Evaluating a Model of Global Psychophysical Judgments: I. Behavioral Properties of Summations and Productions

Ragnar Steingrimsson R. Duncan Luce

June 9, 2003

Send correspondence to:
R. Duncan Luce
Social Science Plaza
University of California
Irvine, CA 92697-5100
Phone: 949-824-6239
FAX: 949-824-3733
E-mail: rdluce@uci.edu

File: SL-1-6c.tex

Evaluating a Model of Global Psychophysical Judgments: I. Behavioral Properties of Summations and Productions[1]

Abstract

The research presented is a partial empirical evaluation of the second author's proposed psychophysical theory (Luce, 2002, 2003a,b). 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 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 structural conditions 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. Despite a number of methodological issues, considerable support is found for particular forms of Ψ for summations and ratio productions separately. Steingrimsson and Luce (2003a) explore the behavioral properties that link summations and productions; Steingrimsson and Luce (2003b) addresses issues regarding the form of the psychophysical and weighting functions.

[1]This article is based, in part, on the first author's Ph. D. dissertation (Steingrimsson, 2002).

Psychophysicists have long been interested in the subjective attributes corresponding to physical intensity. This includes the literature on how a subjective measure, such as is arrived at using jnds or magnitude estimation, grows with intensity (Dzhafarov, 2002; Fechner, 1860; Stevens, 1975) as well as how intensity summates when, e.g., signals are administered to the two ears (Falmagne, 1976; Falmagne et al., 1979; Levelt et al., 1972; Gigerenzer and Strube, 1983; Schneider, 1988). Recently the second author has developed a class of theories whose main conceptual feature is to relate ratio production ideas to summation ones (Luce, 2002, 2003a,b).

The present article reports four experiments that test behavioral aspects of these theories for the two attributes separately. Steingrimsson and Luce (2003a) explore linking relations between summation and production that force a common psychophysical function. Steingrimsson and Luce (2003b) is concerned with the form of the psychophysical and weighting functions and some additional relevant experiments.

The structure is as follows. Section 1 describes the representations of signal summation and a generalization operation of ratio production arrived at in Luce (2003a,b). Section 2 states experimental methods that are common to all experiments. Section 3 examines the questions of the existence of bias in the ears in both one and two-ear matching. Section 4 reports two experiments that test important properties that follow from the individual representations. These properties do not reflect how the two representations relate to one another. Finally, Section 5 summarizes the experimental findings of this article, discusses several methodological issues, and suggests further work to be done.

The theory developed in stages, the first leading to the beginning of the experimental program (Experiments 1-2), and later stages involved modifications in response to the empirical findings and to new experiments to run. This coevolution means that some of the experiments are not as ideally realized as one would like from our current theoretical perspective.

1 Underlying Features of the Theory

The properties are constructed using two psychological "operations." One is a form of summation of which loudness summation is one example. The second is a "production" operation that is a generalized form of ratio production. In Steingrimsson and Luce (2003b) we will explore the method of ratio production further.

The original impetus for Luce's (2002) model of global psychophysical judgments were results of Luce (2000) that were originally developed within the context of utility theory. They were based on an assumption of no bias, or asymmetry, in the summation. Our first data, (Experiments 1 and 2), led him to generalize those results so as to incorporate biased summation. This formulation was subsequently improved and generalized further in Luce (2003a,b). The remaining properties are studied in the second article, and issues about the form of the psychophysical and weighting functions will be taken up in a third

article.

Previous theoretical work of a similar nature to that of Luce (2002) has employed operations analogous to the summation and production operations (e.g. Narens, 1996; Levelt et al., 1972). However, Luce takes the approach, typical of physics, of linking these two operations together—a feature that proves critical for establishing that a common representation exists for both methods (see Steingrimsson and Luce, 2003a).

The theory is formulated deterministically (algebraically) which, of course, is a major idealization of data. It would be more realistic and representative of the data to formulate a probabilistic version. However, we do not know how to formulate probabilistically the interlocking structure that is at the heart of Steingrimsson and Luce (2003a).

1.1 Primitives and Basic Assumptions

1.1.1 *Joint presentations*

The first primitive is the set of ordered pairs (x, u), which in the present study is interpreted to mean that a pure tone x is presented to the left ear and a pure tone u of the same frequency and phase to the right ear. In this context, let ϵ_l and ϵ_r be thresholds for the left and the right ear respectively and let x' and u' be intensities presented in the left and the right ear respectively; then our notation is $x = x' - \epsilon_l$ and $u = u' - \epsilon_r$. Thus, $x = 0$ means the threshold intensity in the left ear and $u = 0$, in the right ear.

The behavioral task used in the experiments is to have participants produce a tone z that in some to-be-specified sense is perceived as equal in loudness to (x, u).

Several alternative interpretations exist, e.g., Schneider (1988) used signal intensities of frequencies separated by more than a critical band, so the stimulus (x, u) has intensity x at frequency f and intensity u at frequency f'.

Temporal summation in which (x, u) is interpreted as the presentation of x for a brief duration followed immediately by u presented equally long is another interpretation of summation.[2]

Interpretations of (x, u) can be extended to other domains: for instance, perception of line-lengths, aspects of vision such as brightness, (e.g. de Weert and Levelt, 1974), cross-modal cases such that x and u belong to different modalities (Ward, 1990), weight lifting, etc.

1.1.2 *Ordering*

The second primitive, $\succsim$, is the ordering of stimuli by loudness: $(x, u) \succsim (y, v)$ means that the stimulus (x, u) is judged as at least as loud as (y, u). The indifference relation $\sim$ 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. We make three assumptions about $\succsim$:

[2] Experiments based on this interpretation using white noise stimuli were reported by Zimmer et al. (2001).

Assumption 1. *Equivalence relation*: The relation $\sim$ on stimuli is an equivalence relation, i.e., for all $(x,u),(y,v),(z,w)$, *it is transitive*,

$$\left.\begin{array}{l}(x,u)\sim(y,v)\\(y,v)\sim(z,w)\end{array}\right\}\Rightarrow(x,u)\sim(z,w), \tag{1}$$

symmetric,

$$(x,u)\sim(y,v)\Leftrightarrow(y,v)\sim(x,u), \tag{2}$$

and *reflexive,*

$$(x,u)\sim(x,u). \tag{3}$$

One possibility, which we do not assume, is that joint presentations satisfy the condition of no bias

$$(x,u)\sim(u,x), \tag{4}$$

which we call *joint-presentation symmetry* or, for short, *jp-symmetry.* Whether or not this holds turns out to matter considerably for the nature of the theory. So our first two experiments focus on this property.

Denote by $\geq$ the usual physical ordering of intensity. The second assumption is a form of compatibility between the physical and subjective orderings: The loudness ordering of two stimuli for which the intensity to one ear is identical agrees exactly with the order of physical intensity of the signals presented in the other ear. Formally,

Assumption 2. *Compatibility of* $\succsim$ *and* $\geq$: For all intensities x,y,u,v,

$$(x,u)\succsim(y,u)\Longleftrightarrow x\geq y, \tag{5}$$

$$(x,u)\succsim(x,v)\Longleftrightarrow u\geq v. \tag{6}$$

Our next assumption reflects the fact that physical intensity can be thought of as a continuous variable and that every stimulus can be matched by either a totally asymmetric stimulus or by a totally symmetric one. Formally,

Assumption 3. *Solvability*: For every stimulus (x,u), there exist intensities denoted $z_i, i=l,r,s$, such that

$$(x,u)\sim(z_l,0),\ (x,u)\sim(0,z_r),\ (x,u)\sim(z_s,z_s). \tag{7}$$

We refer to z_l and z_r as *asymmetric* matches and z_s as a *symmetric* one. It is not difficult to show from Assumptions 1-3 that $\succsim$ is a weak order, i.e., transitive and connected, on the stimulus conjoint structure (Proposition 1 of Luce, 2002).

Although assumption 3 is very plausible within the psychophysical context, in reality when making the single ear matches, z_l and z_r using headphones the tone percept moves from a somewhat middle localization to the matching ear. This seems to create some experimental difficulty. Our impression, not systematically explored, is that people find it difficult to ignore the changes

in localization. This impression motivated us to find ways to use, wherever possible, the symmetric match z_s.

We do not experimentally investigate these three assumptions because they seem to be so well grounded in psychophysical experience.

1.1.3 *Generalized ratio production*

The third primitive is a generalized form of ratio production. Suppose that $x > y \geq 0$ and let $p > 0$ be a positive number. Let (z, z) denote a signal pair that the respondent says makes the subjective "interval"[3] from (y, y) to (z, z) stand in the ratio p to the subjective interval from (y, y) to (x, x). Denote this chosen stimulus by

$$(x, x) \circ_p (y, y) := (z, z). \tag{8}$$

This is a generalization of ordinary *ratio production*[4],[5] which involves the special case $(y, y) = (0, 0)$, i.e., $(x, x) \circ_p (0, 0) = (z, z)$.

Observe that by the previous assumptions, one can replace the symmetric pairs of (8) by non-symmetric ones to get a more general expression

$$(x, u) \circ_p (y, v) := (z, w) \tag{9}$$

and vice-versa. Thus, there is no loss of generality in assuming the symmetric case.

The following assumption is made about a very special case of (8): Consider the judgment of matching the null interval from (x, x) to (x, x). Whatever p is given, we expect the response to be the same null interval. Formally,

Assumption 4. *Idempotence of* $\circ_p$: For every $p > 0$ and all signals x,

$$(x, x) \circ_p (x, x) = (x, x). \tag{10}$$

A second non-controversial property is that if we replace (x, x) by a louder signal (x', x'), the ratio production will also become louder. Formally,

Assumption 5. *Monotonicity and substitutability of* $\circ_p$: For every $p > 0$ and all signals x, x', y, y',

$$(x', x') \succsim (x, x) \Longleftrightarrow (x', x') \circ_p (y, y) \succsim (x, x) \circ_p (y, y)$$

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

[4] In *generalized ratio estimation* the respondent is asked to state the ratio $p = p(x, y, z)$ such that, for $y < x, z$ the distance from (y, y) to (z, z) stands in the ratio p to the distance from (y, y) to (x, x). We have not studied this method experimentally, but see Steingrimsson and Luce (2003b)

[5] A variant on this method is Stevens' (1975) method of magnitude production in which $y = 0$ and no reference x is unspecified. When x is specified, it is usually called magnitude production with a standard. This is explored in some detail in Steingrimsson and Luce (2003b).

and the operation $\circ_p$ is continuous and nowhere constant in the second variable (y, y).

One might have expected the parallel form of monotonicity for the second variable but, as we shall see, that is not correct for some $p > 1$.

1.1.4 *Decomposability*

At the center of the theoretical representation is a psychophysical function Ψ that maps signal pairs to numbers, i.e., $\Psi : \mathbb{R}_+ \times \mathbb{R}_+ \stackrel{onto}{\longrightarrow} \mathbb{R}_+$, and that preserves the ordering $\succsim$. Implicit in assuming that the representation is "onto" the non-negative real line is a structural assumption about the stimuli that seems appropriate in the psychophysical context. Among other things, it means that we assume

$$\Psi(0, 0) = 0, \tag{11}$$

which seems appropriate. It also assumes indefinitely large signals which, of course, is an idealization.

We will be much focussed on the forms of $\Psi(x, u)$ and $\Psi[(x, x) \circ_p (y, y)]$ and behavioral properties that underlie the forms we have come up with. A key assumption underlying the theoretical development is that the representations are decomposable in the following sense:

Assumption 6. *Decomposability:* There exist strictly increasing functions $F, G_p : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $p > 0$ and all intensities x, y, u, v

$$\Psi(x, u) = F[\Psi(x, 0), \Psi(0, u)] \tag{12}$$

$$\Psi[(x, x) \circ_p (y, y)] = G_p[\Psi(x, x), \Psi(y, y)]. \tag{13}$$

This simply means that the roles of the component variables appear only as Ψ-functions of them. The goal of the theory is to specify, first, the forms of the functions F and G_p and then the form of Ψ as a function of physical intensity.

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

We begin by stating in one place the numerical representations that have been derived from behavioral primitives (Luce, 2002, 2003a,b), i.e., the forms of F and G_p. Then we go on to discuss and test certain behavioral properties that are described below and in Steingrimsson and Luce (2003a).

First, using asymmetric matching, z_l or z_r in (7), one can use the behavioral properties to show that there are constants $\delta \geq 0$, $\gamma > 0$ and a distortion of numbers $W : \mathbb{R}_+ \stackrel{onto}{\longrightarrow} \mathbb{R}_+$ such that the following three properties hold:

$$\Psi(x, u) = \Psi(x, 0) + \Psi(0, u) + \delta\Psi(x, 0)\Psi(0, u) \quad (\delta \geq 0), \tag{14}$$

$$\Psi(x, 0) = \gamma\Psi(0, x) \quad (\gamma > 0), \tag{15}$$

$$W(p) = \frac{\Psi\left[(x, x) \circ_p (y, y)\right] - \Psi(y, y)}{\Psi(x, x) - \Psi(y, y)} \quad (x > y \geq 0). \tag{16}$$

Second, using symmetric matching, z_s in (7), one can prove that the summation property (14), restricted to $\delta = 0$, and that (16) hold. However, the very restrictive constant-bias property, (15), may or may not apply under symmetric matching (see below) .

Two important special cases of (14) are: The unbiased case, $\gamma = 1$, which means jp-symmetry, (4), holds, and the case where $\delta = 0$. One can show that if the constant bias property, (15), holds, then $\delta = 0$ if, and only if,

$$\Psi(x,u) = \eta\Psi(x,x) + (1-\eta)\Psi(u,u), \tag{17}$$

where $\eta = \gamma/(1+\gamma)$ is a constant. Observe that $0 < \eta < 1$. As this suggests, a key question to be addressed experimentally is a property that is equivalent to $\delta = 0$. That is taken up in Steingrimsson and Luce (2003a).

As noted earlier, we may replace a stimulus by one that is indifferent to it, so we may write (16) in the apparently more general form

$$\frac{\Psi\left[(x,u) \circ_p (y,v)\right] - \Psi(y,v)}{\Psi(x,u) - \Psi(y,v)} = W(p) \quad [(x,u) \succ (y,v) \gtrsim (0,0)]. \tag{18}$$

The following subsections state behavioral properties that are implied individually by the representations (14) and (16). They are then tested in Sections 3 and 4 . Linking of the representations via a common function Ψ is topic of Steingrimsson and Luce (2003a). The very strong property (15) is studied in Steingrimsson and Luce (2003b); we do not comment on it further here.

1.2.1 *Subjective summation*

First, note that the expression (14) is not as symmetric as it may first seem, in particular, it does not imply (4). Second, this representation can always be transformed into a binary additive conjoint representation of summation of the form:

$$\Phi(x,u) = \Phi_l(x) + \Phi_r(u). \tag{19}$$

When $\delta = 0$ this is obvious using the identifications

$$\Phi(x,u) := \Psi(x,u), \Phi_l(x) := \Psi(x,0), \Phi_r(u) := \Psi(0,u).$$

When $\delta > 0$ it is less obvious that (19) holds, but it is true with

$$\begin{aligned} \Phi(x,u) &: = \ln[1 + \delta\Psi(x,u)], \\ \Phi_l(x) &: = \ln[1 + \delta\Psi(x,0)], \\ \Phi_r(u) &: = \ln[1 + \delta\Psi(0,u)]. \end{aligned}$$

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

$$\left.\begin{array}{l}(x,t) \sim (z,v) \\ (z,u) \sim (y,t)\end{array}\right\} \Longrightarrow (x,u) \sim (y,v), \tag{20}$$

is predicted to hold (Krantz et al., 1971). Note that in a qualitative sense this describes the "additive cancellation" of t and z.

In the presence of monotonicity, solvability, and Archimedeanness, the Thomsen condition, (20), implies the additive representation (19), which in turn implies a stronger property called *double cancellation*. This the same as (20) but with each $\sim$ replaced by $\succsim$. Obviously, double cancellation implies the Thomsen condition, but not conversely except if we have solvability and monotonicity. For details on the above comments, see Krantz et al. (1971).

Some theoretical work has been done in the area of binaural loudness summation. For instance, Levelt et al. (1972) employed aspects of measurement theory to formulate experiments leading them to conclude that loudness summation is additive in the sense of additive conjoint measurement and led to approximately 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 these traditions, 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 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 participant, and Gigerenzer and Strube (1983) with 12 participants. Because of this inconsistency, we re-examine conjoint additivity by testing the Thomsen condition empirically in Experiment 3 (Section 4).

1.2.2 *Production commutativity*

The basic (initial) idea embodied in the representation of generalized ratio production, (16), is that the respondents do what they are told to do using the distortion Ψ of intensities and the distortion W of numbers. An important, although easily demonstrated, consequence of (18) is the behavioral property called *(subjective) production commutativity:* For $p > 0, q > 0$,

$$[(x, x) \circ_p (y, y)] \circ_q (y, y) \sim [(x, x) \circ_q (y, y)] \circ_p (y, y). \tag{21}$$

Observe that the two sides differ only in the order of applying p, q, which is the reason for the term "commutativity." This property with $y = 0$ also arose in Narens' (1996) theory.

1.2.3 *Need for an axiomatization*

We have written the various expressions above involving Ψ, (14), (16), and (17), as if it is the same function in each case.[6] Yet, the testable properties we have stated so far, the Thomsen condition and subjective production commutativity, draw on them individually, not in combination. And so even if both are sustained empirically, we would have no reason to suppose a common Ψ exists. This is reflected in the fact that so far we have not established any link between the joint presentation structure and the subjective production structure. This has been accomplished theoretically; these results are summarized and tested in Steingrimsson and Luce (2003a).

2 Experimental Methods Common to All Experiments

The several experiments reported here and in Steingrimsson and Luce (2003a,b) all have in common a number of testing strategies that are now outlined. Other aspects are described later as relevant (see Appendix A for details on suggested methodological improvements).

2.1 Remark on notation

Recall that the equations for properties that are tested are written in terms of intensity measured above—not relative to—threshold. However, the experimental design and results are all reported using decibels SPL. This practice does not pose problems in the description of the current experiments.

2.2 Signal presentations

The experiments were carried out in the auditory domain using a 1,000 Hz sinusoidal tone presented for 100 ms, which included 10 ms on and off ramps.

2.3 Participants

A total of 33 students—graduate and undergraduate—from the University of California, Irvine, participated in the four experiments of this article. The first author was one of them.[7] Of these 33, three participants 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. All of P9's data involving production judgments are excluded because a strategy was followed that was inconsistent with instructions

[6] This does not apply to constant bias, (15), which follows from the separate asymmetric representations.

[7] We judged this acceptable because no knowledge of the design could affect behavior and because it was very useful in fine tuning the procedures.

(For details see Luce's web page: aris.ss.uci.edu/cogsci/personnel/luce/P9.pdf.), but P9's matching data are included. This leaves 27 individuals, 6 male and 21 female, whose data are reported in full. All participants had normal hearing as assessed through self-reporting and by an audiometric test (Micro Audimetrics EarScan ES-AM).

All participants, 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 designed according to the rules of and received the approval of the UC Irvine's Institutional Review Board.

It would be desirable from an experimental standpoint for each participant to participate in every experiment. However, for various logistical reasons, not least being the fact that the program evolved over two years, this was not achieved (for details on individual participation by experiment for the three articles, see Luce's web page: aris.ss.uci.edu/cogsci/personnel/luce/SLParticpants.pdf).

2.4 *Estimating one-ear and two-ear matches*

The general procedures for obtaining matching data are described here. Those for ratio production are described when first needed.

The three types of matches used are listed in Assumption 3. To indicate the precise form of trial types, let $\langle A, B\rangle$ denote a presentation of A followed by a temporally displaced presentation of B. We used a temporal delay between A and B of 500 ms. Three trial types, corresponding to (7), were used.

$$\langle (x,u),(z_l,0)\rangle, \tag{22}$$

$$\langle (x,u),(0,z_r)\rangle, \tag{23}$$

$$\langle (x,u),(z_s,z_s)\rangle. \tag{24}$$

Following a tone presentation, participants used key presses either to adjust the intensity of z_i, $i = l, r, s$, to repeat the previous trial, or to indicate satisfaction with the loudness match. Intensity adjustments were done in increments of 0.5, 1, 2 or 4 dB. These were named and presented to participants 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 intensity. After an adjustment, the altered tone sequence was played. The previous trial could be repeated by pressing the "r" key. Satisfaction with the match was indicated by pressing the "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 1 sec. 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 participants, the task was explained as that of making the second stimulus pair equal in loudness to the first one. The instructions stressed the importance of paying attention solely to the loudness of the stimuli

and ignoring the subjective sense of tone location.[8] The latter is of particular importance because, for instance, the match of a two-ear stimulus in a single ear clearly requires disregarding the difference in the subjective location quality of the stimuli. For this reason, we use two-ear matches as much as possible.

2.5 Procedure

Experiments were conducted in sessions lasting no more than one hour. Participants typically ran two to four sessions per week and, with rare exceptions, no more than one session per day. The average number of sessions run by participants from whom data are reported was between 18 and 19, which includes training practice, piloting, and experimental sessions.

The initial session was devoted to obtaining the written consent, explaining the practice task—which was a matching task of the type outlined in Experiment 2—and running the practice blocks. Depending on the experiment, practiced participants 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 participants' control.

2.6 Equipment

Stimuli were generated digitally using a personal computer and played through a 16-bit digital-to-analog converter (Quikki; Tucker-Davis Technology), at the converting 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 individual, single-walled, IAC sound booths.

Dr. Bruce G. Berg, UC Irvine, generously made his laboratory and equipment available for the conducting of these experiments for which we are most grateful.

2.7 Statistical method and result presentation

The properties stated are of the form $A = B$, but of course our estimates of A and of B are variable. Thus, to test the theory, we are examining a number of null hypotheses, and 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. As we will see, 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

[8] Preliminary data, not discussed here, suggest that participants may have difficulty ignoring subjective location in some instances.

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, and doing so guided our experimental designs. Even so, in retrospect we know that some of the properties should have been checked slightly differently, but time and resources did not permit us to make all of these improvements. For detailed discussion about this issue, see Appendix A.

Because we do not have a theory that predicts the distributions for these expressions, we felt that the statistical analysis should be nonparametric. Our choice was the Mann-Whitney U-test. A significance level of .05 was used. 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 intensity 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.

We have elected to use averages and standard deviations in reporting the data. Although medians would be preferable if we could accurately estimate them, in fact the discrete nature of the signal values makes the mean a better estimate provided that the distributions are reasonably Gaussian, which they appear to be. The standard deviations are reported except when the results are presented graphically. In those cases, error-bars are normally used to indicate variability but as the statistical method is a non-parametric one, using error-bars based on standard deviations will in some cases seem contradictory to the statistical results, confusing the data presentation. In these cases, the range of the corresponding standard deviations are given numerically for each participant as an approximate indication of data variability.

2.8 Methodological Variations

As mentioned earlier and as is detailed in Appendix A, during the course of experimentation, we learned to improve our procedures. These improvements were incorporated as we realized them. The following two are the most important.

- When possible, avoid matching two-ear stimuli using intensities in only one ear.
- In testing multi-step properties, using individual estimates as input for subsequent estimates instead of only their averages was statistically an improvement. This method was incorporated into Experiments 3 and partially in 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.

Our general assumption is that whether $(x,0) \succ \sim \prec (0,x)$ it is independent of x and we say

$$\begin{aligned} &\text{left bias} \iff (x,0) \succ (0,x) \\ &\text{no bias} \iff (x,0) \sim (x,0) \\ &\text{right bias} \iff (x,0) \prec (0,x). \end{aligned} \tag{25}$$

If (15) is satisfied, then left bias is equivalent to $\gamma > 1$, no bias to $\gamma = 1$, and right bias to $\gamma < 1$.

3.1 Experiment 1: Bias[9] using single ear matches

The initial aim of this experiment was to explore whether the joint-presentation symmetry holds by comparing matches in a single ear. 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. The experiment is based on the matches $(x,y) \sim (z_l,0) \sim (0,z_r)$ of Assumption 3. (We take up using the match $(x,y) \sim (z_s,z_s)$ in Experiment 2.)

3.1.1 *Method*

Three tone intensities were used: a = 58 dB, b = 64 dB, c = 70 dB. These three intensities give rise to six ordered stimulus pairs: (a,b), (a,c), and (b,c) corresponding to the left side of (4) and (b,a), (c,a), and (b,c) corresponding to the right side of (4). 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 (22) and (23), respectively.

Data were collected using two methods.

M1: Left and right ear matches were separated into blocks of trials and sessions were blocked on matching ear condition.

M2: Left and right ear matches were collected within a block and their order alternated.

[9] The hypothesis tested in this and the next experiment is that there is no bias; however, given the nature of the results, namely, that often there is a bias, we decided it would be misleading to title the experiment "no bias."

3.1.2 Results

Six individuals participated. All participants ran using M1 and two participants (P6, P22) ran using M2. The data for the two methods are presented graphically in Figures 1 and 2, respectively. Two graphs, one for the left and one for the right ear match, are given for each participant. In each graph, matching results for stimulus conditions of the form (x, y) are labelled xy and for (y, x) are labelled yx[10].

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 intensity 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.

The range of standard deviations for each participant were:

M	P2	P6	P10	P19	P21	P22
1	1.92–3.56	0.86–1.53	0.81–1.94	1.25–2.51	1.10–3.21	1.02–1.98
2	N/A	1.16–1.86	N/A	N/A	N/A	0.95–1.70

In general, the standard deviation decreases with increasing stimulus intensity—see also result section for Experiment 2. No such differential trend in variability was discerned by matching ear.

Using M1: For P2 and P10 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.

All participants exhibit left bias when matching in the left ear (for P21, the number of observations for left ear matching suffices only to indicate a trend). Two individuals, P6 and P22, both show a shift to a right bias for right ear matches. P2, P10 and P19 show no bias when matching in the right ear although the pattern of data suggests a trend towards left bias for P10, a bias shift for P19, and no trend for P2. Only P21 shows no change in bias based on matching ear.

Using M2: The results for P6 and P22 showed the same bias shift as when using M1. The bias shift is larger, if anything, for P6.

3.1.3 *Discussion*

With respect to the statistical testing, the results for M1 may be divided into two categories:

1. $z_l \leq z'_l$ and $z_r \leq z'_r$ (P2, P10, P19, and P21),

[10] Care must be taken not to confuse the notation xy for multiplication.

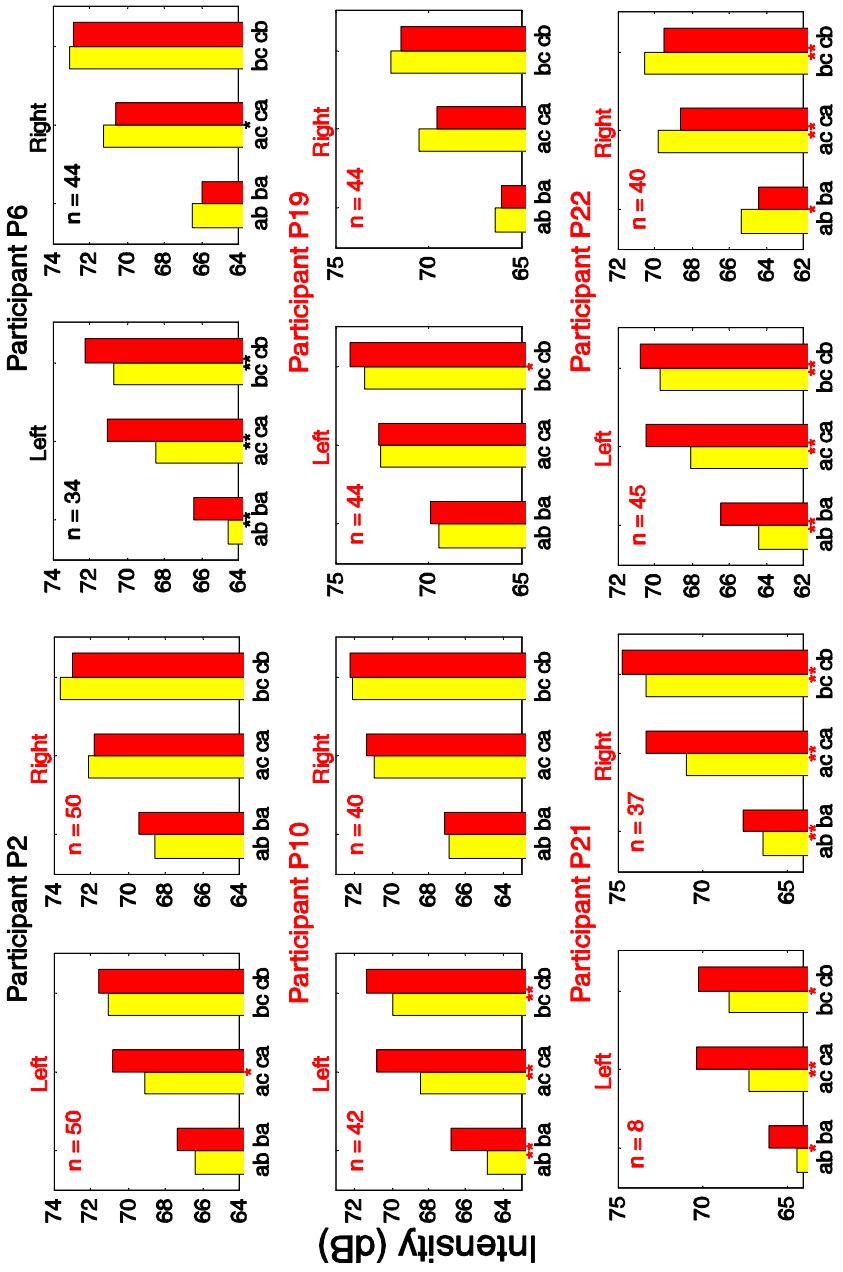

Figure 1: Experiment 1: bias in single ear matches

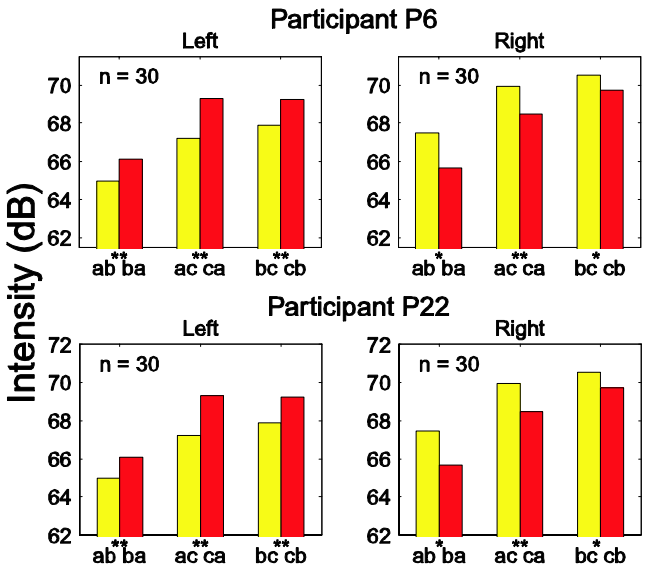

Figure 2: Experiment 1: bias in single ear matches (Method 2).

2. $z_l < z'_l$ and $z_r > z'_r$ (P6 and P22).

The results in the second category are not described within Luce's (2002a,b,c) theoretical framework. If trends in the data are taken into account, P19 would belong to the second category. In fact, only P21 shows, statistically, a 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 seemed prudent to look for methodological explanations.

Sessions were blocked on matching ear. We suspected that 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 M2, where matching ear alternated within a block. In short, the result pattern did not change, lessening the likelihood of the result's being an methodological artifact.

We also considered the use of 2AFC paradigm: The stimulus $\langle (x, 0), (0, x) \rangle$ is presented and the participant'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 (see Appendix A for details) led participants almost always to judge the second tone louder than the first regardless of condition. 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 attempted this.

A clear consequence of these findings is that experimental procedures in which a comparison between the results of left and right ear matching or loudness production are not reliable. For instance, testing for jp-symmetry, $(x, u) \sim (u, x)$, may give two different results depending on which ear the matches are made in. However, a procedure which involves the consistent use of either the left or the right ear is not problematic with respect to these results. Clearly, the results make it desirable to use two-ear matching where ever possible.

The phenomenon of bias shift depending on the matching ear is quite unexpected and the results appear sufficiently robust that they warrant further theoretical development. We take this up, including a new theoretical proposal, in Steingrimsson and Luce (2003b).

3.2 Experiment 2: Bias using two-ear matches

Because of the difficulties with one-ear matching, we elected to explore whether the jp-symmetry, (4), holds using matches of the form $(z_s, z_s) \sim (x, u)$ and $(z'_s, z'_s) \sim (u, x)$ and then compare z_s and z'_sUp to experimental variability, they must be equal if jp-symmetry holds—see Appendix B for details.

3.2.1 *Method*

We used the same stimuli as in Experiment 1. The trial type used is given by (24). The six matching conditions were all run in a block of trials.

3.2.2 *Results*

Fifteen individuals participated. Their data are presented graphically in Figure 3. The range of standard deviations (in dB) for each participant was:

P2	P3	P4	P5	P9	P10	P11	P12
2.84–3.58	1.28–1.72	1.66–2.91	2.14–3.29	1.10–1.39	0.40–0.79	1.69–2.64	1.46–2.15

P14	P16	P17	P20	P22	P23	P24
0.64–1.70	1.14–1.94	1.02–2.02	1.95–2.53	1.07–2.07	1.32–2.39	1.30–2.66

As was the case for in Experiment 1, standard deviation tends to decreases with increasing stimulus intensity. Here, comparing the 30 relevant pairs of stimuli of lower and higher intensity, the standard deviations are higher for the lower stimulus in 29 of 30 cases.

For 12 participants[11], one or more of the three conditions were found to be statistically different. The trend for the remaining conditions is consistent

[11] Ps 2, 4, 5, 9, 10, 11, 14, 16, 17, 22, 23, and 24.

with those found statistically different. In 10 cases[12], the pattern of results was that of $(x,u) \prec (u,x)$, and in two cases[13] the opposite pattern was obtained, i.e. $(x,u) \succ (u,x)$. For three participants[14] the hypothesis of no bias, i.e., $(x,u) \sim (u,x)$, was not rejected.

3.2.3 *Discussion*

In the terms introduced in (25) and in Appendix B, 10 participants 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 participants. So, as a general rule, this symmetry property is rejected. The finding of three participants 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 sample size used.

As we will note in Steingrimsson and Luce (2003a), in the unbiased case the theory predicts associativity as well as commutativity of an induced operation. A small bias that may go undetected when testing jp-symmetry (= commutativity of the operation) may be picked up by the harder-to-pass test for associativity.[15] So, for those participants where commutativity may hold, it would have been desirable to test them also on associativity in order more firmly to establish no bias. By the time we realized this, it was too late to collect these data.

Luce's (Luce, 2002, 2003a,b) 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 (Stanley, 1992, pp 27-33) reports similar phenomenon for the ears. Whether earedness plays a part in the bias we have observed is an open question.[16]

R. P. Hellman (personal communication, December 2001) suggested that the failures of jp-symmetry are caused by a difference in threshold levels for the two ears. She feels that the empirical results of Hellman and Zwislocki (1963) support this view. We are not convinced. One reason was the shift in bias direction for some participants, found in Experiment 1, depending on the matching ear. In that case, differences in threshold levels are not sufficient to explain the data obtained.

[12] Ps 2, 3, 4, 5, 10, 11, 14, 16, 17, 22, and 23.

[13] Ps 9 and 24.

[14] Ps 3, 12, and 20.

[15] Data obtained by Zimmer et al. (2001) in experiments on temporal summation show individuals who pass commutativity but fail associativity.

[16] 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 (Stanley, 1992, pp 27-33).

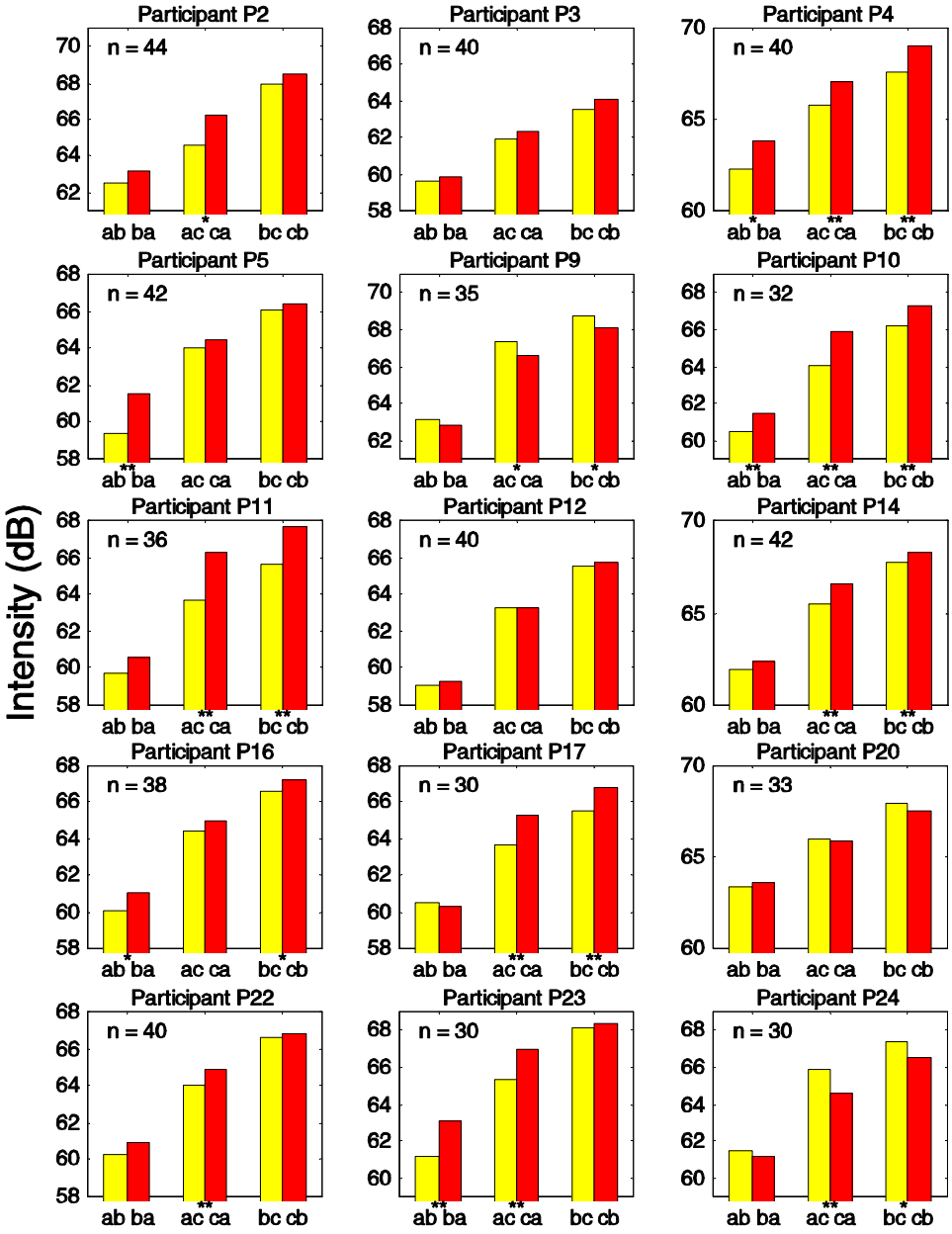

Figure 3: Experiment 2: Bias in two-ear matches

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, (20), and production commutativity, (21). They have the feature that each follows from just one of the two representations, the former from the representation of summation, (14), and the latter from the subjective proportion representation, (16).

4.1 Experiment 3: Thomsen condition

The goal of this experiment is to test the well known necessary condition of a binary additive conjoint representation, the Thomsen condition, (20).

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

4.1.1 *Method*

With reference to (20), the property is tested by successively obtaining the estimates, z', y' and y'' using

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

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 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:

$$\begin{aligned} \mathbf{A} &: \quad x = 66, t = 62, v = 58, \text{ and } u = 70 \text{ dB}, \\ \mathbf{B} &: \quad x = 62, t = 59, v = 47, \text{ and } u = 74 \text{ dB}. \end{aligned}$$

The two stimuli sets were run in separate sessions, with data for stimulus set **A** collected first. For P29 the **A**,**B** pair was run a second time (see discussion).

4.1.2 *Results*

Result for 12 participants are displayed in Table 1. In the table, averages, standard deviations, number of observations, and statistical results are listed for each participant.

Part.	Stim. set	y' (s.d.)	y'' (s.d)	n	p_{stat}	Stat. Trend
P10	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	$y' = y''$
P22	A	71.30 (1.10)	70.88 (0.93)	30	.127	$y' = y''$
	B	72.08 (1.21)	71.88 (1.01)	30	.311	$y' = y''$
P23	A	71.82 (1.20)	71.57 (0.81)	30	.164	$y' = y''$
	B	72.88 (1.05)	73.62 (1.11)	40	.005	$y' \neq y''$
P25	A	72.60 (3.59)	72.00 (2.39)	30	.312	$y' = y''$
	B	76.58 (2.26)	75.59 (1.96)	40	.062	$y' = y''$
P26	A	69.50 (1.14)	69.74 (0.99)	39	.267	$y' = y''$
	B	72.65 (1.21)	72.39 (1.26)	41	.264	$y' = y''$
P27	A	69.48 (1.15)	69.36(1.88)	32	.912	$y' = y''$
	B	71.66 (2.17)	72.19 (2.06)	40	.107	$y' = y''$
P28	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	$y' = y''$
P29	A	70.20 (2.32)	69.58 (1.73)	40	.123	$y' = y''$
	B	72.78 (1.82)	72.24 (1.68)	60	.084	$y' = y''$
P30	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	$y' = y''$
P31	A	69.91 (1.39)	70.21 (0.72)	40	.154	$y' = y''$
	B	70.65 (1.25)	70.90 (0.94)	40	.353	$y' = y''$
P32	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	$y' = y''$
P33	A	73.48 (1.87)	73.75 (1.39)	40	.253	$y' = y''$
	B	69.42 (2.85)	70.15 (1.84)	60	.115	$y' = y''$

Table 1: Experiment 3: Thomsen Condition

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. Participant P29 failed the Thomsen condition for the first presentation of stimulus set A, but it was accepted on the second presentation (see discussion).

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

4.1.3 *Discussion*

As noted earlier, of five published studies concerning the 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 participants. Their experimental and statistical methods were similar to ours.[17] Notably, the stimuli set **B** used stimuli having the same relative intensity relationship as those used by Gigerenzer and Strube (1983), although we used 1,000 Hz whereas they used both 200 Hz and 2,000 Hz—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). 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). 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 verses individual judgments in step 1.*

Gigerenzer and Strube (1983) used median estimates for y'; whereas, by the time this experiment was run, we had concluded that it was better to use each individual observation instead. Doing this meant two things. First, we did not lose the impact of variability on the first estimates affecting later one. Second, it avoided the bias that is necessarily introduced using a point estimate.

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

The intensity ranges were 58-70 dB for **A** and 47-74 dB for **B**. We have encountered localization effects as a function of signal difference, but this suggests that B should have been more adversely affected than A. Gigerenzer and Strube (1983) make essentially the same observation coming to a similar conclusions. However, since we find more rejections for set **A** the for **B**, this observation does not seem to an explanation for the difference in results between the two sets.

- *A possible effect of inadequate practice.*

[17] They employed a statistical method suggested by Ulrich Raatz whose results are identical to that of the Mann-Whitney U. (Raatz, personal communication, August, 2002)

Because performance was more consistent with the Thomsen condition on set **B** than on set **A** and they were run in the order **A**, **B**, possibly the practice was inadequate when **A** was run. Our initial practice varied from 90 to 120 trials (the lower number for experienced observers); we collected 30-60 observations for each of z', y', and y''. Hence, our participants had completed no less than 180 matching trials prior to data collection for stimulus set **B**.

Related to this conjecture is the additional data collected from P29. During the **B** phase we noticed a considerable change in P29's behavior. We were able to collect from P29 additional 40 observations for each of **A** and **B** (4 sessions). During the initial four sessions $|t-t'| = 1.09$ whereas in the additional four sessions this dropped by a factor of more than 6 to 0.17. This result, in addition to other indicators, suggests that very extensive practice may indeed be needed to test the Thompson condition. The additional practice led to P29's not rejecting the condition for the second presentation of set **A** (which is what is reported in the table).

Perhaps also related to practice is the fact that the inter-session variability in this experiment exceeded that for the other tests.

Although Gigerenzer and Strube (1983) state that participants 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 they experienced it is impossible to compare it with our procedures.

- *This experiment differed from all others in the present paper in that only one of a pair of jointly presented tones was adjusted.*

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 one tone being changed, this location moves in an approximate horizontal plane. Some participants indicated they found it harder to ignore the latter effect than the former.

A possible role of additional practice is to reduce the more severe localization effects of testing the Thomsen condition.

- *Differences in signal frequency.*

Gigerenzer and Strube (1983) collected data at 200 and 2,000 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 2,000 Hz. Consistent with all our experiments, we used the oft used 1,000 Hz signals. We see no immediate reason why the Thomsen condition should hold for 1,000 Hz, but not for 200 Hz and 2,000 Hz. However, the result does leave the question of a possible effect of frequency on the results. The impact of frequency on the overall result is an empirically open question.

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 discrepancy in results.

The open questions in the above suggest areas of further exploration of the property, e.g. it would be desirable to collect substantially more 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 tend towards accepting the Thomsen property.

4.2 Experiment 4: Production commutativity

The property of production commutativity is given in (21). Appendix C shows that production commutativity can be tested using the preferred[18] two-ear production, i.e.,

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

This is the first experiment in which the $\circ_p$ operator appears and for that reason we present first a general description for its estimation.

4.2.1 *Estimating the $\circ_p$ operation*

Employing the notation defined in Section 2.4, 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 500 ms. The longer delay was introduced to create a subjective sense of two distinct intervals.

Then, an estimate of $x \circ_{p,i} y = v_i$, where v_i is an intensity under the participant's control, is, in the case of $i = s$, obtained using the trial type

$$\langle\langle (y,y),(x,x)\rangle, \langle (y,y),(v_s,v_s)\rangle\rangle. \tag{26}$$

The value of 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.

Other aspects of this process were identical to that for joint presentations, described in Section 2.4.

Instructions to participants took the form of a verbal description of the task coupled with graphical examples. Participants 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

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

as a line segment twice as long as the first and having the same starting point. A comparable example was given for $p = \frac{2}{3}$.

The special case of $y = 0$ with $i = s$ was estimated using the trial type

$$\langle (x, x), (v_s, v_s) \rangle. \tag{27}$$

Participants were instructed to make the loudness of the first tone x stand in proportion p to the second tone v_s. These instructions were analogous to those for the $u > 0$ case.

Note: In constructing trials involving $\circ_p$ care should be taken to chose intensities such that no intensity interval becomes so small that the participants have a problem perceiving it. In addition, to reduce substantial inter-session variability, it is desirable that when testing a particular property that all needed ratio productions be done within the same session. See Appendix A for details.

4.2.2 *Method*

The testing required making four estimates in two steps. The first consisted of

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

and in the second of

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

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

The intensities $x = 64$ dB and $u = 70$ dB and the proportions $p = 2$ and $q = 3$ were used, giving rise to four trial condition in each step. The four trial forms used were as described by expression (26).

For P22 and P25, 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 P26, P27 we switched to using the individual estimates v and v' in the subsequent estimates of w and w'. (See Appendix A for details on the rational for this change). The two estimation steps were done within a block of trials, hence a block contained all eight trial conditions. Trial order was randomized but in such a way that w or w' was only estimated if a prior instance of v or v' was available.

4.2.3 *Results*

Four participants completed this experiment. Their data are presented in Table 4.

Participant	Mean (s.d.)		p_{stat}	n	Statistical Trend
	T_1	T_2			
P22	77.43 (1.56)	77.70 (1.86)	.574	30	$T_1 = T_2$
P25	79.90 (2.21)	80.17 (2.32)	.662	30	$T_1 = T_2$
P27	77.57 (2.70)	78.48 (2.83)	.102	30	$T_1 = T_2$
P28	74.15 (4.55)	74.41 (1.98)	.301	30	$T_1 = T_2$

Table 2: Experiment 4: Proportion commutativity

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

The property was not rejected by any of the four participants.

4.2.4 *Discussion*

Ellermeier and Faulhammer (2000) investigated the related property, threshold-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 (21) in which $y = 0$. They used an experimental paradigm similar to the one employed in the present study, and tested the property using $p, q > 1$ and found it to hold. Theirs and our results provide good initial support for the property.

A theoretical prediction of threshold-production commutativity is that the properties hold for both $p, q < 1$ as well as $p, q \geq 1$. Neither that nor the stronger properties have been tested for $p, q < 1$ or 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 following three properties are satisfied:

$$\begin{aligned} \Psi(x,u) &= \Psi(x,0) + \Psi(0,u) + \delta\Psi(x,0)\Psi(0,u) \quad (\delta \geq 0) \\ \Psi(x,0) &= \gamma\Psi(0,x) \quad (\gamma > 0), \\ W(p) &= \frac{\Psi\left[(x,u) \circ_p (y,v)\right] - \Psi(y,v)}{\Psi(x,u) - \Psi(y,v)}. \end{aligned}$$

For symmetric matches, the first equation with $\delta = 0$ and the third equation both hold, but the theory for symmetric matches does not predict the constant bias of the second equation.

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 testing of the

Ex. #	Name	#P	#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	1 P extra practice
4	Prop. comm.	4	4	0	

Table 3: Summary of experimental results

properties derived from the first and third expression separately. 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. 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 (2003a).

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, (4), does not generally hold, and a majority of the participants were left biased. This result dictated both further theory construction and experimentation towards testing the property of the biased theoretical solution of Luce (2002, 2003a,b).

More specifically, single-ear matching of Experiment 1 found two of six participants statistically 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, hence it quite surprised us. Changes in the experimental design aimed at eliminating a plausible procedural explanation had no effect on the results, making the result less likely to be an artifact of the experimental design. This empirical results warrants further theoretical development, which will appear in Steingrimsson and Luce (2003b). We were able to bypass this problem for bias testing by using the two-ear matching of Experiment 2.

The summation representation implies the Thomsen condition (Experiment 3). As noted, the literature is split on this property. We found, especially after considerable exposure to the task, more evidence for it than against it. Comparisons between our and other's data suggest the possible need for participants' receiving more extensive practice on the task than has otherwise seemed needed. Furthermore, data from Gigerenzer and Strube (1983) suggest that it would be prudent to investigate the property at several frequencies.

The subjective proportion representation implies production commutativity for $p, q > 1$ (Experiment 4). [Threshold-production commutativity was established by Ellermeier and Faulhammer (2000) for $p, q > 1$]. Production commutativity and its threshold version, have yet to be tested for $p, q < 1$.

5.2 Further work

In Steingrimsson and Luce (2003a,b) we take up the testing of further axioms of Luce's theory. However, present research suggests some additional avenues to explore and work be done.

Direct improvements on our results include:

- To increase confidence in the results obtained, it is desirable to increase the number of people tested.
- Production commutativity and its threshold version should be tested with $p, q < 1$.
- One-ear matching test (Experiment 1) showed reversal of bias direction when matching ear was changed. Because this is unexpected and somewhat problematic, further theoretical and empirical exploration of the issue is warranted (see Steingrimsson and Luce, 2003b).

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

- The preliminary success of the present theory in the auditory domain, warrants extending the work done here into other domains and/or to other interpretations of the summation and production operators.
- 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 1,000 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 difference in our results compared to those of Gigerenzer and Strube (1983). Luce's theory makes no predictions about response variation across frequency, but in pilot date we collected using 200, 1,000, 6,000 Hz we observed consistent and significant left bias across all three frequencies conditions (two people), right bias in the 200 Hz condition, left bias in the 1,000 Hz condition, and no significant bias in the 6,000 Hz condition (one person), and no consistent bias (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.

References

Dzhafarov, E. (2002). Multidimensional fechnerian scaling: probability-distance hypothesis. *Journal of Mathematical Psychology*, 45:352–374.

Ellermeier, W. and 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., and 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. 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. and 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. and Zwislocki, J. (1963). Monaural loudness function at 1000 cps and interaural summation. *Journal of the Acoustical Society of America*, 35:856–865.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement*, volume 1. Academic Press.

Levelt, W. J. M., Riemersma, J. B., and 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*. 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. (2003a). Symmetric and asymmmetric matching of joint presentations. *Psychological Review*. Revision under review.

Luce, R. D. (2003b). Increasing increment generalizations of rank-dependent theories. Manuscript submitted for publication.

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.

Stanley, C. (1992). *The left-hander syndrome: the causes and consequences of left-handedness*. The Free Press.

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. and Luce, R. D. (2003a). Evaluating a model of global psychophysical judgments: II. Behavioral properties linking summations and productions. Manuscript submitted for publication.

Steingrimsson, R. and Luce, R. D. (2003b). Evaluating a model of global psychophysical judgments: III. Invariance and forms for the psychophysical and weighting functions. In preparation.

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

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

Zimmer, K., Luce, R. D., and Ellermeier, W. (2001). Temporal integration of loudness: Commutative and associative properties. Paper presented at the meeting of the European Mathematical Psychology Group, Lisbon, Portugal.

Appendices

A Methodological improvement and recommended procedures

Over the two year the experiments reported here and in Steingrimsson and Luce (2003a,b), plus more not reported (see Steingrimsson, 2002), we learned to improve our methodology an various ways. These improvements were incorporated as we realized them. The following are the main points.

Method of adjustment:

In all experiments we use a free-adjustment method where participants can increase and decrease intensity until they are satisfied with the result. Initially, we used the adaptive procedure PEST (Findlay, 1978) to obtain loudness estimates. However, comparison of results using PEST to the free-intensity adjustment method showed the end results to be comparable but with the free-adjustment method requiring substantially shorter experimental time than PEST—the five individuals who experienced both methods uniformly preferred the free-adjustment method.

Issues with production judgments:

- In constructing trials involving $\circ_p$ care should be taken to chose intensities such that no intensity interval becomes so small that the participants had a problem perceiving it. Experience from experiments not reported here (see Steingrimsson, 2002) suggests that where this is unavoidable, explicit instructions to participants about the possibility of this situation should be provided.

- 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 productions be done within the same session.

Asymmetric matches and localization:

Early on (Experiments 1 and 2) we found out that left and right matches are problematic because of the major changes in the localization of the sound when using ear phones. To the extent possible, we recommend that experimenters use symmetric matches. However, as evidenced by Appendices B and C, considerable care must be taken in figuring out how to test properties using symmetric matches.

Variance propagation:

Testing most of the properties involved using estimated values to estimate subsequent values. Since estimates were taken to be the average of obtained data, each has some variance and the resulting bias, but not the variance, propagates to the next stage of estimation. 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 that the variances in production judgments tend to be larger that those arising in matching. Thus when testing properties involving production judgments variance propagation 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. This will presumably lead to variance being accumulated from step to step. We followed this practice in Experiment 3 and for some participants in Experiment 4

Time order errors:

A potentially problematic time-order effect is one that is exemplified by participants' tendency to judge the second tone in the pair $\langle (x,x),(x,x)\rangle$ louder than the first (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. In that case, 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 participants heard $\langle (x,0),(0,x)\rangle$, in a fully counterbalanced fashion, but the results seemed completely dominated by the time order such that the second tone was almost always experienced louder than the first. Likewise, in a pilot experiment where counterbalancing was achieved through equal number of adjustment of the first and second tone, they resulted in very large variance in the data, requiring substantially larger sample size than we normal use.

B Testing bias using two-ear matches

The object is to show that testing for bias can always be done with two-ear matching. Let

$$(x,u) \sim (t,t) \quad (u,x) \sim (t',t'),$$

of which $x = 0$ or $u = 0$ are special cases.

Of course, if the symmetric matching theory is correct, one may use symmetric matches. The only issue is when the asymmetric theory is being tested. If that theory is correct, we know that both (14) and (15)) are satisfied and so

they may be assumed. Using them and doing a bit of algebra,

$$\frac{\Psi(t,t)-\Psi(t',t')}{\Psi(x,0)-\Psi(u,0)}=\frac{\Psi(x,u)-\Psi(u,x)}{\Psi(x,0)-\Psi(u,0)}=\frac{\gamma-1}{\gamma}.$$

So, for $x>u$, $\Psi(x,u)>\Psi(u,x)\Longleftrightarrow\gamma<1$. Hence, we see

$$\begin{array}{rcl}\text{left bias} & \Leftrightarrow & \gamma>1\Leftrightarrow(x,0)\succ(0,x)\Leftrightarrow(x,u)\succ(u,x)\Leftrightarrow t>t',\\ \text{no bias} & \Leftrightarrow & \gamma=1\Leftrightarrow(x,0)\sim(0,x)\Leftrightarrow(x,u)\sim(u,x)\Leftrightarrow t=t',\\ \text{right bias} & \Leftrightarrow & \gamma<1\Leftrightarrow(x,0)\prec(0,x)\Leftrightarrow(x,u)\prec(u,x)\Leftrightarrow t<t'.\end{array}$$

For $x<u$, left bias corresponds to $t<t'$, etc. So we conclude that two-ear matching is sufficient to test for bias.

C Testing production commutativity using two-ear matches

Our goal is to show that we may test production commutativity using two-ear matches. With no loss of generality [see remark leading to (18)] we can take the commutativity property in the form

$$[(x,x)\circ_p(y,y)]\circ_q(y,y)=[(x,x)\circ_q(y,y)]\circ_p(y,y). \tag{28}$$

Define

$$\begin{array}{rcl}(x,x)\circ_p(y,y) & \sim & (v,v),\\ (v,v)\circ_q(y,y) & \sim & (t,t),\\ (x,x)\circ_q(y,y) & \sim & (w,w),\\ (w,w)\circ_p(y,y) & \sim & (t',t').\end{array}$$

We show that (28) is equivalent to $t=t'$. Using

$$\Psi[(x,x)\circ_p(y,y)]-\Psi(y,y)=[\Psi(x,x)-\Psi(y,y)]\,W(p), \tag{29}$$

we have

$$\begin{array}{rcl}\Psi(t,t)-\Psi(y,y) & = & \Psi[(v,v)\circ_p(y,y)]-\Psi(y,y)\\ & = & [\Psi(v,v)-\Psi(y,y)]\,W(q)\\ & = & [\Psi(x,x)-\Psi(y,y)]\,W(q)W(p).\end{array}$$

Similarly,

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

By the commutativity of multiplication

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

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