# Evaluation of Time-Order Error Predictions from a Model of Global Psychophysics

Ragnar Steingrimsson and R. Duncan Luce
Institute for Mathematical Behavioral Science
University of California, Irvine

Ragnar Steingrimsson
Institute for Mathematical Behavioral Science
University of California, Irvine
Irvine, CA, 92697-5100
e-mail: ragnar@uci.edu

**Abstract**

The well known time-order error (TOE) phenomenon is studied in the light of a model arising from Luce's (2004) global psychophysical theory which asserts that when a respondent matches $z$ to be $p$ times a reference signal $x$, this amounts to satisfying the equation

$$W(p) = \frac{\psi(z) - \psi(\rho)}{\psi(x) - \psi(\rho)},$$

where $\psi$ is a psychophysical function, $W$ is a numerical weighting function, and $\rho$ is a reference signal. The experimental procedure that seems closest to this modeling is free-adjustment (FA) matching. We do not have a theory for the selection of the reference level, but it appears that separated blocks of standards ($S$) better tend to stabilize $\rho$ within a block than do interleaved ($I$) standards. Experiment 1 focuses on FA auditory data. The FA-$S$ data are consistent with this theory with the respondents split 50:50 between two predicted patterns. The FA-$I$ data accord less well with theory. Experiment 2 explores the 2-interval force-choice (2IFC) procedure, and these data are even less well fit by the theory. We believe this is because the reference point fails to stabilize under either $S$ or $I$. Creative new research is needed on reference levels, and data should be collected in other prothetic domains, e.g., brightness.
*Keywords:* weighting function, time-order error, audition, forced choice, loudness, matching, magnitude production.

The so-called time-order error, or *TOE,* for continua of physical intensities can be characterized as the tendency to perceive a second stimulus of two physically identical

ones presented successively as either more or less intense than the first.[1] Cast in terms of audition, if a tone of intensity $x$ (the standard) is presented first and a tone of intensity $z$ at the same frequency is, in some sense, matched to $x$ (a point of subjective equality), then a measure of a TOE when all measures are in decibels[2] is usually defined to be

$$\kappa_{dB} := z_{dB} - x_{dB}. \tag{1}$$

Under this definition, the TOE is said to be negative if $\kappa_{dB} < 0$ and positive otherwise.

This quantity has been estimated and studied using forced-choice procedures. Within that context, a complex of experimental findings is summarized by Hellström (1985) and more recent data and an ad hoc theory for them is offered by Hellström (2003).

However, the experiment can be done in a quite different way. One can present $x$ and have the respondent produce the $z$ that seems equally loud to him or her using, e.g., a free-adjustment (FA) procedure which arrives at the point of subjective equality faster than by constructing forced-choice psychometric functions. In either method, $x$ can be in the temporally first interval and $z$ in the second one, written $\langle x, z\rangle$, or in the opposite order, i.e., $\langle z, x\rangle$. We denote these two $z$'s as, respectively, $z_2$ and $z_1$.

This article has three purposes:

1. To describe a theoretical development for TOE that is based upon a model suggested by Luce (2002, 2004, 2008 erratum) and – favorably – evaluated experimentally by Steingrimsson and Luce (2005a, 2005b, 2006, 2007). See Luce and Steingrimsson (2008) for a fairly minor erratum that applies to the 4 empirical studies, which is also described in the section "Forms of the unknown functions."

2. In Experiment 1, we explore how well this theory accounts for free-adjustment (FA) data from 6 respondents who completed all experimental phases. Two sub-procedures are used: (i) Separated ($S$) standards, in which matches with the same standard are grouped into blocks of standards of increasing intensities, seems to stabilize the reference signal better than the other procedures that we have tried. (ii) Randomly interleaved ($I$) standards within a block, which has frequently been recommended as a way of averaging out errors, but which seem not to stabilize the reference signal.

3. In Experiment 2, we explore the more commonly used 2-interval forced choice (2IFC) using an Up-Down staircase procedure (Levitt, 1971) to match a tone to a standard, for both separated ($S$) and interleaved ($I$) staircases. Neither procedure, especially the often recommended interleaved ($I$) staircase method, seems to permit stabilization of the reference level.

[1] An analogue holds for signals presented at the same time but in different locations—each to an ear or in two spatial locations in vision. Any matching difference there are called space-order errors (SOE).

[2] Because the theory we discuss is formulated in terms of physical intensity, it is important to distinguish the two measures, $x$ in units of intensity and in decibels, $x_{dB}$.

This research was supported in part by National Science Foundation grants SBR-9808057 & BCS-0720288 to the University of California, Irvine. Any opinion, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional financial support was provided by the School of Social Sciences and the Department of Cognitive Sciences at UC Irvine. Discussions with several people have contributed to our efforts and we thank them: Joetta Gobell, Louis Narens, Hal Stern, and A. J. Marley.

Hellström (2003) laments how relatively little work on the TOE phenomenon has been done in recent years, especially when one recognizes its potential to affect materially the experimental outcomes. Indeed, we too have been remiss in this regard: In a series of four articles, Steingrimsson and Luce (2005a,b; 2006, 2007, 2008 erratum), which employed auditory matching and ratio production, we built into our procedures averaging over respondents in an attempt to wash out possible effects of the TOE on the experimental outcomes. But we undertook no systematic exploration of TOE in our experimental paradigm and, to our knowledge, that has not been done by anyone else. So another aim is to remedy this lacuna to some extent.

In brief, we show that:

• The individual FA matching data with separate blocks of standards are consistent with the theory, whereas the data averaged over respondents are not. In part, averaging FA-$S$ fails because two quite different types of individuals exist when matching the first signal presented to the second one. Which type depends upon how the respondent interprets the instructions.

• Although the data from FA-$S$ matching are well accounted for by the theory, the data from the other 3 cases (FA-$I$, 2IFC-$S$, and 2IFC-$I$) are not well explained. In particular, the 2IFC data of individual respondents are quite varied and, for the most part, do not seem very consistent with the theory.

• We certainly do not understand very well what controls the reference points that arise in the theory, but we show that they vary considerably with procedural changes. We very much need a theory for them since we believe that their selection may account for the observed differences.

• Appendices A and B show for FA and 2IFC the average dB data, which is often used in the literature; these are neither very consistent with the theory nor, are in any way, representative of the individual data. Appendix C demonstrates the existence of a time order-error phenomenon in brightness.

## Background

*Historical*

In his comprehensive review of the TOE, Hellström (1985), writes

> Among factors that have been shown to be of importance for the direction and magnitude of the TOE are level of stimulus magnitude (e.g., Bartlett, 1939; Needham, 1935; Woodrow, 1933), length of ISI (e.g., Needham, 1935), and intensity of stimulation interpolated into the ISI (e.g., Ellis, 1973b; Lauenstein, 1933). These factors interact in a complex way with amount of training (Köhler, 1923; Needham, 1934a; Woodrow, 1933), and stimulus duration (Inomata, 1959), as well as with the particular set of ISIs used in the experiment (Wada, 1937). [...] The picture is thus very complicated, and the outcome of a TOE experiment can indeed be hard to predict. Besides, the TOE effects are often (but not always) rather small, and they vary considerably from subject to subject. (p. 36)

Hellström's (1985) notes that, in general, the TOE tends to be negative, i.e. in (1) $\kappa_{\mathrm{dB}} < 0$, for the higher intensities and positive for the lower ones. In audition, the TOE is

generally assumed to be negative, i.e. in (1), $\kappa_{\rm dB} < 0$ across the intensity spectrum (e.g. Stevens, 1975, pp. 139-141), but that does not really seem to be correct.

Hellström does not mention any TOE data based upon a free-adjustment matching procedure; thus, it is unclear from the literature what will be found using that type of matching. That is one of our topics, but first we summarize the major representations derived from testable behavioral assumptions by Luce (2002, 2004, 2008 erratum) and use that to arrive at some TOE predictions.

*A psychophysical theory*

In an auditory realization of the general situation, let $x$ denote the intensity of a pure tone less the threshold intensity (not the dB difference) that is presented to the left ear. Using the same frequency and phase for the right ear, denote by $u$ that intensity less the right ear threshold intensity. Pairs of auditory signal intensities $(x, u)$ are ordered by loudness. In principle, we could study the full binaural case, but in practice that greatly increases the theoretical complexities and we decided not to pursue it further here and focus only on the two monaural cases $(x, 0)$ and $(0, u)$.

Another primitive is a form of magnitude production where the respondent is asked to adjust, say, the left ear intensity $z$ so that it seems to be $p$ times as loud as $x$. Clearly, $(z, 0)$ is a function of $x, \rho, p$ which can be made clear by using an operator notation $(x, 0) \circ_p (\rho, 0) := (z, 0)$. We use both notations, the more explicit one when we think it improves clarity. For a right ear procedure, the notation change is trivial.

The representation that Luce (2004, 2008 erratum) axiomatized, establishes the existence of an order-preserving psychophysical function $\Psi(x, u)$, i.e.,

$$(x, u) \succsim (y, u) \Leftrightarrow \Psi(x, u) \geq \Psi(y, u), \tag{2}$$

$$\Psi(0, 0) = 0, \tag{3}$$

and a numerical distortion function $W(p)$ such that two major relations hold. To simplify the writing we use the abbreviated notations

$$\psi_l(x) := \Psi(x, 0) \tag{4}$$

$$\psi_r(x) := \Psi(0, x) \tag{5}$$

The relation of binaural matches to monaural ones was shown to be

$$\Psi(x, u) = \psi_l(x) + \psi_r(u) + \delta\psi_l(x)\psi_r(u) \quad (\delta = 0, 1). \tag{6}$$

For $\delta = 0$, this is pure additivity, whereas for $\delta = 1$ it is not and is called p-additive[3,4]. The second result captures the ratio nature of the productions which for monaural signal presentations as:

$$W(p) = \frac{\psi_i\left(x \circ_{i,p} \rho\right) - \psi_i(\rho)}{\psi_i(x) - \psi_i(\rho)} \quad (i = l, r,\ x > \rho \geq 0). \tag{7}$$

[3]This term stands for polynomial additive because this is the only polynomial with $\Psi(0, 0) = 0$ that can be transformed by $\Psi^* := \ln(1 + \Psi)$ into an additive representation.

[4]This part of the theory, by itself, also admits the case of $\delta = -1$, but other features of the psychophysical theory rule out that case (Luce, 2004).

Because we want to be consistent in using lower case Greek letters for model parameters and functions, we denote the reference point by $\rho$, which is explained more fully in the section "A Theory of Time-Order Errors." The signal $\rho$, which may either be presented by the experimenter or be a "creation" of the respondent, is used to establish the "intervals" to be compared. Steingrimsson and Luce (2005a,b) provided empirical tests of the behavioral properties giving rise to the above representation, and they seemed to be sustained.[5]

This article draws primarily on (7).

*Forms of the unknown functions*

*Form of $\psi_i$.* Steingrimsson and Luce (2006) examined one possible mathematical form for $\psi_i, i = l, r$, namely power functions. Such an assumption can be defended in several ways. A qualitative condition, called multiplicative invariance, was tested and the power function form was sustained for 12 of 22 respondents (Steingrimsson & Luce, 2006). Additional data collected for this article, ended up with the power form holding for 19 of 32 respondents. We are reporting here data for 6 of these 19. The others failed to complete all of the experiments or participated in pilot studies only.

The error Luce made in the 2004 article was failing to realize that the behavioral property of bisymmetry holds not just for $\delta = 0$ but also for $\delta = 1$ (Luce, 20008). In the latter case, there theory shows that the existence a constant $\eta$ such that

$$1 + \psi_l(x) = (1 + \psi_r(x))^{\eta}. \tag{8}$$

So when the $\psi_i$ are power functions and $x$ is sufficiently large, this is approximately the same as

$$\psi_l(x) = \alpha_l x^{\beta_l} \text{ and } \psi_r(x) = \alpha_r x^{\beta_r} \text{ where } \beta_r = \eta\beta_l.$$

*Forms of $W$.* Steingrimsson and Luce (2007) focused on the form of the weighting function, and one of their important realizations was that it is wrong to assume that $W(1) = 1$, which others and, initially, we had done, leading to some dubious inferences based on that assumption (Ellermeier & Faulhammer, 2000; Narens, 1996; Zimmer, 2005). Once we recognized what our data were telling us, namely, $W(1) \neq 1$, we realized that TOE should appear in production data. Not all respondents' data were fit by power functions, but the remainder were fit by the more general Prelec function $\exp\left(-\beta\left(-\ln p\right)^{\alpha}\right)$. Here we restrict attention to those for which $W$ is a power function.

## A Theory of Time-Order Errors

Somewhat in accord with the traditional approach to TOE, (1), suppose that we consider measuring TOE in terms of the psychophysical function, namely,

$$\kappa_{\psi}(x) := \psi(z) - \psi(x), \tag{9}$$

[5]For those familiar with utility theory (7) has a familiar flavor. Solving for the matching term:

$$\psi_i\left[x \circ_{i,p} \rho\right] = W(p)\psi_i(x) + [1 - W(p)]\psi_i(\rho).$$

With $p \leq 1$, this is the weighted utility representation of a binary gamble in which $x$ occurs with probability $p$ and $\rho$ with probability $1 - p$. In utility theory, usually $W(1) = 1$.

rather than in terms of dB. Recall that the model (7) asserts,

$$W(p) = \frac{\psi_i(z) - \psi_i(\rho)}{\psi_i(x) - \psi_i(\rho)} \quad (i = l, r), \tag{10}$$

where the experimenter presents $x$ and $p$ to the respondent and asks him or her to find the $z$ that makes the ratio seem to be $p$.

It is clear from (10) that $z$ and $\kappa_\psi$ also depend upon the reference signal $\rho$, but as a parameter, not a variable. So we mostly suppress $\rho$ in the notation for $z$.

*Second stimulus matched to first*

In (10), set $p = 1$, and define $\varpi := W(1)$, and the subscript 2 on $z_2$ and $\rho_2$ emphasizes that the respondent is adjusting the second signal to the first one. This yields

$$\frac{\psi_i(z_2) - \psi_i(x)}{\psi_i(x) - \psi_i(\rho_2)} = \varpi - 1,$$

which rearranged becomes

$$\kappa_{\psi,2,i}(x) = \psi_i(z_2) - \psi_i(x) = (\varpi - 1)\left(\psi_i(x) - \psi_i(\rho_2)\right). \quad (i = l, r) \tag{11}$$

Thus, (11) predicts that $\kappa_{\psi,2,i}(x)$ is linear with $\psi_i(x) - \psi_i(\rho_2)$ with slope $\varpi - 1$, facts of importance in evaluating the theory.

Of course, (11) is the same as the dB measure usually used if and only if the psychophysical functions $\psi_i$ are logarithms (Fechner's law).[6] Instead, we will study power functions where $\beta_i$ is considerably less than 1, which means the $\psi_i$ are not too greatly different from logarithms.

*First stimulus matched to second*

In contrast to the previous section the present case is somewhat more complex to analyze because in the representation (10) is ambiguous as to whether the numerator corresponds to the second presentation or to the signal $z$ that matches $x$. Quite different predictions follow for these two interpretations and, as we will see, they seem to correspond to a difference among respondents. Put another way, a respondent confronted by two tones has the option of which is taken as the standard.

*Model numerator corresponds to matching signal.* Assume that the matching signal, $z_1$, corresponds to the numerator, then the calculation is identical to (11) but with $z_2$ replaced by $z_1$ and $\rho_2$ by double subscript $\rho$, to be explained The first subscript, 1, reflects that $\rho$ corresponds to the location of $z_1$ in the first interval. The second subscript, in this case 1, indicates that signal 1 is in the matching role. Solving,

$$\begin{aligned}
&\frac{1}{\varpi} = \frac{\psi_i(z_1) - \psi_i(\rho_{11})}{\psi_i(x) - \psi_i(\rho_{11})} \\
&\iff \varpi\left[\psi_i(x) - \psi_i(\rho_{11})\right] = \psi_i(z_1) - \psi_i(\rho_{11}) \\
&\iff \kappa_{\psi,2,i}(x) = \psi_i(z_1) - \psi_i(x) = (\varpi - 1)\left(\psi_i(x) - \psi_i(\rho_{1,1})\right). \quad (i = l, r)
\end{aligned} \tag{12}$$

[6] Using the log measure, all degrees of freedom in the representation cancel, whereas with power functions the exponent $\beta$ remains.

Thus, we expect the plots for (11) and (12) to agree. Denote the estimated slope $\widehat{S}_{1,1}$, which is our estimate of $\varpi - 1$, and is positive when $\varpi > 1$ but negative when $\varpi < 1$. This is in contrast to the next version.

*Model numerator corresponds to second presentation.* Because the adjusted signal $z_1$ is presented prior to presenting the standard $x$, a calculation similar to that of (11) yields:

$$\kappa_{\psi,1,i}(x) = \psi_i(z_1) - \psi_i(x) = \left(\frac{1-\varpi}{\varpi}\right)\left(\psi_i(x) - \psi_i(\rho_{1,2})\right), \quad (i = l, r) \tag{13}$$

and $\rho_{1,2}$ is the reference signal when the variable signal is in the first interval but the numerator corresponds to the second presentation. So the estimated slope $\widehat{S}_{1,2}$ approximates $\frac{1-\varpi}{\varpi}$, which is, positive when $\varpi < 1$ and negative when $\varpi > 1$.

We explore experimentally how well these equations, (11) and (13), account for what we find experimentally.

*A slope prediction from (11) and (13) .* To that end, if we multiply together the slopes of (11) and (13) and take into account the fact that $\varpi > 0$, we see that

$$(\varpi - 1)\left(\frac{1-\varpi}{\varpi}\right) = -\frac{(\varpi - 1)^2}{\varpi} \leq 0. \tag{14}$$

Thus, either $\varpi = 1$, which means both slopes are 0, or the two slopes must be of opposite signs – one positive and one negative. We make heavy use of that fact.

*Qualitative properties predicted.* In each case, we predict that once $\beta$ and the reference point $\rho$ are estimated we should see straight lines. So consider these estimates.

Hellström (1985) notes: "...[T]he TOE effects...vary considerably from subject to subject." (p. 36). So far as we are aware, this has been documented only via estimates of variances in the data. We, at least, had not fully understood just how large the differences really can be. This is demonstrated in Experiments 1 and 2 where the individual and averaged data for the 6 respondents are presented (Figures 1–4). The TOE measure $\kappa_k^\beta = {z_k}^\beta - x_n^\beta$ is plotted directly a function of $x_n^\beta - \rho_i^\beta$, where $n = 1, 2, 3$.

To explore whether the data come out according to the predictions of the sections on matching the second/first stimulus matched to first/second we should find $\kappa_k^\beta = {z_k}^\beta - x_n^\beta$ as a function of $x_n^\beta - \rho_i^\beta$ forms approximately straight lines. Thus the plotting requires an estimate of $\beta$ and $\rho_i$. Despite devising and trying several methods to estimate $\beta$ from our data for individuals, nothing useful resulted. However, because group averages from a myriad of studies have established $\beta \approx 1/3$ when intensity is measured in sound pressure (see Stevens, 1975, p. 15, for an overview). Presumably the individual fits would be improved a bit by making $\beta$ vary somewhat by respondent.

The reference signal can easily be estimated by finding

$$z^\beta - x^\beta = 0 = x^\beta - \rho^\beta$$
$$\Leftrightarrow z^\beta = x^\beta = \rho^\beta.$$

This usually requires interpolation or extrapolation, which is easy to do when the data are nearly linear.

Next we examine the slopes of the data both to estimate $\varpi$ and to decide which of the models seems to describe the data. Because we will collect both kinds of data – matching the second to first signal, and matching the first to the second signal – we should see one of two patterns:

1. Both lines arising in (11) and (12) are samples from a single line with slope $\varpi - 1$. So the estimated slopes should satisfy $\widehat{S}_{1,1} = \widehat{S}_{2,2}$.

2. When (11) and (13) apply, one slope is positive and the other is negative and the value of $\varpi$ is estimated using the calculated product

$$\sigma = \widehat{S}_{2,2}\widehat{S}_{1,2}. \tag{15}$$

The variability of the data shows up in the estimate $\sigma$. Putting (15) in (14) yields

$$\sigma = -\frac{(\varpi - 1)^2}{\varpi}, \tag{16}$$

which yields

$$\varpi = \frac{2 - \sigma \pm \sqrt{\sigma(\sigma - 4)}}{2}. \tag{17}$$

There are two integer roots for Eq. (17), namely $\sigma = 4$, $\varpi = -1$ and $\sigma = 0$, $\varpi = 1$. Of course, there are additional non-integer roots.

## Experiments

We present results as two experiments in which subjective equality is sought through two procedures: matching using a free adjustment (FA) procedure and a 2-interval forced choice (2IFC) procedure where, in loudness, a variable tone is compared to a standard tone. In each case there are two variants depending upon how we handle the presentation order of standards, separated ($S$) and randomly interleaved ($I$).

### *Common experimental methods*

*Respondents.* A total of 19 students – graduate and undergraduate – from the University of California, Irvine, and coauthor R22 participated in the a series of pilot studies that ultimately led to our final design. Of these, we report only the data of the 6 respondents who provided data in all of the studies being reported in this article – two of these also provided data for one control condition. The remaining respondents participated in pilot studies only.

All respondents reported normal or corrected-to-normal hearing. To facilitate group analysis of the data, respondents were evaluated for having psychophysical functions well approximated by a power-function form (as evaluated by multiplicative invariance, see Steingrimsson & Luce, 2006, for details— in particular, method section of their Experiment 1 (§ 3.2.1)).

All respondents, except the participating coauthor, 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, 2002). Consent forms and procedures were approved by UC Irvine's Institutional Review Board.

*Stimulus.* The basic signal was a 1000 Hz sinusoidal tone presented for 100 ms, which included 10 ms on and off ramps. The basic stimulus consisted of two signals (tones) separated by 450 ms.

Recall that the theory is cast in terms of intensities less threshold. So for a left ear threshold of $x_T$ and a right ear one of $u_T$, the effective stimulus $(x, u)$ consists of $x = x' - x_T$ and $u = u' - u_T$ where $(x', u')$ are the actual intensities presented. However, because all signals were well above threshold and the respondents were selected for normal hearing, the error in reporting intensities $(x', u')$ in dB SPL (henceforth abbreviated dB) is negligible.

*Apparatus.* Stimuli were generated digitally using a personal computer and played through a 24-bit digital-to-analog converter (RP2.1 Real-time processor, Tucker-Davis Technology). Presentation level was controlled by built-in features of the RP2.1 and stimuli were presented over Sennheiser HD265L headphones to the listener seated in an individual, single-walled IAC sound booth located in a quiet lab-room. A safety ceiling of 90 dB was imposed in all experiments.

*Procedure.* Experiments were conducted in sessions that lasted at most one hour each. The initial session was devoted to obtaining written consent, explaining the task, and running practice trials. All respondents trained for one session on each task. Rest periods were encouraged with both their frequency and duration under respondent control.

*Interleaved and separated conditions.* Typically, experimenters have claimed that interleaving experimental conditions is a form of error averaging; however, in the context of the present theory it has more to do with disrupting the stabilization of the reference signal. This fact led us to also study two distinct procedures. Assuming standards $x_1, x_2 ...., x_n$, they were:

**Interleaved:** Standards were interleaved in a random fashion within a single block.

**Separated:** Standards were presented in individual blocks of trials and in an ascending order of intensity within a session.

Thus, in the interleaved condition, respondents had frequent changes in standards, whereas in the separated condition, the changes in standard occurred infrequently.

*Experiment 1: Free adjustment matches*

The goal of this experiment is to use a free adjustment procedure (FA) to obtain matches of the first tone to the second and of the second to the first. An advantage of this procedure is that it is more rapid than estimating the psychometric function of 2IFC. We will see that there are additional advantages.

*Method.*

• Matches were of two types. In the first the first tone is the standard and the second one is varied (matching $2^{nd}$); in the second, the first tone is varied and the second is the standard (matching $1^{st}$).

• For the matching $2^{nd}$ – matching $1^{st}$ is analogous – two tones (see the Stimulus section) were presented, the standard being the first and the variable one second. The respondent had the choice of either repeating the tone sequence, increasing it, or decreasing it using key presses to select increments of *big*, *medium*, *small*, and *extra-small* steps, which corresponded to changes of $4, 2, 1,$ or 0.5 dB respectively. Following the key-press, the adjusted tone sequence was played. Respondents made as many adjustments as desired except to adjust intensity beyond 90 dB safety limit, until a satisfactory match was obtained, indicated by a key-press.

• Three standards were used: $x_1 = 58$, $x_2 = 66$, and $x_3 = 74$ dB. The three standards and two matching conditions give rise to 6 matching conditions.

• In the Interleaved ($I$) condition, the three standards were randomized within the block of trials; the matching conditions ($1^{st}$ or $2^{nd}$) were run in separate blocks. In the Separated ($S$) condition, 24 matches were collected for $x_1$ standard, followed by 24 of $x_2$ and then 24 if $x_3$. The matching conditions were run in strictly alternating order within the block.

• Thus there are a total of four conditions, FA-$I_i$ and FA-$S_i$, where $i = 1, 2$ indicates the matching condition ($1^{st}$ or $2^{nd}$).

*Individual and averaged FA-S data.* Data from six respondents are depicted in Figure 1. In the figure, the data from the FA-$S$ procedure are presented by both plots for each individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$ as detailed in the section "Qualitative properties predicted."

Observe that these data form two distinct patterns, exactly as characterized in the section "First stimulus matched to second". The left column of 3 respondents has the crossed pattern expected when the model numerator corresponds to second presentation. In contrast, the right column has closely similar negative slopes as the section "Second stimulus matched to first" and on when the model numerator corresponds to the matching signal. The FA-$S$ data show that half of the respondents ($R$ $10, 22, 35$) seem to accord with the $z_1$ model described in "Model numerator corresponds to matching signal," whereas the other three ($R$ $47, 59, 60$) agree with that described n the following section, "Model numerator corresponds to second presentation." It is also very evident that the averaged data are a very misleading summary of the situation.

For R10, 22, and 35, $\varpi$ is arrived at by way of Eqs. (15) and (17), which yields two estimates for $\varpi$. Since for all three $\widehat{S}_{2,2} > 0$ and $\widehat{S}_{1,2} < 0$, we conclude that the slope patterns indicate the solution of $\varpi > 1$. Table 1 provides for the FA-$S_i$ data, estimates of the reference point $\rho$ and $\varpi = W(1)$ assuming $\beta = 0.3$.

*Individual and averaged FA-I data.* Data from six respondents are depicted in Figure 2. In the figure, the data from the FA-$S$ procedure are presented by both plots for each individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$ as detailed in the section "Qualitative properties predicted."

By contrast with the FA-$S$ data, the FA-$I$ data, except for R10, exhibit no such consistent pattern.

*Discussion of FA data.* For FA data there is considerable difference between the interleaved and separated adjustments. We suspect that these procedures affect sharply the

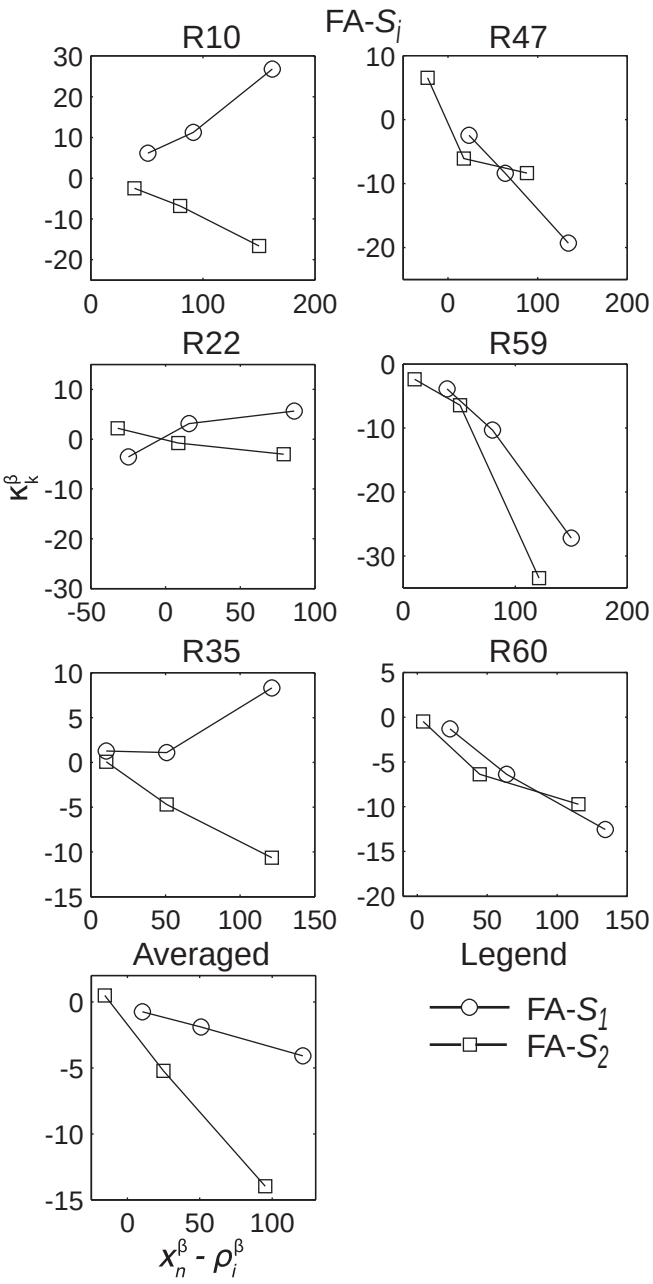


*Figure 1.* Data from the FA-$S$ procedure are presented by plots of individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legend is given in the lower right corner.

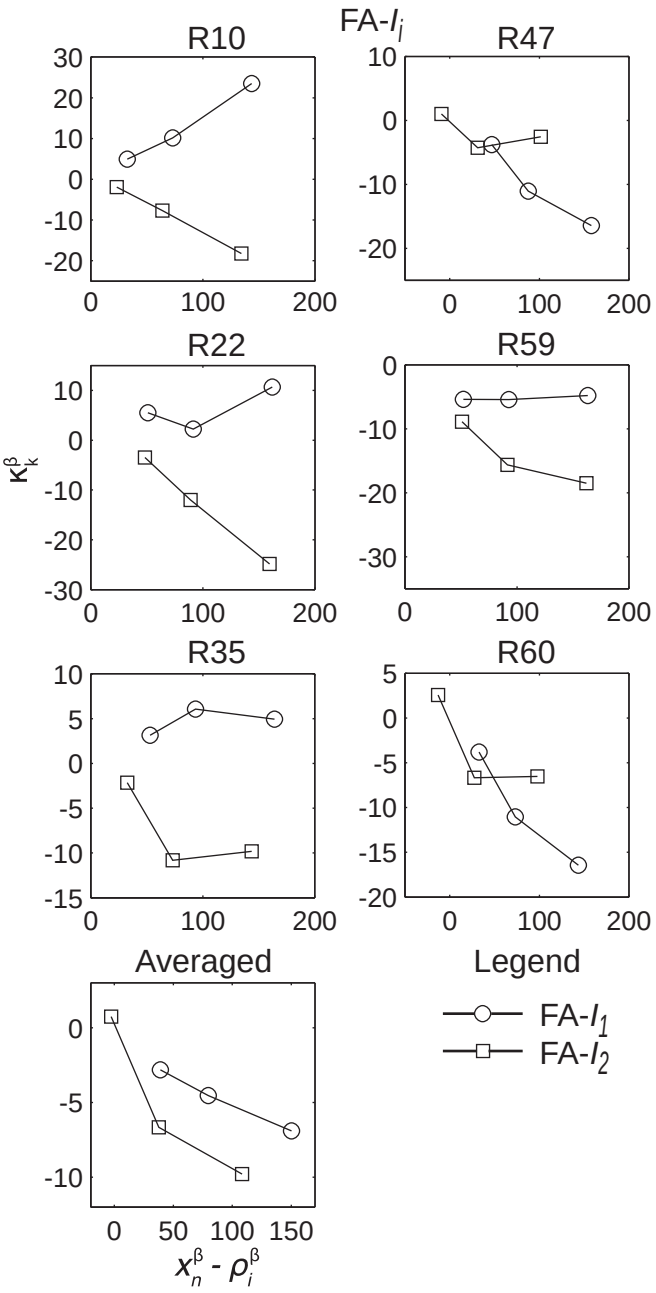


*Figure 2.* Data from the FA-$I$ procedure are presented by plots of individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legend is given in the lower right corner.

| Respondent | $\varpi$ | $\rho$ in dB |
|---|---|---|
| 10 | 1.17 | 30.0 |
| 22 | 1.06 | 64.0 |
| 35 | 1.08 | 55.0 |
| 47 | .86 | 53.4 |
| 59 | .75 | 47.5 |
| 60 | .91 | 56.5 |

Table 1: The table lists, for each respondents, the estimated $\varpi$ and $\rho$ in dB for the FA-$S$ data. The method of estimation is detailed in the section "Qualitative properties predicted."

value of the reference point $\rho$ that is being used. There can be no doubt that we badly need greater understanding – a theory – of how these reference point are "chosen". Nonetheless, the qualitative analysis gives quite strong support for the theory in the case of FA-$S$.

*Experiment 2: Two-interval, forced-choice matches*

The goal of this experiment was to seek a match (point of subjective equality) of either the first tone to the second or the second to the first using 2-interval forced-choice procedure, which is typical of earlier estimates in the literature.

*Method.* Matches of a variable tone to a standard were collected using a 2-interval forced-choice paradigm (2IFC) based on the standard Up-Down method (Levitt, 1971). Respondents hear two tones and indicate, using one of two keys, whether the first or the second tone was louder.

• Matching conditions were of two types. In the first, the first tone is the standard and the second is varied (match $1^{st}$) in the second, the first tone is varied and the second is standard (match $2^{nd}$). In the case where the first tone is standard and the respondent indicates the first tone is louder, the second tone in increased by 1 dB, the other three cases are analogous.

• Two stimulus condition sets were used, a broad range one consisting of $x_1 = 58$, $x_2 = 64$, $x_3 = 70$, $x_4 = 76$, $x_5 = 82$ dB, and a narrow one consisting of $x_1 = 70$, $x_2 = 73$, $x_3 = 76$, $x_4 = 79$, $x_5 = 82$ dB. The narrow range is simply the broader range clipped from below.

• A separate staircase was run for each standard and matching condition, or 10 staircases for each of the two range condition. Each staircase consisted of 65 trials. The two range conditions were run in separate sessions.

• In the Interleaved ($I$) condition, the 10 staircases (of a given range condition) were randomly interleaved within a session. In the Separated ($S$) session, the two staircases with the same standard were randomly interwoven and presented separately such that the standards were in strictly increasing intensity order.

• Labeling the matching conditions $i = 1, 2$, the two range conditions, $j = 1, 2$, there were a total of eight conditions, 2IFC-$I_{i,j}$ and 2IFC-$S_{i,j}$. For 2IFC-$I_{i,j}$ the conditions are as follows:

| Range/Matching | Broad: $j = 1$ | Narrow: $j = 2$ |
|---|---|---|
| Match $1^{st}$ : $i = 1$ | 2IFC-$I_{1,1}$ | 2IFC-$I_{1,2}$ |
| Match $2^{nd}$ : $i = 2$ | 2IFC-$I_{2,1}$ | 2IFC-$I_{2,2}$ |

The 2IFC-$S_{i,j}$ conditions are analogous.

• An estimate from a staircase was arrived at by discarding the first 15 trials, with the average of the rest taken as the final estimate.

*Individual and averaged 2IFC data.* Data from six respondents are depicted in Figures 3 and 4. In the figures, the data from the 2IFC-$S$ and 2IFC-$I$ procedures are presented by both plots for each individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$ as detailed in the section "Qualitative properties predicted."

R10 seems to exhibit a fairly consistent pattern independent of procedure, and one that is broadly consistent with our theory as are the averaged data for the 2IFC-$I$ procedure. However, none of the other 5 respondents seems to exhibit such a consistent pattern. In fact, the individual data reveal such large variability that they do not seem to exhibit any one clear pattern.

Lu, Williamson, and Kaufman (1992) hypothesized that respondents' reference gravitated to the central tendency of the intensity environment of the experiment. They collected data using a 2IFC procedure much like the current one. A direct prediction from this hypothesis is that maximal magnitude of the TOE should diminish as the stimulus range becomes smaller. The 2IFC-$S$ condition places respondent in a nearly constant stimulus range environment for extended period of time. Therefore, the TOE should be substantially diminished in this condition. Again, for R10, this prediction holds and though the slopes for 2IFC-$S$ are smaller than for 2IFC-$I$the effect is neither consistent nor large. Another exception would be that the narrower range would simultaneously show smaller TOE as well as shift the reference point higher. Again, the data are not broadly consistent with this prediction.

Lu et al. (1992) discussed the TOE in a context of decay of echoic memory. They did note that their data were sensitive to sequential effects. For this reason, we were led to try giving the respondents a prescribed reference point.

*Reference tone.* The apparent importance of the reference tone $\rho$ motivated us to carry out what amounts to pilot work in which we presented a reference tone prior to each staircase in the 2IFC procedure.

The procedure was the same as in the case of 2IFC-$I_{i,1}$. There were two tonal conditions.

**Fixed Reference** A tone of 70 dB of 100 ms duration is presented 600 ms before each trial (hence inter-tone between reference and the first tone of 500 ms). This is labeled 2IFC-$FR_{i,1}$.

**Random Reference** The same as the Fixed Reference, except the intensity of the presented tone was an a chosen as a random integer value for intensity, from 54 to 84 dB. This is labeled 2IFC-$RR_{i,1}$.

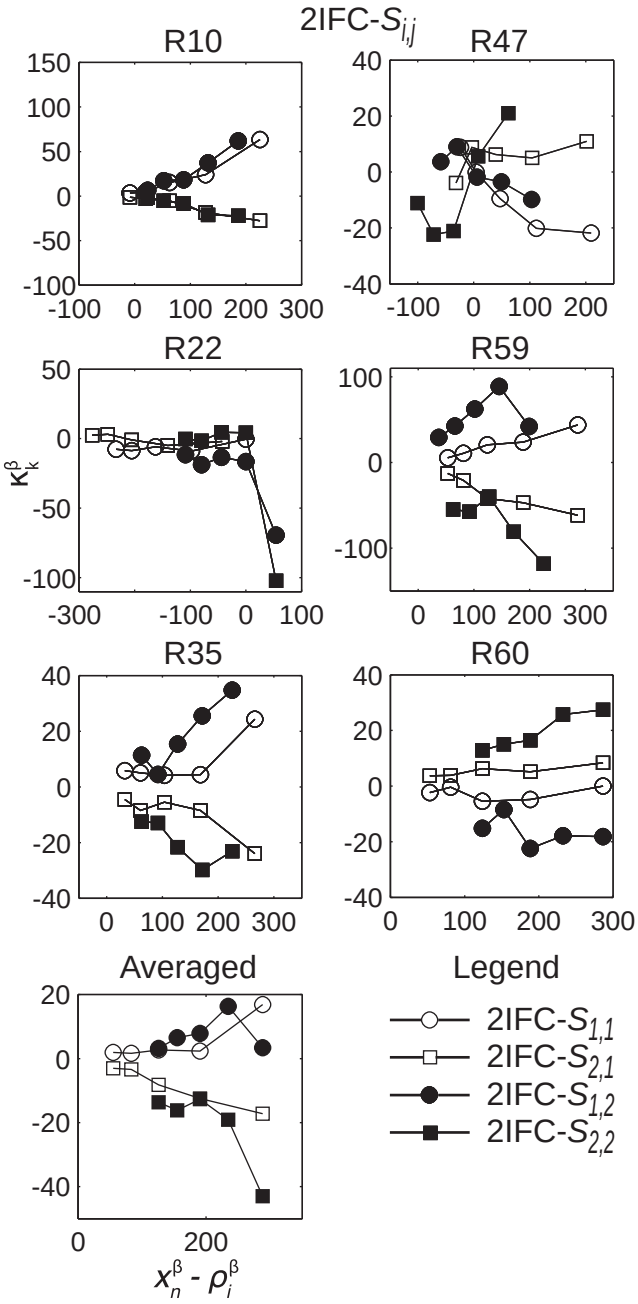


*Figure 3.* Data from the 2IFC-$S$ procedure are presented by plots of individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legend is given in the lower right corner.

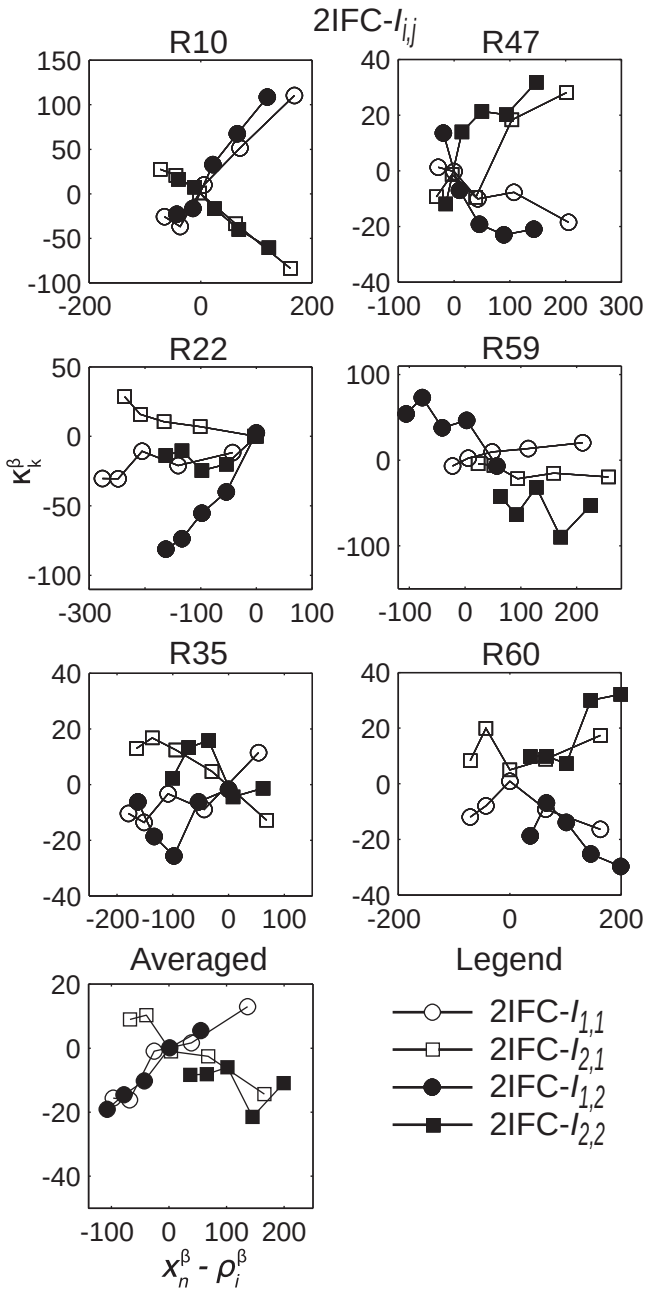


*Figure 4.* Data from the 2IFC-$I$ procedure are presented by plots of individual data as well as their averages. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legend is given in the lower right corner.

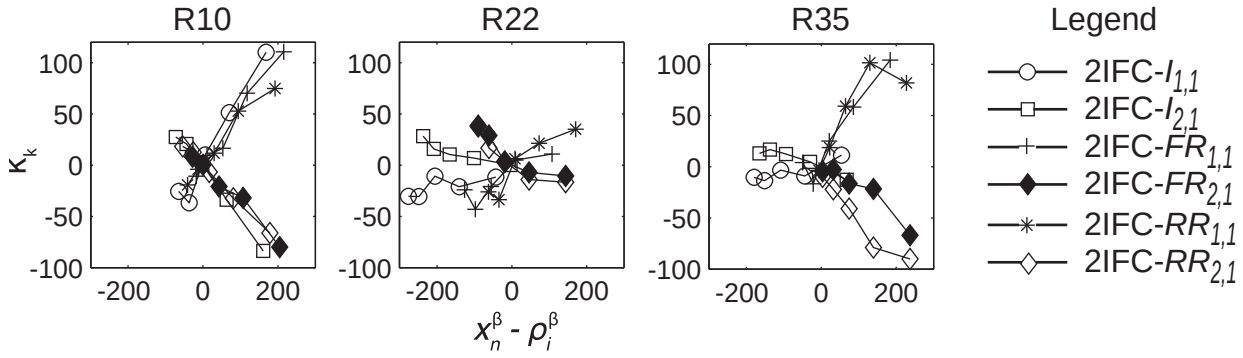


*Figure 5.* Data for the reference condition are presented by plots of individual data as well as their 2IFC-$I_{i.1}$ counterpart. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legend is given in the lower right corner.

Three of the 6 respondents provided data and the results are plotted in Figure 5.

Once again, R10 exhibits the same pattern as before whereas for the other two the pattern is less clear. However, it is quite clear that the manipulation has no clear effect on the TOE and hence is not a solution towards stabilizing the reference point.

*Qualitative analysis.* On examining Figures 1 and 2 from the FA experiment and Figures 3 and 4 from the 2IFC experiment in terms of the qualitative analysis described as items 1 and 2 in the section "Qualitative properties predicted," it is immediately clear that none of the 2IFC data are very well described by the theory. Those of FA-$I$ are somewhat better described, but not nearly as well as are the FA-$S$ data. Using the estimation method described in the section "Qualitative properties predicted", we presented in Table 1 estimates of $\rho$ and $\varpi$ from the FA-S data.

*Discussion of 2IFC data.* The 2IFC method is very commonly used in psychology. There are therefore two causes for concern.

The first is simply that neither 2IFC-$S$ nor 2IFC-$I$ is in very good accord with the model, whereas the FA-$S$ data do seem to be. That suggests that FA methods may be more satisfactory despite the lore of the field that 2IFC is the preferred method.

Assume that the problem, at least for the 2IFC method, is that we have failed to achieve a suitable way to stabilize the reference point $\rho$. It was that realization that led us to run 2IFC with experimental prescribed $\rho$'s.

The second concern is the fact that in 2IFC-$I$, the magnitude of the TOE reached $\sim 6$ dB (see Fig. B1, Appendix B) and that the TOE behavior for 2IFC-$S$ were radically different, which revealed how dependent the TOE is on experimental context. Randomization of conditions (along with other experiment specific considerations) is commonly taken as a way to avoid context effects in psychological research; however, our results show unequivocally that randomization of conditions is not a "cure" for the TOE.

A further worry is that since the TOE can be so large and so varied depending upon context, an experimental outcome can potentially be radically altered by the simple act of

adding or removing conditions, or the ordering of conditions.

Since the averaged data show some regularities, it may be possible to quantify in a given experiment the potential influence of the TOE and then extract in some fashion from the averaged results. Yet, that is a cumbersome avenue, involving a great deal of variability. Clearly the TOE is quite different in the free-adjustment task of Experiment 1 hence, one way to ascertain context independence in an experimental results may be to carry the experiment out using different methods. Should the results concur, there is a good chance they are context-independent. When this is not feasible or obvious methodologically, the same experiment may be carried out using very different sets of stimuli and/or orderings of conditions, comparing the results of those.

Because averaging over individuals entails the implicit assumption that respondent's data are no less than linearly related (e.g., Luce, 1995), therefore, the individual differences of the TOE are of great interest.

## Conclusions

Our data demonstrate clearly that the TOE is an important factor in experiments, and potentially large enough to alter conclusions radically. It is an effect that can be neither averaged nor randomized out of consideration. Of the four procedures we have run, it is very clear that FA-$S$ yields data are well described by our global theory, for which there is considerable independent, supporting evidence. Evidently, that procedure, more than the other three, induces a stable reference point $\rho$. We are in need of a far better theoretical understanding of the nature of reference point selection and how best to stabilize it in experimental practice. These are major open problems.

However, in none of the cases did we eliminate the TOE. Providing a reference tone on each trial had either no or an unclear effect, suggesting that whatever benefit, in terms of the magnitude of the TOE, arises from a homogenous intensity environment that operates on a relatively long time-scale. The free adjustment matches suggest that this time-scale may begin to affect judgments on the order of a few seconds, but many minutes only serve to make the TOE similar across intensities.

Of course, the FA-$S$ method runs counter to what is widely believed to be an important practice of randomizing conditions. In this condition, the respondent finds him/herself in a homogenous intensity environment for about 10 minutes at a time and the subsequent change in environment is always ascending with relatively small intensity change. In this sense, each point in the averaged or individual data is collected in a homogenous intensity environment, whereas the opposite is true for both of the 2IFC conditions.

Of course, our inability to discover a suitable estimation procedure for the exponent $\beta$ of the power function psychophysical function for individual respondents may have kept us from giving more sensitive evaluation of our theory. Despite that fact, we feel that some of the dramatic effects we have demonstrated make clear that all too many studies have swept the TOE under the carpet when, in fact, it may have had a serious impact.

In conclusion, we add that some very limited pilot work on brightness shows that TOE to be an equally important phenomenon as for audition. We suspect that this holds for any scale of sensory intensity (prothetic scale).

## References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, **57**, 1060–1073.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436..

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

Hellström, Å. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference *Perception & Psychophysics*, **65**, 1161-1177.

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

Lu, Z.-L.,Williamson, S. J., Kaufman, L. (1992). Behavioral Lifetime of Human Auditory Sensory Memory Predicted by Physiological Measures. *Science*, **258**, 1668–1670.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467–477.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Reviews of Psychology*, **46**, 1–26.

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 asymmetric matching of joint presentations. *Psychological Review*, **111**, 446–454.

Luce, R. D. (2008). Correction to Luce (2004). *Psychological Review*, **115**, 601.

Luce, R. D., Steingrimsson, R. (2008). Note on a changed empirical inference in several Steingrimsson and Luce articles due to C. T. Ng's correction of an error in Luce (2004). *Journal of Mathematical Psychology*, **52**, 263–264.

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

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

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

Steingrimsson, R., & Luce, R. D. (2006). Empirical evaluation of a model of global psychophysical judgments: III. A form for the psychophysical function and intensity filtering. *Journal of Mathematical Psychology*, **50**, 15–29.

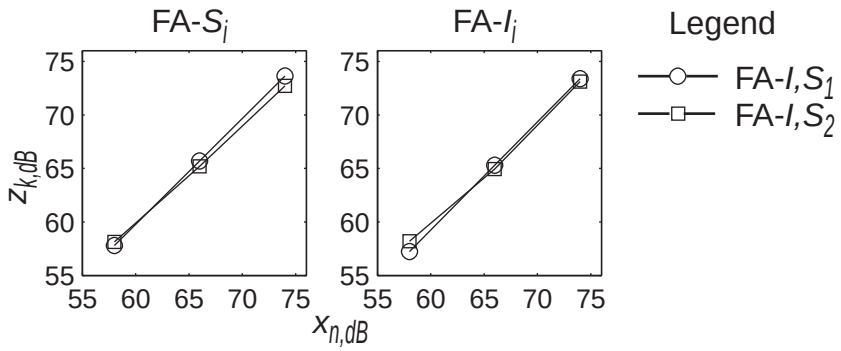


*Figure A1.* The averaged data from the FA procedure are presented. Plotted is $z_{k,dB}$ as a function of $x_{n,dB}$.

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

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

## Appendix A
## Average dB data from FA

Averaged data for 6 respondents are presented. In Figure A1 $z_{i,dB}(x_n)$ is plotted against $x_{n,dB}$ separately for the separate condition (FA-$S$) and the interleaving (FA-$I$) one.

As before, the TOE is any deviation from the diagonal. In this form, it is not clear that the TOE is an important factor in the procedure, hence to magnify its effect, we look at the TOE expressed as in (1), namely

$$\kappa_{k,dB} = z_{k,dB}(x_n) - x_{n,dB}, \tag{18}$$

and in Figure A2, the TOE, $\kappa_{k,dB}$, is plotted directly a function of $x_{n,dB}$.

Figure A2 shows the magnitude of the TOE reaches nearly 2 dB, which is generally speaking a significant statistical effect, meaning the TOE is far from insignificant factor in the task. The effect varies by adjustment condition, and appears more pronounced for the matching $2^{nd}$ condition.

## Appendix B
## Average dB data from 2IFC

In Figure B1 $z_{i,dB}(x_n)$ is plotted against $x_{n,dB}$ for the separated condition (2IFC-$S$) and for the interleaved one (2IFC-$I$).

In Figure B2 the TOE is any deviation from the diagonal, and at first it appears to be small and fairly constant in dB space especially for the 2IFC-$S$ condition. There is perhaps some deviation for the lower intensities in the 2IFC-$I$ condition. Overall, there does not

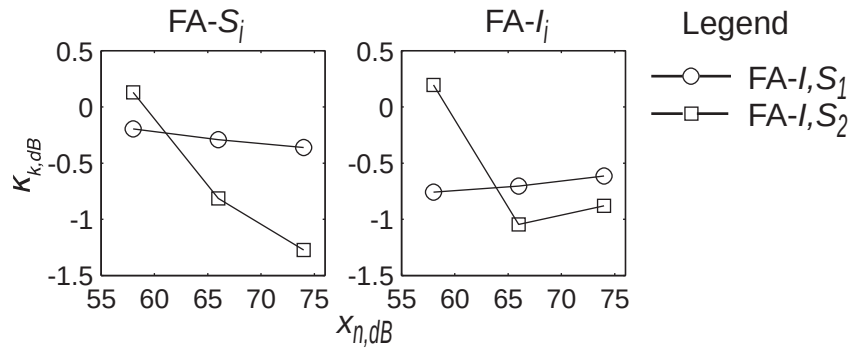


*Figure A2.* The averaged data from the FA procedure are presented. Plotted is $\kappa_{k,dB}$ as a function of $x_{n,dB}$.

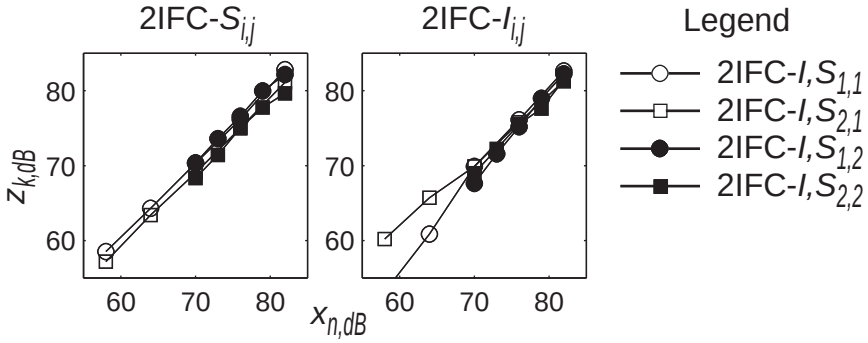


*Figure B1.* The averaged data from the 2IFC procedure are presented. Plotted is $Z_{k,dB}$ as a function of $x_{n,dB}$.

appear to be a large difference between the two conditions. Let us zoom in on the effect by plotting the TOE, $\kappa_{k,dB}$, Eq. (18), directly a function of $x_{n,dB}$.

Viewed this way, there are clear differences between the conditions. First, for 2IFC-$I_{1,1}$ the magnitude of the TOE reaches 6 dB. For 2IFC-$I$ the overall pattern between the broad and narrow range is similar. However, if the range manipulation has no or minimal effect, the TOE should be roughly equal for equal intensity. For 2IFC-$I_{2,1}$ and 2IFC-$I_{2,2}$ this is broadly true, but not for 2IFC-$I_{1,1}$ and 2IFC-$I_{1,2}$. Here, adjusting the $1^{st}$ signal shows far larger TOE than does adjusting the $2^{nd}$ one.

Turning to 2IFC-$S$, the results for all but the 2IFC-$S_{1,2}$ appear dramatically different from those of 2IFC-$I$. The broad range results show a quite small and fairly constant (in dB) TOE, where as the narrow range ones show similar non-constant TOE. Also notable is that the broad range curves, although different, are nearly parallel and the same holds for the narrow range curves.

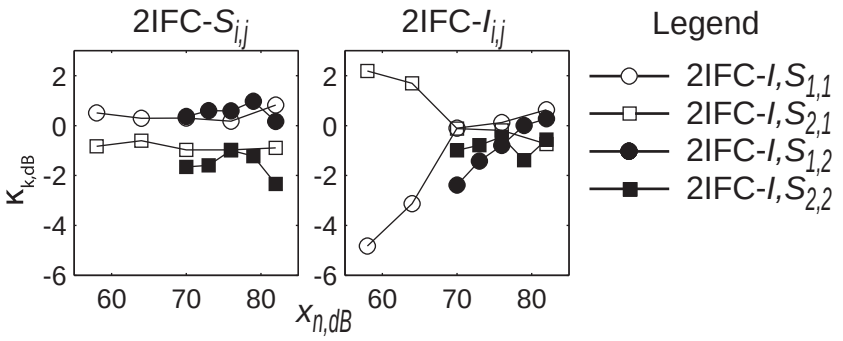


*Figure B2.* The averaged data from the 2IFC procedure are presented. Plotted is $\kappa_{k,dB}$ as a function of $x_{n,dB}$.

## Appendix C
## Existence of brightness TOE

We have not carried out a systematic study of the TOE in any other domain than loudness, however, we sought to verify that the phenomenon was not restricted to that domain. For this reason, we collected data for brightness.

The procedure used was identical that of Experiment 2 with the following differences. The stimulus was an achromatic square, subtending 20 degrees of visual angle, presented on an 18" NEC Multisync FE 950+ with a resolution of $1024 \times 768$ pixels and refresh rate of 75 Hz, and generated by an Apple G4. The Experiment was conducted in a dark and light-insulated room. The stimuli for the broad range were $x_1 = 12.0$, $x_2 = 20.0$, $x_3 = 30.7$, $x_4 = 44.4$, $x_5 = 61.6$ $cd/m^2$ and for the narrow range, they were $x_1 = 30.7$, $x_2 = 37.4$, $x_3 = 44.4$, $x_4 = 52.2$, $x_5 = 61.5$ $cd/m^2$ all presented on a background of 3.4 $cd/m^2$.

Data for two respondents are presented in Figure C1. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$ as detailed in the section "Qualitative properties predicted."

The data have the main similarity with the auditory ones that there is a great deal of individual variability, they are not well described by our model, and the TOE is smaller for the 2IFC-$S$ than for the 2IFC$-I$. However, the main goal of presenting these data is the simple conclusion that the TOE is not restricted to loudness.

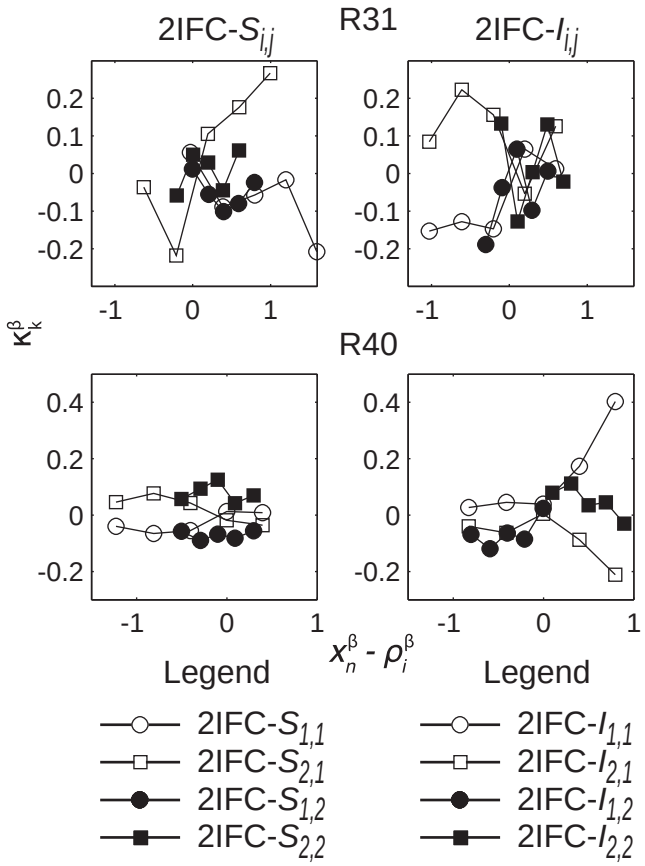


*Figure C1.* Data from the 2IFC-$S$ and 2IFC-$I$ procedures are presented by plots of individual data for two respondents. Plotted is $\kappa_k^\beta$ as a function of $x_n^\beta - \rho_i^\beta$. The figures legends are given below each column for which it applies.