinguished within the power of the statistical method employed.

To address this issue we carried out Monte Carlo simulations based on the bootstrap technique (Efron & Tibshirani, 1993). We assessed the accuracy of our statistical results by simulating repeated experimental data collections and recomputed the associated Mann–Whitney statistic. Specifically, if A and B are indeed drawn from the same distribution, as acceptance of the null hypothesis implies, then all of the collected data combined forms a single estimate of the common underlying distribution. Based on this observation, one creates two samples A_i, B_i, $i = 1, \ldots, n$ by sampling randomly from $A \cup B$ with replacement. For each such pair, one calculates the p_i-value corresponding to the test of the hypothesis that $A_i = B_i$. Under the null hypotheses, these n values of p_i are normally distributed with a mean $\bar{p_i} = \frac{1}{2}$. We can thus estimate a confidence interval, say the 95% one, for that simulated distribution of p_i values. If the null hypothesis is true, then p, of the original test statistic and $\bar{p_i}$ are both instances of the sampling distribution of p_i. This hypothesis can be tested by verifying that both values fall within (or outside) the confidence interval, in which case we accept (reject) the null hypothesis. As long as the actual test-statistics fell within for $p > 0.05$ or outside for $p \leqslant 0.05$ of a 95% confidence interval for the statistic, the test statistic was accepted.

In Experiments 1 and 2, three tests of the same condition are conducted for each participant. When a single hypothesis of no effect is tested using more than one test and the hypothesis is rejected if one of the tests shows statistical significance, the probability of rejecting by chance increases with the number of tests conducted. Hence, it is often appropriate to correct for multiple comparisons, using e.g., a Bonferroni correction.[6] Two factors make the use of such correction problematic in our statistical analysis. First, we aim to accept a number of null hypotheses and a correction for multiple comparisons will make this easier, not harder. Second, we do not strictly reject a property based on a single rejection, but rather the overall pattern of results. For these reasons we will not report results corrected for multiple comparisons. However, we explored the effect such a test would have had on all the experiments conducted. In Experiments 1 and 2 we found a total of four tests out of 93 that would be affected, but in no case would these affect the overall pattern or results.

2.7. Multi-step testing

Testing in Experiments 3 and 4 requires one estimate to be used as stimulus in a subsequent trial. Initially, we used the average (as our best estimate of the median) of the first estimate as the input in the subsequent testing step. However, we realized that using individual estimates as input for subsequent estimates instead of only their averages was a statistical improvement—this latter method allows the variance in the first estimate to propagate through the testing process. This method was incorporated into Experiment 3 and partially into 4.

3. Existence of bias

When this project began, the theory was a reinterpretation of the second author's utility theory which was for an unbiased case (Luce, 2000). So our first effort was to see if, among young people at least, a substantial fraction are unbiased. As will become clear from the first two experiments of this section, bias is the norm. Thus, the theory was expanded to cover that case.

3.1. Experiment 1: bias or not using single-ear matches

The initial aim of this experiment was to explore whether the joint-presentation symmetry $(x, u) \sim (u, x)$ holds by comparing matches in a single ear (asymmetric matching). However, pilot data suggested that bias behavior might be influenced by the choice of matching ear (left or right). Consequently, investigation of this phenomenon became the focus of this experiment. We explore it theoretically elsewhere.

[6] Here that would effectively mean the use of a rejection limit $\alpha = 0.05/n$ where n is the number of tests of each condition.

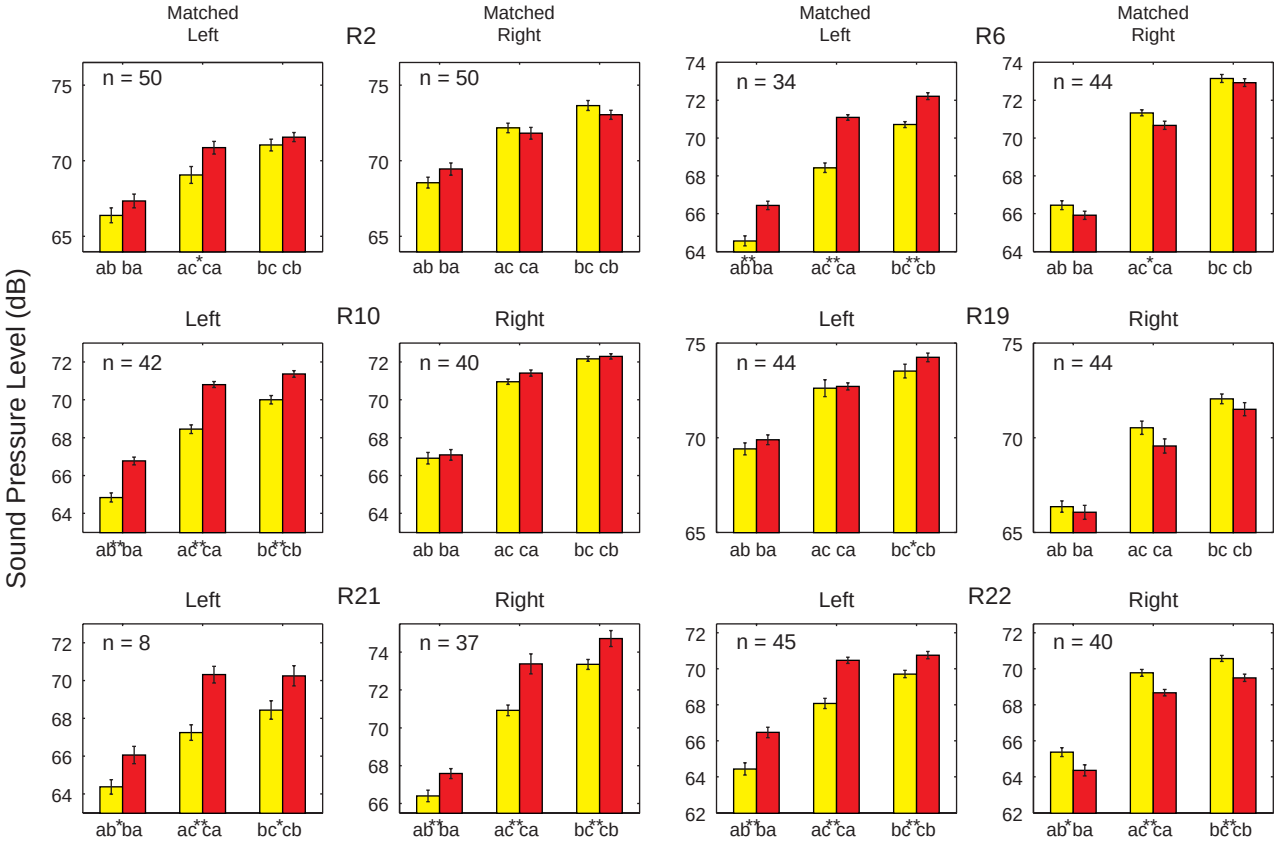

Fig. 1. Experiment 1: bias in single-ear matches (Method 1).

3.1.1. Method

The experiment involves obtaining estimates z_l and z_r using the matches $(x,u)\sim(z_l,0)$ and $(x,u)\sim(0,z_r)$ of (2). Three sound pressure levels were used: $a = 58\,\mathrm{dB}$, $b = 64\,\mathrm{dB}$, $c = 70\,\mathrm{dB}$. These three sound pressure levels give rise to six ordered stimulus pairs: (a,b), (a,c), and (b,c) corresponding to the left side of (1) and (b,a), (c,a), and (b,c) corresponding to the right side of (1). Each of these six stimuli were matched in both the left ear and the right ear. The left and right matches were obtained using the two trial forms (11) and (12), respectively.

Data collection was blocked on the matching ear (Method 1). For two respondents, data were also collected using blocks in which left and right ear matching was alternated (Method 2). Other aspects of the procedure were carried out as outlined in Section 2.

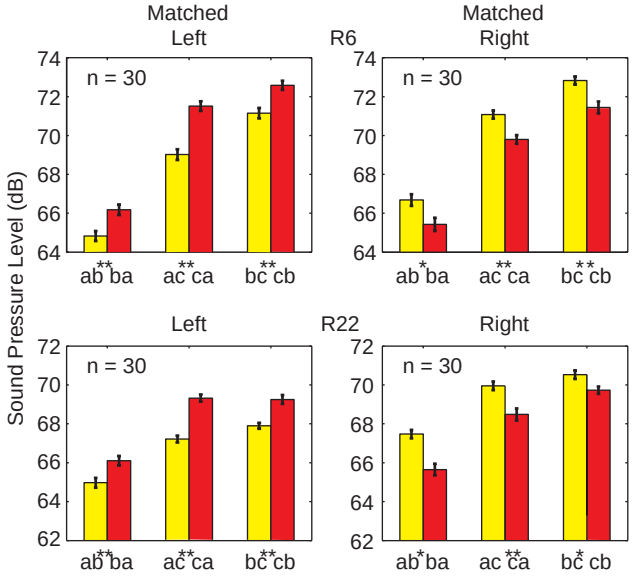

Fig. 2. Experiment 1: bias in single-ear matches (Method 2).

3.1.2. Results

Six individuals participated. The data are presented graphically in Fig. 1; the additional data in which the matching ear was alternated within a block are presented in Fig. 2. Two graphs, one for the left and one for the right ear match, are given for each respondent. In each graph, matching results for stimulus conditions of the form (x,u) are labeled xu and for (u,x) are labeled ux.[7]

[7]Care must be taken not to confuse the notation xu for multiplication.

The statistical hypothesis is that $(x,u)\sim(u,x)\Leftrightarrow xu=ux$. Hence, for each ear there are three statistical hypotheses to be tested, namely whether $ab=ba$, $ac=ca$, and $bc=cb$, which are marked on the abscissa. Average sound pressure level is marked on the ordinate. Sample size is indicated in the upper left portion of each graph.

The result of the statistical test is indicated on the abscissa, above the label of the relevant conditions, with $\star$ denoting rejection at the 0.05 level, $\star\star$ at the 0.01 level, and unmarked meaning apparent acceptance.

For R10 a consistent left bias is obtained in left ear matching. However, no statistically significant bias is found in the right ear matching although visual inspection reveals the trend in the data to be toward left bias. For R2 a statistically significant left bias is obtained in one of three conditions in left ear matching with the other two showing a consistent trend towards left bias. Conversely, for right ear matching, no statistical significant bias is obtained and the trend appears mixed.

All respondents exhibit evidence of left bias when matching in the left ear (for R21, the number of observations for left ear matching suffices only to indicate a trend). Two individuals, R6 and R22, show a shift to a right bias for right ear matches. R2, R10 and R19 show no bias when matching in the right ear although the pattern of data suggests a trend towards left bias for R10, and a bias shift for R19. Only R21 shows no clear change in bias based on matching ear.

As evidenced in Fig. 2, including both left and right ear matches within a block did not alter the results for R6 and R22—the bias shift is larger, if anything, for R6.

3.1.3. Discussion

With respect to the statistical testing, the results may be divided into two categories:

(1) $z_l\leqslant z'_l$ and $z_r\leqslant z'_r$ (R2, R10, R19, and R21),
(2) $z_l<z'_l$ and $z_r>z'_r$ (R6 and R22).

The results in the second category are not described within Luce's (2002; 2004) theoretical framework. If trends in the data are taken into account, R19 may also belong to the second category. In fact, only R21 shows a statistically consistent bias regardless of matching ear. Hence, there is substantial evidence of some sort of change in bias depending on the matching ear. These results were unexpected and we are not aware of similar results in the psychoacoustic literature. Hence, it is prudent to look for methodological explanations.

Sessions were blocked on matching ear (Method 1). We thought this repeated matching in the same ear might play a role in bringing about the unexpected results. Hence, we ran the two individuals who clearly showed the bias shift using blocks in which the matching ear alternated within a block (Method 2). In short, the pattern of results did not change (Fig. 2).

We also considered the use of a 2AFC paradigm: The stimulus $\langle(x,0),(0,x)\rangle$ is presented and the respondent's task is to judge which tone, the first or the second, is louder. Assuming additional design features such as counter-balancing of presentation order, an unbiased person is in principle equally likely to choose either tone. Using this approach, tones are heard both in the left and the right ear within a trial. However, pilot data using this method quickly showed that the time-order error led respondents almost always to judge the second tone louder than the first regardless of condition. See Appendix A for details on the time-order error.

A similar procedure, using matching, is to present a tone in a single ear but match it in the opposite ear. We have not yet attempted this.

A clear consequence of these findings is that an experimental procedure that relies on bias being constant, regardless of matching ear, is not reliable. For instance, testing for jp-symmetry, $(x,u)\sim(u,x)$, yields two different results depending in which ear the matches are made. However, any other single-ear matching procedure is not problematic with respect to these results.

The phenomenon of bias shift depending on the matching ear was quite unexpected to us and the results appear sufficiently robust that they warrant further theoretical development, which we are currently carrying out.

3.2. Experiment 2: bias or not using two-ear matches

In light of the results of Experiment 1, we explored whether the jp-symmetry, (1), $(x,u)\sim(u,x)$, holds using symmetric matching.

3.2.1. Method

The two matches $(z_s,z_s)\sim(x,u)$ and $(z'_s,z'_s)\sim(u,x)$ were obtained. It is trivial to show that the property can be judged to hold as long as z_s and z'_s are found statistically indifferent. We used the same stimuli as in Experiment 1. The trial type used is given by (13). The six matching conditions were all run within a block of trials.

3.2.2. Results

Fifteen individuals participated. Their data are presented graphically in Fig. 3 in a manner analogous to Fig. 1.

For 12 of 15 respondents, one or more of the three conditions were found to be statistically different, with the trend for the remaining conditions consistent with this statistical difference. In 10 cases (Rs 2, 4, 5, 10, 11, 14, 16, 17, 22, 23), the pattern of results was that of $(x,u)\prec(u,x)$, and in two cases (Rs 9 and 24) the

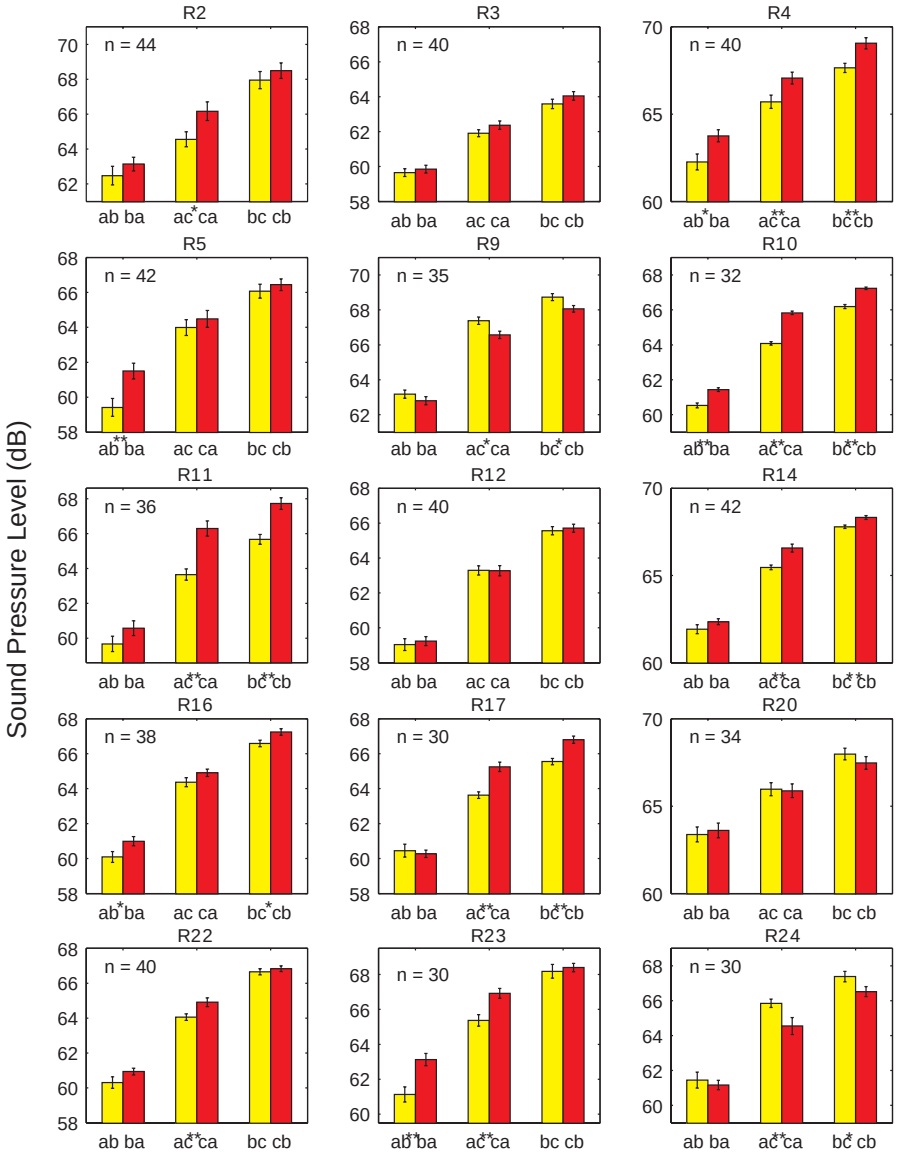

Fig. 3. Experiment 2: bias in two-ear matches.

opposite pattern was obtained, i.e., $(x,u) \succ (u,x)$. For three respondents (Rs 3, 12, 20) the hypothesis of no bias, i.e., $(x,u) \sim (u,x)$, was not rejected.

3.2.3. Discussion

In the terms introduced earlier, 10 respondents exhibited left bias, two right bias, and three possibly no bias, i.e., joint-presentation symmetry.

The results suggest that jp-symmetry does not hold for at least 12 of the 15 respondents. So, as a general rule, this symmetry property is rejected. The finding of three respondents for whom jp-symmetry was not rejected may mean that some people truly are unbiased or it may mean that the deviations are simply too small to be detected with the number of observations collected.

Luce's (2002; 2004) theory admits bias in either direction, but the theory makes no attempt to explain the proportions of people who are left and right biased. Dominance of side (left vs. right) is common human feature. Beyond the familiar handedness, most people exhibit eye dominance and Coren (1992, pp. 27–33) reports a similar phenomenon for the ears. Whether earedness plays a part in the bias we have observed is an open question.[8]

[8]Earedness—the tendency to prefer one ear over another in such tasks as talking on the telephone or listening for sounds through a wall—is observed in the population with one study showing about 60% to be right eared (right handedness is observed in ca. 90%). Earedness is similar to eyedness in that the most sensitive ear (as determined by, e.g., a hearing test) is not always dominant (Coren, 1992, pp. 27–33).

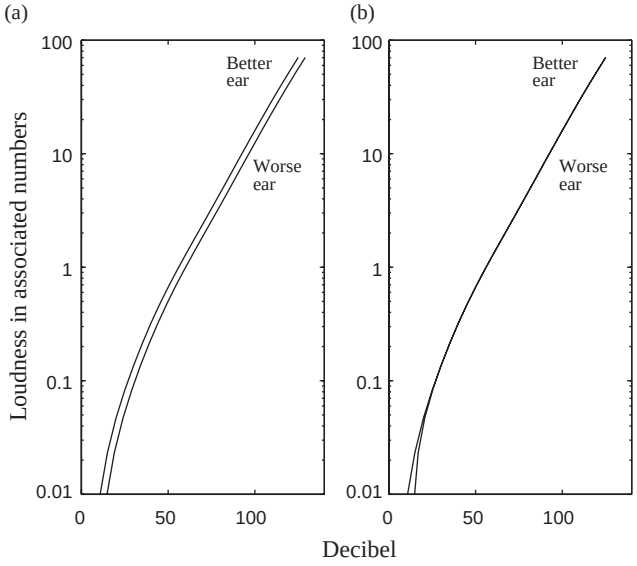

Fig. 4. (a) Recreates two curves from Hellman and Zwislocki (1963, Fig. 12), i.e., $10\log\psi_b(x')$ and $10\log\psi_w(x')$ as a function of x' in dB SPL and under the assumption that the two function differ by a constant dB term. (b) shows the two functions are assumed to differ by a constant intensity term and we plot $10\log\psi_b(x-\varepsilon_b)$ and $10\log\psi_w(x-\varepsilon_w)$ as a function of dB SPL.

R.P. Hellman (pers. comm., December 2001) suggested that the failures of jp-symmetry are caused by a difference in threshold levels for the two ears. She cited the empirical results of Hellman and Zwislocki (1963) in support of this view.

Those results are embodied in their Fig. 12, whose relevant portion we have reproduced in Fig. 4a. This concerns the two curves labeled "better ear" and "worse ear", respectively, where *better* and *worse* refers to the respective threshold sensitivity of the two ears.

Hellman and Zwislocki (1963) collected monaural magnitude estimates (their Table 1) and fitted those data to the better-ear curve shown in Fig. 4a.[9]

If we denote by ψ_b the mean magnitude estimates of the better (i.e., more sensitive) ear fitted to a smooth curve, ρ is a dB constant representing average threshold difference between the two ears, and x' is the signal intensity in physical terms, then the second function they plot is ψ_w, the magnitude estimate of the worse (i.e., less sensitive), which they assume to be

$$10\log\psi_w(x') = 10\log\psi_b(x') + \rho.$$

Thus, their two curves represent the plots of $10\log\psi_b(x')$ and $10\log\psi_w(x')$, respectively, against x' in decibels, estimating that $\rho = 4\,\mathrm{dB}$.

[9]We do not know what fitting procedure they used, but for our own reproduction we used the data points provided in their Table 1 and used a polynomial fitting algorithm to recreate their smooth curve. The results appear, for our purposes, identical.

Indeed, when the monaural data for the better and worse ear are plotted under the assumption of the two differing only by a dB-constant, there does seem to be a correlation between the threshold sensitivity of each of the two ears and their sensitivity at well above threshold intensities.

We consider their plot to be problematic because the difference in ear sensitivity at threshold represents a vanishingly small portion of the total sound pressure energy at, e.g., normal speech level. Thus, the assumption that the decibel difference at all intensities is constant implies a proportional effect on the transformation of energy into sensation that is equal to sensitivity differences at threshold.

Recall that in the present theory we have formulated things in terms of intensity x less the threshold ε, and so if ε_b and ε_w denote the better and worse thresholds, respectively, then threshold sensitivity differences can be expressed as $\varepsilon_b - \varepsilon_w = \rho'$, where $10\log\rho' = \rho$. Thus, in our view, the two curves should be plotted under an assumption of constant differences in intensities, or in the same decibel terms as above

$$10\log\psi_b(x+\varepsilon_b) - 10\log\psi_w(x+\varepsilon_w) = \Delta(x).$$

What is immediately obvious is that $\Delta(x)$ approaches 0 for large x, but as x approaches 0 from above, it approaches $10\log\psi_b(\varepsilon_b) - 10\log\psi_w(\varepsilon_w) = \rho$.

Fig. 4b replots the two curves of Fig. 4a under the assumption of constant differences in intensity, that is $10\log\psi_b(x-\varepsilon_b)$ and $10\log\psi_w(x-\varepsilon_w)$, respectively, against x in dB and shifted such that the first curve of Fig. 4a (better ear) coincides with the first in Fig. 4b.

The replot in Fig. 4b suggests that the threshold difference between the two ears does not result in any material way the sensitivity difference between the ears except close to threshold. This result strikes us as both inconsistent with Hellman's conjecture and with our data that showed asymmetries for signals well above thresholds. This suggests that there may be no correlation at all between bias between the two ears and threshold differences.

Further doubt is cast on Hellman's view by the shift in bias direction depending upon the matching ear that we found for some respondents in Experiment 1. In those cases, differences in threshold levels are certainly not sufficient to explain the data obtained.

4. Tests of predictions of summations and productions separately

Recall that the testable properties that we have arrived at so far are the Thomsen condition, (9), and production commutativity, (10). They have the feature that each follows from just one of the two representations, the former from the representation of summation,

(5), and the latter from the subjective proportion representation, (6).

4.1. Experiment 3: Thomsen condition

The goal of this experiment is to test the well-known necessary condition of binary additive conjoint representation (Krantz et al., 1971; Michell, 1990, pp. 68–73), the Thomsen condition, (9),

$$\left.\begin{array}{l}(x,t)\sim(z,v)\\(z,u)\sim(y,t)\end{array}\right\}\Longrightarrow(x,u)\sim(y,v).$$

As mentioned earlier, the results of the existing studies are inconsistent. Of these the study by Gigerenzer and Strube (1983) was the most extensive. Our methodology is most similar to that study. For this reason we made methodological choices facilitating a direct comparison with their results.

4.1.1. Method

With reference to (9) and the notation of (13), the property is tested by successively obtaining the estimates, z', y' and y'' using

$$\langle(x,t),(z',v)\rangle,$$
$$\langle(z',u),(y',t)\rangle,$$
$$\langle(x,u),(y'',v)\rangle.$$

The property is said to hold if y' and y'' are found to be statistically indifferent.

All three trial types were run twice within a block in a pseudo-randomized order. Individual estimates of z' were used for subsequent estimates of y'. (The software insured that the second trial type was run only when a prior estimate of a z' was available).

Two stimuli sets were used:

A: $x = 66,\ t = 62,\ v = 58,$ and $u = 70$ dB,

B: $x = 62,\ t = 59,\ v = 47,$ and $u = 74$ dB.

The two stimuli sets were run in separate sessions, with data for stimulus set A collected first. For R29 the A, B pair was run a second time (see discussion) and the latter result is reported.

4.1.2. Results

Result for 12 respondents are displayed in Table 1. In the table, averages, standard deviations, number of observations, and statistical results are listed for each respondent. Sound pressure levels are in dB SPL.

The property is found to hold in 19 of 24 tests; however, the failures were primarily within set A (four of five) with only one within set B. Respondent R29 failed the Thomsen condition for the first presentation of stimulus set A, but it was accepted on the second presentation (see discussion below).

Relevant to the discussion, we find no systematic biases between y' and y'' in the data, namely $y' > y''$ in 14 of 24 cases, i.e., 58%.

Table 1
Experiment 3: Thomsen condition

Resp.	Stim. set	y' (s.d.)	y'' (s.d)	n	p_{stat}	Stat. concl.
R10	A	69.03 (1.25)	69.85 (1.01)	40	.007	$y' \neq y''$
	B	72.35 (1.23)	72.67 (1.04)	30	.468	$y' = y''$
R22	A	71.30 (1.10)	70.88 (0.93)	30	.127	$y' = y''$
	B	72.08 (1.21)	71.88 (1.01)	30	.311	$y' = y''$
R23	A	71.82 (1.20)	71.57 (0.81)	30	.164	$y' = y''$
	B	72.88 (1.05)	73.62 (1.11)	40	.005	$y' \neq y''$
R25	A	72.60 (3.59)	72.00 (2.39)	30	.312	$y' = y''$
	B	76.58 (2.26)	75.59 (1.96)	40	.062	$y' = y''$
R26	A	69.50 (1.14)	69.74 (0.99)	39	.267	$y' = y''$
	B	72.65 (1.21)	72.39 (1.26)	41	.264	$y' = y''$
R27	A	69.48 (1.15)	69.36(1.88)	32	.912	$y' = y''$
	B	71.66 (2.17)	72.19 (2.06)	40	.107	$y' = y''$
R28	A	72.19 (0.98)	71.63 (0.85)	40	.009	$y' \neq y''$
	B	73.75 (1.24)	73.57 (1.53)	54	.270	$y' = y''$
R29	A	70.20 (2.32)	69.58 (1.73)	40	.123	$y' = y''$
	B	72.78 (1.82)	72.24 (1.68)	60	.084	$y' = y''$
R30	A	70.70 (2.69)	72.00 (1.71)	40	.013	$y' \neq y''$
	B	74.06 (3.05)	74.84 (3.39)	40	.139	$y' = y''$
R31	A	69.91 (1.39)	70.21 (0.72)	40	.154	$y' = y''$
	B	70.65 (1.25)	70.90 (0.94)	40	.353	$y' = y''$
R32	A	72.71 (1.56)	71.66 (1.13)	38	.001	$y' \neq y''$
	B	74.08 (2.35)	73.94 (1.29)	60	.645	$y' = y''$
R33	A	73.48 (1.87)	73.75 (1.39)	40	.253	$y' = y''$
	B	69.42 (2.85)	70.15 (1.84)	60	.115	$y' = y''$

4.1.3. Discussion

As noted earlier, of five published studies concerning double cancellation, a close relative of the Thomsen condition, three accepted and two rejected the property. Of these Gigerenzer and Strube (1983) was the most comprehensive, with 12 respondents. Their experimental and statistical methods were similar to ours.[10] Notably, the stimulus set B was chosen to have stimuli with the same relative intensity relationship as those used by Gigerenzer and Strube (1983), although we used a 1000 Hz tone whereas they used both 200 Hz and 2000 Hz tones—a fact that may be relevant.

Gigerenzer and Strube (1983) rejected the condition in 40 of 48 cases. Not only did they conclude that $|y' - y''| \neq 0$ but they argued that there is a systematic relationship, namely, $y' > y''$ (33 of 48 cases, i.e., 69%). In contrast, we find that the Thomsen condition is rejected in five of 24 cases of which four were in set A and one in set B, and find no evidence for $y' > y''$ (14 of 24 cases, i.e., 58%). Thus, our results provide considerably stronger evidence favoring the Thomsen condition.

The question is what factors of methodological nature may have given rise to the difference.

- *Median versus individual judgments in step* 1.

Gigerenzer and Strube (1983) used median estimates for y', whereas we used each individual observation. Doing this meant two things. First, the variability of the first estimates affected later ones. Second, it avoided the bias that is necessarily introduced using a point estimate. Indeed, Gigerenzer and Strube (1983) observed a systematic bias in their results whereas we do not.

- *Stimulus sets A and B differ in range.*

The intensity ranges were 58–70 dB for A and 47–74 dB for B. The larger intensity differential in B than A will, when using headphones, result in larger subjective changes in tone-localization between successive presentation. While we, as in innumerable other studies, do not address this issue directly, we note that if this phenomenon caused problems in the current study, it should affect B more adversely than A, which may account for variability being somewhat higher in B for eight of 12 respondents. However, the overall variability is neither unusual nor dramatic so it is not easy to determine whether it could explain the statistical evidence of fewer failures in B compared to A. Gigerenzer and Strube (1983) make essentially the same observation coming to a similar conclusion. Hence, the difference between sets A and B does not seem to explain why more rejections were found for A than for B.

- *The possibility of inadequate practice.*

Because the Thomsen condition was rejected in fewer of the B cases than in A cases and they were run in the order A, B, possibly the practice was inadequate when A was run. Our initial practice experience varied from 90 to 120 trials (the lower number for experienced observers); then we collected 30–60 observations for each of z', y', and y''. Hence, our respondents had completed no less than 180 matching trials prior to data collection for stimulus set B. Related to this conjecture are data from R29 who ran and failed the test for A, then ran B, where the property held, and subsequently ran multiple sessions of both A and B in which the property was consistently accepted.

One reason practice might be more important in this experiment than in others relates to tone localization changes (see Section 2). When both tones are adjusted equally, the subjective location of the tone moves on a line either away from or towards a head-centered place. However, with only one tone being changed, this location moves in an approximate horizontal plane. Some respondents indicated they found it harder to ignore the latter than the former effect. Perhaps practice makes it easier to ignore these changes.

Although Gigerenzer and Strube (1983) state that respondents engaged in extensive initial practice and for experimental trials they collected 41 observations for z' and 20 for y' and y''. Without additional information about the order and practice their respondents experienced, it is impossible to compare their level of practice with our procedures.

- *Differences in signal frequency.*

Gigerenzer and Strube (1983) collected data at 200 and 2000 Hz with stimuli having the same intensity relationship as our stimulus set B but at two different base levels, 20 dB apart. These four conditions were run separately in an unknown order. They rejected double cancellation in 12 of 24 tests at 200 Hz, and 18 of 24 at 2000 Hz. In all of our experiments, we presented the often used 1000 Hz signals. We see no immediate reason why the Thomsen condition should hold for 1000 Hz, but not for 200 and 2000 Hz. However, absent further information, this is an empirically open question.

- *In conclusion*

In this analysis, we do not find any methodological flaws in our procedures. Indeed ours may in some respect be an improvement on those used by Gigerenzer and Strube (1983), which is possibly enough to explain the

[10] They employed a statistical method suggested by Dr. Ulrich Raatz whose results are identical to that of the Mann–Whitney U-test (Raatz, pers. comm., August, 2002).

discrepancy in results. Notably, Gigerenzer and Strube (1983) found a systematic bias in their data, a bias we do not replicate.

This discussion suggests areas for further exploration of the property. For example, it would be desirable to collect data where stimulus sets are interwoven, to explore the changes in estimates over time, and to assess the role of stimulus frequency on the results. However, these manipulations do not in any obvious way diminish the support we find for the property, hence we do not feel compelled to carry out these additional experiments at this time.

In conclusion, our results lead us to accept the Thomsen property as a good working hypothesis.

4.2. Experiment 4: production commutativity

Recall that the property of production commutativity, (10), using the respondent-preferred[11] two-ear production, is given by

$$[(x,x)\circ_p(y,y)]\circ_q(y,y)\sim[(x,x)\circ_q(y,y)]\circ_p(y,y).$$

Appendix B shows that using the two-ear version does not reduce generality.

4.2.1. Method

The testing requires four estimates in two steps. The first consisted of estimating v and then w in

$$(x,x)\circ_p(y,y)\sim(v,v),$$
$$(v,v)\circ_q(y,y)\sim(w,w),$$

and the second of estimating v' and then w' in

$$(x,x)\circ_q(y,y)\sim(v',v'),$$
$$(v',v')\circ_p(y,y)\sim(w',w').$$

The property is considered to hold if w and w' are found to be statistically equivalent.

Employing the notation defined in Section 2.3, the basic trial form is $\langle\langle A,B\rangle, \langle A,C\rangle\rangle$ where $\langle A,B\rangle$ and $\langle A,C\rangle$ represent the first and the second intensity interval, respectively. The temporal delay between $\langle A,B\rangle$ and $\langle A,C\rangle$ was 750 ms, and between A and B (and A and C) the delay was 450 ms. The longer delay between the pairs was introduced to create a subjective sense of two distinct intervals.

Then, the first estimate, where v is under respondents' control, is obtained using the trial type

$$\langle\langle(y,y),(x,x)\rangle, \langle(y,y),(v,v)\rangle\rangle. \quad (14)$$

In practice, observers heard two tones separated by 450 ms, representing the first loudness interval. After 750 ms, another set of two tones, separated by 450 ms, was heard, representing the second loudness interval. The first tone in both intervals is the same and quieter than the second tone. Respondents controlled the loudness of the second tone in the second interval.

The value of the proportion p was displayed on the monitor prior to the onset of a new ratio production and then remained there until the estimate was completed. For instance, with $p=2$, the phrase "Proportion is 2" was displayed. The other estimates were performed using trials analogous to that described by expression (14). Other aspects of this process were identical to that for joint presentations, described in Section 2.3.

The sound pressure levels $y=64$ dB and $x=70$ dB and the proportions $p=2$ and $q=3$ were used, giving rise to four trial condition in each step.

Instructions to respondents took the form of a verbal description of the task coupled with graphical examples. Respondents were told that they would hear four tones and that the tones formed two loudness intervals separated by a time delay. On paper, a coordinate system was drawn with intensity indicated on the ordinate. The first interval was represented by a line segment starting at $y>0$. Using $p=2$ as an example, it was explained that the task was to produce a second interval that was twice as loud as the first. This interval was depicted as a line segment twice as long as the first and having the same starting point. Comparable examples were given for $p=\frac{2}{3}$ (note that $p=\frac{2}{3}$ was not a stimulus condition, it was only used as an example) and $p=3$.

For R22 and R25, the averages of v and v' were used in the subsequent estimation of w and w'. The trials from each step were organized into separate blocks each containing two instances of each trial condition. Both block types were run sequentially within a session.

For R27 and R28 we switched to using the individual estimates v and v' in the subsequent estimates of w and w'. (See Section 2.7 and Appendix A for details). The two estimation steps were carried out within a block of trials, hence a block contained all eight trial conditions. Trial order was pseudo-randomized in such a way that w or w' was only estimated if a prior instance of v or v' was available.

4.2.2. Results

Four respondents completed this experiment. Their data are presented in Table 2.

In the table, T_1 and T_2 stand for the means of w and w' respectively; standard deviations are given in parentheses. All sound pressure levels are given in dB SPL. The p-values indicate statistical test results.

The property was not rejected for any of the four respondents.

4.2.3. Discussion

Ellermeier and Faulhammer (2000) and Zimmer (in press) investigated the related property, threshold-

[11] Not only did we, as experimenters, prefer it, but it was uniformly preferred by respondents during pilot studies.

Table 2
Experiment 4: proportion commutativity

Respondent	Mean (s.d.)		p_{stat}	n	Statistical conclusion
	T_1	T_2			
R22	77.43 (1.56)	77.70 (1.86)	.574	30	$T_1 = T_2$
R25	79.90 (2.21)	80.17 (2.32)	.662	30	$T_1 = T_2$
R27	77.57 (2.70)	78.48 (2.83)	.102	30	$T_1 = T_2$
R28	74.15 (4.55)	74.41 (1.98)	.301	30	$T_1 = T_2$

production commutativity, namely,

$$[(x,x)\circ_p(0,0)]\circ_q(0,0) \sim [(x,x)\circ_q(0,0)]\circ_p(0,0),$$

which is the special case of (10) in which $y = 0$. They used an experimental paradigm and stimuli similar to those employed in the present study. Ellermeier and Faulhammer (2000) tested the property using $p, q > 1$ and Zimmer (in press) $p, q < 1$ and both found it to hold.

A theoretical prediction of production commutativity is that the property holds for both $p, q < 1$ as well as $p, q \geqslant 1$. The property remains to be tested for $p, q < 1$ and neither it nor its threshold version have been tested for $p < 1 < q$.

5. Conclusions

The topic has been a theory of global psychophysical judgments leading to the two representation classes. For asymmetric matches, the theory leads to representations satisfying the following three properties:

$$\Psi(x,u) = \Psi(x,0) + \Psi(0,u) + \delta\Psi(x,0)\Psi(0,u) \quad (\delta \geqslant 0), \tag{5}$$

$$W(p) = \frac{\Psi[(x,u)\circ_p(y,v)] - \Psi(y,v)}{\Psi(x,u) - \Psi(y,v)} \quad [(x,u) \succ (y,v) \gtrsim (0,0)], \tag{6}$$

$$\Psi(x,0) = \gamma\Psi(0,x) \quad (\gamma > 0). \tag{7}$$

For symmetric matches, (5) with $\delta = 0$ and (6) both hold, but the theory for symmetric matches does not predict the constant bias of (7).

These representations have a number of necessary consequences (behavioral properties) that in turn are sufficient under certain structural conditions to give rise to the representations. The focus of this article has been the separate testing of the properties derived from the first and second expression. Note that the results of the present paper can, at best, support (5) and (6) with different psychophysical functions, $\Psi_{\oplus}$ and $\Psi_{\circ_p}$. Our overall conclusion from the four experiments here is that the summation and production forms of Luce's (2002) theory are separately supported in the auditory domain.

Table 3
Summary of experimental results

Ex. #	Name	#R	#Tests	#Fail	Comment
1	Bias-1 ear	6	48	33	L-R inconsistency
2	Bias-2 ear	15	45	23	
3	Thomsen cond.	12	24	5	Extra practice for 1 R
4	Prop. comm.	4	4	0	

We have not, as yet, presented the evidence supporting the hypothesis that the same function Ψ holds in each case. We do so in Steingrimsson and Luce (in press).

5.1. Summary of main results

The test results are summarized in Table 3.

Experiments 1 and 2 strongly establish that for joint presentations to the two ears, jp-symmetry, (1), does not generally hold, and a majority of the respondents were left biased. This result dictated both further theory construction and experimentation towards testing the property of the biased theoretical solution of Luce (2002, 2004).

More specifically, single-ear matching of Experiment 1 found two of six respondents reversed the bias direction when the matching ear was changed. This is not predicted by the theoretical framework, nor have instances of this behavior been reported in the psychoacoustic literature. Changes in the experimental design aimed at eliminating a plausible procedural explanation had no effect on the results, making it less likely to be an artifact of the experimental design. This empirical result warrants further theoretical development, which is currently underway. We were able to bypass this problem for bias testing here by using the two-ear matching of Experiment 2.

The summation representation implies the Thomsen condition (Section 1.3.1 and Experiment 3). As noted, the literature is split on the support for this property. We found more evidence for it than against it. Furthermore, data from Gigerenzer and Strube (1983) suggest that it would be prudent to investigate the property at several frequencies.

The subjective proportion representation, (6), implies production commutativity, which we tested for $p, q > 1$ (Experiment 4) and found to hold. Threshold-production commutativity was established by Ellermeier and Faulhammer (2000) for $p, q > 1$ and Zimmer (in press) for $p, q < 1$. Production commutativity has yet to be tested for $p, q < 1$ and neither has been tested with $p > 1 > q$.

5.2. Further work

As we noted earlier, the unexpected reversal of bias direction when the matching ear is changed bears further study. And production commutativity should be studied with $p>1>q$.

Current results suggest several new avenues of research. We mention two:

- The preliminary success of the present theory in the auditory domain, warrants extending the work to other domains and/or to other interpretations of the summation and production operators. Currently, the first author is carrying out a parallel study of brightness perception in which luminance is varied across the two eyes.
- It is well-documented that factors such as hearing sensitivity and loudness sensation are affected by signal frequency (see, e.g., Stevens, 1975, p. 97). We consequently ran our tests using a fixed, 1000 Hz, frequency stimulus. In discussing tests of the Thomsen condition (Experiment 3) we noted the possibility, although not the expectation, that stimulus frequency might play a role in explaining the difference between our results and those of Gigerenzer and Strube (1983). Luce's theory makes no predictions about response variation across frequency, but in pilot data we collected using 200, 1000, and 6000 Hz we observed consistent and significant left bias across all three frequencies conditions for two people, right bias in the 200 Hz condition, left bias in the 1000 Hz condition, and no significant bias in the 6000 Hz condition for one person, and no consistent bias at all for one person. With this degree of inconsistency, the sample is far too small to reach any conclusions, but it raises a question of empirical interest: are aspects of the bias observed in Experiment 2, such as direction and magnitude, and other response patterns independent of signal frequency?

Acknowledgments

This research was supported in part by National Science Foundation Grant SBR-9808057 to the University of California, Irvine. Additional financial support was provided by the School of Social Sciences and the Department of Cognitive Sciences at UC Irvine. We are especially grateful to Dr. Bruce Berg for unfettered access to his laboratory, for technical assistance, and for help resolving a number of issues concerning psychoacoustical methodology. We appreciate many helpful comments of Dr. Joetta Gobell on earlier versions. We would also like to thank ones, several anonymous ones and Donald Laming (as a reviewer), for their many valuable and thoughtful suggestions for improvement of this paper.

Appendix A. Methodological improvement and recommended procedures

Over the 2 years during which the experiments reported here and in Steingrimsson and Luce (in press) were performed, plus more not reported here (see Steingrimsson, 2002), we improved our methodology in various ways. These improvements were incorporated as we realized their value. We have included the methodological elements that are directly relevant to the current article in the main text, but the following additional points may be of interest to some experimenters.

A.1. Method of adjustment

In all experiments we use a free-adjustment method where respondents can increase and decrease intensity until they are satisfied with the result. Initially, we used the adaptive procedure PEST (Taylor & Creelman, 1967; Findlay, 1978) to obtain loudness estimates. However, a comparison of results using PEST to the free-intensity adjustment method showed the results to be comparable, and the free-adjustment method required substantially less experimental time than PEST. Moreover, the five individuals who experienced both methods uniformly preferred the free-adjustment method.

A.2. Issues with production judgments

- In constructing trials involving $\circ_p$ care should be taken to chose sound pressure levels such that no intensity interval becomes so small that the respondents have a problem perceiving it. Experience from experiments not reported here (see Steingrimsson, 2002) suggests that, where this is unavoidable, explicit instructions to the respondents should be provided about the possibility of this situation.
- Pilot data from Experiment 4 revealed that loudness productions show—relative to matching—substantial inter-session variability. Therefore, it is desirable that when testing a particular property that all needed ratio production conditions be collected within the same session. Of course, when data for these conditions are pooled across sessions, inter-session effects will manifest themselves in increased variability. However, for a statistical test such as the Mann–Whitney, only intra-session variability is of statistical concern.

A.3. Asymmetric matches and localization

Results from Experiment 1 suggest that the bias is not independent of matching ear. This fact must be kept in mind when designing experiments. This means that any experimental procedure that relies on bias being constant independent of matching ear is not reliable. In some cases, the issue can be avoided by using symmetric

matches as was the case in Experiment 2, where we were able to use symmetric matches to assess bias. However, as evidenced by Appendix B, considerable care must be taken in determining how to test properties using symmetric matches.

A.4. Variance propagation

Some experiments require the use of an estimate from one step as input to a second step. If a median or an average from the first step is used as the input in a second step, then whatever error it contains is necessarily carried over into that second step, but all information about the variance is lost. The statistical test used, the Mann–Whitney, provides no way to include information about accumulated variance and bias, an unfortunate feature when estimates are made using prior estimates.

Steingrimsson (2002) reported data strongly suggesting, not surprisingly, that the variances in production judgments tend to be larger that those arising in matching. Thus, when testing properties involving production judgments, variance propagation—or lack of it—is a greater nuisance than with those properties that just involve matching judgments.

A standard approach to deal with variability is to increase sample size and this is certainly an avenue. However, another option is simply to use individual data points and let them "propagate" through the data collection steps. Then the errors of the first estimate are carried into the subsequent step and average out there, while the variance information is preserved throughout the estimation process.

We followed this practice in Experiment 3 and for some respondents in Experiment 4.

A.5. Time-order errors

A potentially problematic time-order effect is one that is exemplified by respondents' tendency to judge the second tone in the pair $\langle (x,x),(x,x)\rangle$ louder than the first (Hellström, 1985; Stevens, 1975, pp. 139–140). In practice, however, it is only a problem when for a statistical test of $A = B$, the error cannot be assumed approximately equal or insignificant in both A and B. Indeed, pilot experiments suggest that for most respondents, the magnitude of the bias caused by the time-order error is constant in decibels across frequencies. In cases where this bias is not constant, complete counterbalancing is needed. This did not seem necessary for the present experiments. However, it should be noted that even with counterbalancing, time-order errors can be a great nuisance. For instance, related to Experiment 1, we did a pilot study where respondents heard $\langle (x,0),(0,x)\rangle$, in a fully counterbalanced fashion, but the results seemed to be completely dominated by the time order such that the second tone was almost always experienced as louder than the first. Likewise, a pilot experiment, where counterbalancing was achieved through equal numbers of adjustments of the first and second tone, produced results with a very large variance, requiring substantially larger sample size than we normal use.

Appendix B. Testing production commutativity using two-ear matches

Our goal is to show that if (6) holds, then we may test production commutativity using two-ear matches. With no loss of generality [see remark leading to (6)] we can take the commutativity property in the form

$$[(x,x)\circ_p(y,y)]\circ_q(y,y) = [(x,x)\circ_q(y,y)]\circ_p(y,y). \tag{B.1}$$

Define

$$(x,x)\circ_p(y,y)\sim(v,v), \tag{B.2}$$
$$(v,v)\circ_q(y,y)\sim(t,t), \tag{B.3}$$
$$(x,x)\circ_q(y,y)\sim(w,w), \tag{B.4}$$
$$(w,w)\circ_p(y,y)\sim(t',t'). \tag{B.5}$$

We show that (B.1) is equivalent to $t = t'$. Using

$$\Psi[(x,x)\circ_p(y,y)] - \Psi(y,y) = [\Psi(x,x) - \Psi(y,y)]W(p), \tag{B.6}$$

we have

$$\Psi(t,t) - \Psi(y,y) = \Psi[(v,v)\circ_p(y,y)] - \Psi(y,y) \tag{B.3}$$
$$= [\Psi(v,v) - \Psi(y,y)]W(q) \tag{B.6}$$
$$= [\Psi(x,x) - \Psi(y,y)]W(q)W(p). \tag{B.2}$$

Similarly, using (B.5), (B.6), and (B.4),

$$\Psi(t',t') - \Psi(y,y) = [\Psi(x,x) - \Psi(y,y)]W(p)W(q).$$

By the commutativity of multiplication

$$\Psi(t,t) = \Psi(t',t')$$

and so by the strict monotonicity of Ψ in each variable, we have $t = t'$.

References

American Psychological Association (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, *47*(12), 1597–1611.

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

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

Coren, S. (1992). *The left-hander syndrome: The causes and consequences of left-handedness*. New York: The Free Press.

Dzhafarov, E. (2002). Multidimensional Fechnerian scaling: pairwise comparisons, regular minimality, and nonconstant self-similarity. *Journal of Mathematical Psychology*, *46*, 583–608.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

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

Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, *83*, 65–79.

Falmagne, J.-C., Iverson, G., & Marcovici, S. (1979). Binaural "loudness" summation: Probabilistic theory and data. *Psychological Review*, *86*, 25–43.

Fechner, G. T., 1966/1860. Elements of psychophysics. Holt, Rinehart, and Wilson, New York, translated from *Elemente der Psychophysik* (1860) by H. E. Adler.

Findlay, J. M. (1978). Estimates on probability functions: A more virulent PEST. *Perception & Psychophysics*, *23*, 181–185.

Gigerenzer, G., & Strube, G. (1983). Are there limits to binaural additivity of loudness? *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 126–136.

Hellman, R. P., & Zwislocki, J. (1963). Monaural loudness function at 1000 cps and interaural summation. *Journal of the Acoustical Society of America*, *35*, 856–865.

Hellström, A. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*, 35–61.

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

Levelt, W. J. M., Riemersma, J. B., & Bunt, A. A. (1972). Binaural additivity of loudness. *British Journal of Mathematical and Statistical Psychology*, *25*, 51–68.

Luce, R. D. (1977). A note on sums of power functions. *Journal of Mathematical Psychology*, *16*, 91–93.

Luce, R. D. (2000). *Utility of gains and losses: Measurement theoretical and experimental approaches*. Mahwah, NJ: Erlbaum errata: see Luce's web page at www.socsci.uci.edu.

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

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

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In: Atkinson, R. C., Herrnstein, R. J., Lindzey, G., Luce, R. D. (Eds.), *Stevens' handbook of experimental psychology* (vol. 1 and 2) (2nd ed.) (pp. 3–74). New York: Wiley.

Michell, J. (1990). *An introduction to the logic of measurement*. Hillsdale, NJ: Erlbaum.

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

Schneider, B. (1988). The additivity of loudness across critical bands: A conjoint measurement approach. *Perception & Psychophysics*, *43*, 211–222.

Schneider, B. A. (1980). A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika*, *45*, 357–372.

Steingrimsson, R. (2002). *Contributions to measuring three psychophysical attributes*: *Testing behavioral axioms for loudness*, *response time as an independent variable*, *and attentional intensity*. Psychology Ph.D., University of California, Irvine, available at aris.ss.uci.edu/~ragnar/thesis.html.

Steingrimsson, R., & Luce, R. D. (in press). Evaluating a model of global psychophysical judgments: II. Behavioral properties linking summations and productions. *Journal of Mathematical Psychology*, accepted for publication.

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

Taylor, M. M., & Creelman, C. D. (1967). Pest: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, *41*(4), 782–787.

Ward, L. M. (1990). Cross-modal additive conjoint structures and psychophysical scale convergence. *Journal of Experimental Psychology: General*, *119*, 161–175.

de Weert, C. M. M., & Levelt, W. J. M. (1974). Binocular brightness combinations: Additive and nonadditive aspects. *Perception & Psychophysics*, *15*, 551–562.

Zimmer, K., Luce, R. D., & Ellermeier, W. (2001). Testing a new theory of psychophysical scaling: Temporal loudness integration. Fechner Day 2001. *Proceedings of the 17th annual meeting of the international society for psychophysics*. Lengerirch, Germany: Pabst.

Zimmer, K. (in press). Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics*.