Time order errors in visual length discrimination

by

David G. McGill, B.A.

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Arts.

Department of Psychology

Carleton University

Ottawa, Ontario

August, 2004

© copyright

David McGill, 2004

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-98984-4

Our file Notre référence

ISBN: 0-612-98984-4

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

In each of 2 experiments, participants compared successively presented visual extents at inter-stimulus intervals (ISIs) of 500, 3000, and 6000 ms. Lines in each of the 3 comparison pairs were short, mid-range or long. On half the trials, the lines were presented in the short-long order, and in long-short order on the other half. The TOE index is negative when discrimination is better in the short-long order than in the long-short order, and positive contrariwise. The observed TOE was found to be dependent upon pair magnitude and ISI. In a third experiment, lines were interpolated into varying locations within a fixed ISI, without effect. Singular attempts to explain presentation order effects in terms of sensory fatigue, peripheral or decisional biases, assimilation-contrast, or wave theory are incompatible with the data. The differential sensation weighting approach developed by Hellström (1985) provides a full account of the overall configuration of findings.

Acknowledgements

During the writing of this thesis, I have benefited from the expertise of many people. The comments and corrections made by Dr. Joseph Baranski, Dr. Craig Leth-Steensen and Dr. Robert West were immensely helpful.

Thanks are also due to Dr. Samuel Shaki, a fellow lab member, who showed me how to program experiments, and familiarized me with Carleton's experimentation protocol. My advisor, Dr. William Petrusic, provided support in myriad ways throughout my degree. He assisted with funding, helped arrange several presentations, and served as an invaluable reference for all of my thesis-related questions.

This research was supported by a Natural Sciences and Engineering Research Council of Canada Discovery grant to Dr. William Petrusic.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
List of tables.....	vi
List of figures.....	vii
List of appendices.....	viii
Introduction.....	9
Fechner's discovery of TOE.....	9
TOE indices.....	10
The relevance of TOE to comparative judgment.....	12
Ecological validity: chimerical fancy?.....	14
TOE as a pairing effect.....	15
Fading trace theories.....	19
Assimilation.....	21
Adaptation level.....	22
Contrast effects.....	24
The category effect.....	25
Simple response bias.....	27
Decisional biases.....	28
Wave theory.....	30
Thurstonian scaling.....	31
Differential sensation weighting.....	35

Derivation of a new DSW model	44
The current investigation	48
Experiment 1	49
Experiment 2	60
Experiment 3	66
General Discussion	70
References	73
Appendices	78
Footnotes	99
Tables	103
Figures	107

List of Tables

Table	Title	Page
1	Coefficients for the empirically fit DSW model in experiment 1	103
2	Distributional characteristics of s_1/s_2 values at each ISI in experiment 1	104
3	Coefficients for the empirically fit DSW model in experient 2	105
4	Distributional characteristics of s_1/s_2 values at each ISI in experiment 2	106

List of Figures

Figure	Title	Page
1	A three-way interaction between order, instruction and inter-stimulus interval for experiment 1	107
2	A three-way interaction between pair, order and inter-stimulus interval for experiment 1	108
3	The sample distribution of relational levels for experiment 1	109
4	A three-way interaction between pair, order and inter-stimulus interval	110
5	The sample distribution of relational levels for experiment 2	111
6	The pair by interpolated length interaction for experiment 3	112
7	The non-significant pair by interpolated length by time-to-interpolation interaction for experiment 3	113
8	The instruction by interpolated length interaction in terms of T_e for experiment 3	114

List of Appendices

Appendix	Description	Page
A	TOE in the context of classical psychophysics	78
B	Proofs	84
C	A rationale for $T_e = \overline{D} / R_D$	86
D	Data sets	89
E	Syntax and equations used for computer analyses	91
F	Using R_{press}^2 for model validation	93
G	Participant plots for the pair by ISI interaction for T_e	94

Introduction

Fechner's discovery of TOE

Psychologists have studied the time-order error (TOE) phenomenon since Gustav Theodor Fechner established psychophysics in the mid-nineteenth century. Fechner (1966) served as his own participant in a series of experiments involving the comparison of lifted weights. Using the method of constant stimuli, he observed that the second-lifted comparison weights in his experiments tended to be either over- or underestimated relative to the first-lifted standard¹ weights. Thus, discriminative judgment was affected by the order in which two weights were compared.

Overall mass of the weight pairs influenced the discriminative process. Increasing magnitude of the to-be-compared weights produced more “second heavier” judgments. Fechner suggested that fatigue was responsible for this result. He theorized that the muscular enervation brought about by lifting the standard weight, increased absolute sensitivity to the comparison weight.

Although, Fechner's theory primarily explains negative TOE, he observed both positive and negative TOEs. Negative TOE is defined as the systematic underestimation of the first stimulus relative to the second in a comparative judgment task. Positive TOE consists of the systematic overestimation of the first stimulus relative to the second. TOE is a type of asymmetric error since successful discrimination between two stimuli depends, in part, upon the order in which they are presented. If discriminative accuracy is equal for both presentation orders, then none of the error in judgment can be said to be asymmetric. TOE is also a type of presentation order effect.

Since Fechner's day, TOE has been further documented in comparative judgment experiments involving weights (Woodrow, 1933; Hellström, 1985), sound intensities (Ellis, 1973a; Inomata, 1964; Needham, 1934a, 1935a), temporal durations (Hellström, 1977; Jamieson & Petrusic 1975a, 1975b; Woodrow, 1935), line lengths (Baranski & Petrusic, 1992; Leth-Steensen, 1994; Inomata, 1959; Petrusic, Harrison & Baranski, 2004), area for geometric shapes (Tresselt, 1948; Inomata, 1959), electric shock intensity (Geertsma, 1958), dot numerosity (Rambo & Johnson, 1964), and the aesthetic appeal of color patterns (Hellström, 2001). TOE is most reliably found for those judgment classes belonging to Steven's (1957) prothetic continua (e.g., sound intensity and brightness). Nonetheless, it has occasionally been found for metathetic dimensions, such as pitch (Tresselt, 1948).

TOE indices

Several different TOE indices exist. Using the method of constant stimuli, the TOE is computed as the difference between the physical value of the standard stimulus and the point of subjective equality (PSE)². This difference is expressed in terms of the physical units of the standard. Traditionally, this index is called "constant error" (CE) rather than TOE; $CE = PSE - \text{Standard}$. This name is certainly inapt. As noted by Luce, Bush and Galanter (1963), asymmetric error regularly varies over experimental conditions. It is not constant.

Investigators have frequently used a percentage difference (D%) index of TOE. D% is a percentage based upon the relative frequency of "second less" and "second greater" judgments. Using D% rather than CE to represent TOE is advantageous insofar

as it allows one to compare the TOE found in either of the prevailing psychometric methods: the method of constant stimuli and the method of paired comparisons. The latter method lacks an accepted procedure for calculating a PSE.

$$D\% = 100 \frac{(L - G)}{(L + G)} \quad (1)$$

In Equation 1, L represents the total number of incidences in which the second stimulus is judged less than the first. Meanwhile, G is the total number of incidences in which the second stimulus is judged greater than the first. Values between -100% and +100% can be obtained with this index.

Jamieson and Petrusic (1975a) devised a measure of TOE based upon error differences in discriminative accuracy between presentation orders. This index may be interpreted as the average decrement in discriminative accuracy that is attributable to response asymmetry. Thus far, it has been applied only to either the complete or the incomplete method of paired comparisons. Nonetheless, it is also applicable to the method of constant stimuli.

$$T_e = \frac{1}{2} [p(\text{correct|GL}) - p(\text{correct|LG})] \quad (2)$$

The probability of a correct response in the lesser-greater stimulus configuration is subtracted from the probability of a correct response in the greater-lesser stimulus configuration. The resultant difference is next halved. Values between -.50 and +.50 can be obtained with this index. The measure was designed so as to be congruent with Fechner's notion of TOE. Fechner held that pair discriminability is fundamentally the same between presentation orders. In one presentation order, TOE is added to the base

discriminability. Meanwhile, TOE is subtracted from the base discriminability in the other presentation order.

For each of these three indices, negative TOE produces negative numbers and positive TOE produces positive numbers. CE reflects shifts in reference level. $D\%$ reflects response asymmetry, and T_e is a performance measure that indicates the extent to which discriminative accuracy varies between presentation orders.

The relevance of TOE to comparative judgment

TOE effects have been studied within comparative judgment experiments based upon the method of constant stimuli since Fechner's time. The method of constant stimuli is a classical method of psychophysics, which has endured for over a century. The longstanding popularity of this experimental method is partly attributable to its association with the just noticeable difference (JND). The JND's validity, as an index of discriminability, is compromised by the presence of TOE effects. As such, psychophysicists who make use of the method of constant stimuli should be concerned with presentation order effects. A more elaborate handling of this topic is provided in Appendix A.

TOE effects can profoundly influence the results of any comparative judgment experiment – not only those based upon the method of constant stimuli. Jamieson and Petrusic (1975b) found that the percentage of correct responses in a duration discrimination experiment changed by as much as 80% with a shift in presentation order. Ellis (1973a) found TOE of equal magnitude for comparisons of sound intensity. The second of two equal stimuli was judged louder in 83% of the experimental trials. If these

magnitudes of effect were to be reliably found in all stimulus dimensions (durations, sound intensities, etc.), then it is probable that presentation order effects would be a focal area of psychophysical research. This is not the case. The last review article dealing with TOE theories was published almost 20 years ago (Hellström, 1985).

TOE has defied explanation because the magnitude and direction of the comparison error are influenced by a multitude of factors: stimulus dimension, stimulus magnitude, stimulus range³, inter-stimulus interval (ISI)⁴ and individual differences. A glut of deficient theories is testament to the enigmatic nature of TOE.

Although the vast majority of TOE theories can be regarded as insufficient, this point by no means insinuates that they lack value. An adequate theory of TOE must dauntingly cut across sensory domains to outline the fundamental processes underlying comparative judgment. Consequently, TOE is not an insular, unsolved mystery of psychophysics. Waylaid theories of presentation order effects have inexorably contributed to the broader topic of comparative judgment.

Furthermore, the existence of two, successful, quantitative models of presentation order effects – Helson's adaptation-level model (Helson, Michels & Sturgeon, 1954; Helson, 1964) and Hellström's differential sensation weighting model (Hellström, 1977, 1978, 1979, 1985, 2000) – suggests that TOE may be discussed in the mathematical medium. Thurstonian scaling is of direct relevance to the study of TOE if the phenomenon is, indeed, perceptual. Those areas of psychological inquiry, which can be discussed in precise, quantitative language, have patent merit.

Ecological validity: chimerical fancy?

Showing that TOE is relevant to psychophysicists working in laboratory settings in no way obliges ecological validity. The vast majority of TOE research does not address how the error might permeate everyday judgments. There are several plausible reasons for this avoidance:

1. TOE is poorly understood, even in experimental settings where extraneous variables are strictly controlled.
2. The temporal interval separating to-be-compared items is crucial to the magnitude and direction of the TOE. Gestalt co-founder Wolfgang Köhler observed that the three-second ISI tends to be a locus for changes between positive and negative TOE (Needham, 1934b). Lengthening the ISI to mimic judgments that are not immediate might fail to capture the dynamic, temporal aspect of TOE.
3. Research into TOE has focused primarily upon quantitative, perceptual comparisons. Everyday comparisons are frequently of the qualitative sort. Thurstone's (1927) attempt to bridge the gap between quantitative and qualitative judgments has not been fully realized.
4. Primacy and recency effects are well known⁵. One can account for asymmetrical response patterns in comparative judgment with reference to primacy and recency (e.g., Schab & Crowder, 1988).

Hellström (2001) found and empirically fit TOE in an experiment that involved comparisons of aesthetic appeal. His theory of TOE effects is perceptually rooted. Thus, perceptual theories of TOE might provide insight into a broad field of inquiry. It is an

intriguing possibility that perceptual explanations of TOE may be harmonized with primacy/recency explanations. Such reconciliation might constitute a link between perceptual and semantic memory.

TOE as a pairing effect

Stimuli may be over- or underestimated. This obvious truism does not, in itself, provide a sufficient basis for TOE. TOE is present when stimuli are compared and the result of this comparison is the over- or underestimation of one stimulus relative to another. There are two requisite conditions for the occurrence of asymmetric comparison error: a) two stimuli are compared and b) stimuli are presented separately in time⁶. Absolute judgment experiments do not technically evaluate TOE, because comparison is not involved. Nonetheless, these experiments may provide valuable insight into the factors responsible for asymmetric comparison error.

The mapping of knowledge from absolute to comparative settings will be uncomplicated if two assumptions are valid: a) absolute and comparative judgments are mediated by the same set of global processes, and b) the set of comparative judgment processes encompasses, or is superordinate to the set of absolute judgment processes. Neither of these assumptions is self-evident, but both are frequently made⁷. One should be careful when comparing the results of absolute judgment, and comparative judgment experiments.

Needham (1935b) evaluated asymmetric responses for a series of absolute judgments. In his experiment, participants judged four sound intensities differing by a step value of 1 db. The stimulus intensities were focused around a sound intensity of 40

db⁸. Stimuli were randomly presented at 4, 6.5, or 14-second intervals and participants provided ratings of “weak” or “loud”. The D% index was applied to the results by corresponding responses of weak and loud with “less” and “greater”. Loud judgments were most prevalent at all ISI’s. This propensity to judge stimuli as loud was found to increase with lengthening of the ISI and decrease as a function of practice trials.

One might explain Needham’s results in terms of participants having a simple bias towards louder judgments. However, there is no intuitive reason why bias should shift with either ISI or practice when feedback is absent. One might alternatively attribute the TOE to participants’ flawed percepts. Such a perceptual account of TOE must incorporate perceptual memory processes, because TOE changed with ISI. One need then determine if the perceptual trace of one stimulus is somehow dependent upon the perceptual traces of other series stimuli.

If previously experienced stimuli affect the perceptual trace of a judged stimulus, then differential carryover effects are present. The presence of these effects would mean that “absolute” judgments are not absolute in the traditional sense. Consequently, we are left with two difficult questions. Is absolute judgment more basic than comparative judgment? What processes are being examined when one contrasts the results obtained from absolute and comparative judgment experiments?

Fernberger (1931) performed an experiment in which participants made both comparative and absolute judgments for the same series of weights. The ISI was fixed for both types of trials. Firstly, absolute judgments of the form “light”, “intermediate”, or “heavy” were made on a series of seven weights. Secondly, the method of constant

stimuli, utilizing one standard and (the same) seven variable weights, was employed.

Responses of “lighter”, “heavier” or “equal” were given in the comparative component.

The PSE for the comparative experiment was less than both the physical standard and the PSE for the absolute judgment experiment. Two conclusions may be drawn from these results: a) PSEs may underlie the arithmetic mean of a series and b) the comparison process acts upon the PSE (comparison depressed the PSE in this experiment).

Jamieson and Petrusic (1975c) verified and elaborated upon this result in an experiment involving the reproduction of paired or unpaired temporal durations. For any given experimental trial, participants were unaware as to whether one (unpaired) or two, identical (paired) temporal durations would be presented. Only after the presentation of the first duration did this information become available. If the trial consisted of a singly presented duration, then participants immediately reproduced this duration. If the trial consisted of two durations, separated by a variable ISI, then participants reproduced either the first or the second duration. They were not informed as to which (first or second) duration they would be reproducing until both stimuli had been presented. This dual presentation scenario sought to mimic the pairing effects of a comparative judgment task.

To compare the results under discussion to those of Fernberger (1931), we must make the assumption that reproduction values represent direct estimates of PSEs. Since Fernberger had a fixed ISI, we shall also collapse across the three different ISI's used by Jamieson and Petrusic in their first experiment. If Fernberger's weight -lifting results generalize to duration discrimination, then we should expect the reproduced durations to

underestimate actual durations (this result is analogous to negative TOE). We should also expect reproduced values for both comparison stimuli to be less than the reproduced values for single stimuli in the absolute condition. Each of these results was observed.

Further conclusions can be drawn from Jamieson and Petrusic's (1975c) experiment when one considers the full spectrum of factors for the dual presentation situation. The reproduction value of the first stimulus tended to increase with ISI. The reproduction value of the second stimulus tended to decrease with ISI. These findings imply that subjective magnitude for first and second pair members are differentially coupled to ISI. In both comparison and absolute judgment conditions, longer time spent in memory translated into greater perceived magnitude. McClelland (1943) obtained a similar result for an experiment involving absolute judgments made upon line lengths.

These results suggest that the perception of two temporally coupled stimuli is not, by itself, synonymous with comparative judgment. Based upon the results, it can logically be concluded that subjective magnitude increases with time of memory retention. It can also be concluded that the subjective magnitude of second stimuli diminishes as ISI lengthens. If these two postulates are true, then TOE should become positive as ISI lengthens. However, this is not the case. When temporal durations are compared, TOE becomes increasingly negative as ISI lengthens (Ellis, 1973a; Jamieson & Petrusic, 1975a). When line lengths are compared, TOE directionality is more complicated. The TOE for lesser line pairs has been found to increase as ISI lengthens, while the TOE for greater line pairs has been found to decrease (Inomata, 1964).

Fading trace theories

Köhler (as cited in Needham, 1934b) theorized that each physical stimulus is internally represented via a memory trace. He further supposed that this trace increases immediately after stimulus presentation, quickly reaches asymptote, and then begins to fade (as cited in Peak, 1940). Köhler's trace strength is synonymous with subjective magnitude. Accordingly, the first comparison stimulus will tend to be overestimated if a judgment is made while its trace is near the initial asymptote. This stimulus will be underestimated if a judgment is made when the trace is in a state of decline.

Köhler's theory was physiological insofar as he hypothesized the existence of a second-order isomorphism between stimuli and internal, electro-chemical states. Every stimulus was posited to map onto a dynamic concentration of ions located in a distinct region of the brain. These ion concentrations were posited to wax and then wane during the course of a stimulus presentation.

This fading trace theory readily explains findings of positive TOE when an ISI is brief. It also explains findings that TOE becomes increasingly negative as an ISI lengthens. Köhler predicted that TOE should be monotonically related to ISI. Increasing the interval between to-be-compared stimuli should push TOE in the negative direction. However, the theory makes no dispensation for a relationship between stimulus magnitude and TOE. Köhler's fading trace theory of TOE can be refuted on two grounds. Positive TOE frequently becomes increasingly positive as ISI lengthens (e.g., Geertsma, 1958; Inomata, 1959), and stimulus magnitude regularly affects the directionality and magnitude of TOEs.

Lauenstein developed an alternative trace theory (as cited in Peak, 1940). He defined the trace to be a field of neural activity, and suggested that first and second stimuli traces move towards each other (Tresselt & Mayzner, 1967). The investigator supposed that stimuli assimilate towards an interpolated interval. Lauenstein asserted that traces for both stimuli decrease when an ISI is unfilled. He also contended that first stimuli are affected by an unfilled ISI to a greater extent than second stimuli; first stimuli are retained in memory for a greater amount of time. If a third stimulus were to be inserted into an ISI, then traces for both stimuli would move towards the interpolated value. Needham (1934b) criticized Lauenstein's trace theory on the grounds that an interpolated event would erase the trace of the first stimulus and spoil the successive comparison. This criticism is tenuously based upon the premise that fields of neural activity cannot coexist.

Pratt (1933) also found a flaw in Lauenstein's account of TOE. Pratt used the method of constant stimuli to study the comparison of sound intensities. Two stimuli, which were quieter or louder than the standard, were alternately interpolated into an ISI. In a third condition, an ISI was left blank.

Lauenstein's theory predicts that positive TOE will result when an interpolated sound is louder than either of the comparison stimuli. His theory implies that negative TOE will result when an interpolated stimulus is softer than both comparison stimuli. According to Lauenstein's trace theory, the greatest, negative TOE should have characterized that condition in Pratt's experiment, which featured an unfilled ISI.

Lauenstein rationalized that a first stimulus' trace will assimilate towards a value of zero if an ISI is empty.

Pratt (1933) confirmed the first two predictions, but obtained results discrepant with the last hypothesis. The TOE was less negative for Pratt's control condition than for the condition featuring a low-intensity interpolated stimulus. It did not appear that first stimuli were assimilating towards a value of zero. Pratt concluded that traces assimilate towards interpolated stimuli, but added a proviso to this conclusion: an empty ISI cannot be conceived as an interpolated stimulus that has a value of zero. Pratt did not address the issue that plagued Köhler (positive, increasing TOE).

Assimilation

Pratt was not the only researcher to find interpolation effects and explain them in terms of an assimilative mechanism. A standard finding, which cuts across sensory domains, is that intense interpolations produce positive TOE, while weak interpolations produce negative TOE. This result holds true for comparisons of lifted weights (Guilford & Park, 1931), and sound intensities (Needham, 1935a). In this context, assimilation may be said to take place between series stimuli.

Assimilation might also take place between stimuli and background. Researchers have generally assumed that assimilation is facilitated by increased similarity between background and stimuli. McClelland (1943) performed a comparative judgment experiment involving line lengths and found that assimilation is limited to certain circumstances. Background illumination did not produce an assimilatory effect. However, presenting the lines within a square frame caused TOE to become less

negative. Tressalt (1948) studied comparisons of both line length and square area. He concluded that similarity fosters assimilation while heterogeneity lends itself to contrast.

In addition to assimilation between stimuli, and assimilation between stimuli and backgrounds, there may be assimilation between stimuli and some psychological reference value. Woodrow (1935) postulated that stimulus traces assimilate towards two remote standards: a) an absolute indifference interval⁹ and b) a weighted series average.

Given the prevalence of negative TOE, Woodrow (1935) concluded that the absolute indifference interval tends to underlie the physical standard. The commonly observed reduction in TOE effects accompanying practice was attributed to the shifting of influence between the two remote standards. Woodrow suggested that the weighted series average gains potency as experimental trials progress. Several other researchers have also found that the absolute magnitude of TOE tends to decrease as a function of practice trials (Needham, 1934a; Rambo & Johnson, 1964).

Adaptation level

Helson (1964) quantitatively described how stimuli might assimilate to a remote anchor. This remote anchor, or adaptation level (AL), was posited to be a weighted geometric mean of a stimulus series. Any one stimulus' contribution to the AL is determined by its' recency of occurrence. Helson's AL is synonymous with the term PSE. The crux is that the AL is determined by an entire set of experimental stimuli.

The quantitative AL model is applicable to results obtained from comparative rating experiments. Michels and Helson (1954) used the method of constant stimuli to study comparisons of lifted weights. The investigators translated the verbal response

categories of “very much heavier; much heavier; heavier; a little heavier; equal; a little lighter; lighter; much lighter; very much lighter” into a 9 point numerical scale. The formula for predicting judgment category in the standard-variable order was

$$J = C + K \log (X/A) - K s \log (S/A) \quad (3)$$

In this formula, C represents the numerical value associated with “equal”, K is a constant and X is the comparison stimulus (in physical units). The value A represents adaptation level (for the entire series proceeding comparison), and S represents the standard stimulus (also in physical units). The value s is a weight that determines the extent to which a standard stimulus contributes to the formation of a distinct, comparatal AL. It should be noted that an s value of one is commensurate with a standard stimulus being accurately perceived. The authors did not calculate the value of A using Equation 3. Instead, they regressed category rating upon stimulus magnitude, for both presentation orders. The stimulus magnitude located at the interception of regression lines was taken to be A^{10} .

The adaptation level model provided an excellent fit to the data of the experiment under discussion. Michels and Helson (1954) encountered similar success when performing another experiment involving the comparison of lifted weights. Helson and Mawr (1947) also found that the model could account for judgments made when an anchor was interpolated into the ISI. Rambo and Johnson (1964) applied the model to comparisons of dot numerosity and found that the comparatal AL becomes more stable across trials.

The AL model has been criticized for exclusively focusing upon assimilatory mechanisms. Hellström (2000) assessed the AL model as being limited for a different reason. It is only the first member of a comparison pair that is affected by the AL. Hellström's own model of TOE accounts for distorted perception of both the first and second stimuli. Additionally, the AL model does not comprehensively explain the effects of ISI upon TOE. The model is also premised upon Fechner's law. Undoubtedly, all of these factors contribute to the contemporary inattention to the AL model.

Contrast effects

Helson's AL theory does not account for contrast. Peak (1940) categorized contrast as being the exaggeration of physical differences between stimuli. Stimuli interpolated into an ISI tend to have profound influence upon TOE. The effects of interpolation have been alternately attributed to assimilation and contrast. Needham (1935b) performed a comparative judgment experiment with sound intensities and described his findings in terms of contrast between the second stimulus and the interpolated value. This could alternatively be likened to assimilation between the first stimulus and interpolated value.

Ellis (1973b) employed anchoring stimuli that were either interpolated into the comparison, or presented before the first stimuli (which were sound intensities). TOE was affected by the order in which the anchor was presented. This result could not readily be explained in terms of assimilation. Thus, the author concluded that interpolation effects are due to proactive contrast. Yet, the author's findings – self-admittedly – were not the norm. Gleitman (1957) performed a similar experiment with

sound intensities and found that anchor position did not influence TOE. There is no overwhelming evidence to suggest that proactive contrast is responsible for comparison error.

The category effect

Stevens' (1957) dismissed TOE as being nothing more than a by-product of an experimental method. He did not consider TOE to be a genuine perceptual phenomenon that involved the distortion of subjective scale values. Instead, Stevens argued that TOE is produced by certain characteristic response tendencies, which are associated with the issuance of category ratings. The method of constant stimuli requires that subjects respond with category ratings. The author attributed TOE to a "category effect."

Stevens claimed that subjects change response categories more rapidly when discriminative ability improves. He further asserted that discriminative ability is not equal across different stimulus magnitudes. For example, a subject might be able to compare small stimuli to a standard stimulus more accurately than she can compare large stimuli to the same standard. If this were to be the case, then the average rating given to the standard stimulus would not be equal to the median response category (e.g., "3" of a 5 point Likert scale). Consequently, the calculated PSE would not equal the physical value of the standard stimulus. Hence, TOE would be found. However, this TOE would not be due to the under- or overestimation of either comparison stimulus.

Stevens and Galanter (1957) implicitly predicted minimal TOE for comparisons of line length. They found that category and magnitude scales for visual extent were linearly related. In other words, discriminative ability did not vary over stimulus range

for the dimension of line length. This unusual finding was attributed to subjects' proficiency in judging length. Inomata (1964) also explained his experimental results in terms of the category effect. Notwithstanding, few researchers have adopted this approach.

Stevens did not address how changing ISI might affect TOE. Presumably, the relationship between category and magnitude scales must be sensitive to the time course of judgment if TOE is to vary over ISI. This supposition has not been evaluated.

Stevens' interpretation of TOE effects is questionable insofar as the author presumed that magnitude scales are a stable benchmark against which category scales may be evaluated. This is not necessarily the case. Markley, Ayers and Rule (1969) found the power law exponent to be 1.1 for similarity judgments of line length. Other experimenters have found the power law exponent for line length to be approximately 0.5 (Parker, Schneider & Kanow, 1975; Petrusic, Baranski & Kennedy, 1998).

Furthermore, there may be considerable variation in the power law exponents for individual participants within the same experiment. Petrusic, et al. (1998) found the range of participants' power law exponents for line length comparisons to be 1.01. Hence, it is not clear that TOE can be defined as a difference between biased category scales and unbiased magnitude scales. Magnitude scales are a tenuous benchmark.

Even if magnitude scales are a stable benchmark against which category biases may be evaluated, the category effect remains a demonstrably insufficient account of TOE. Massive TOE effects often appear in experiments based upon the method of paired

comparisons (Jamieson & Petrusic, 1975b). One does not expect a category effect to be possible when responses are binary (e.g., lesser or greater).

Simple response bias

A simple response bias exists when a participant prefers certain modes of response. In the comparative judgment paradigm, there are typically two possible responses (lesser and greater). If TOE were the product of a participant's preference for one of two possible responses, then changing experimental instructions would cause a reversal in TOE directionality. For example, suppose that a participant is asked to compare two lifted weights. This hypothetical experiment shall require that the second weight be judged relative to the first, and that responses be either 'lighter' or 'heavier'. Negative TOE will result if the participant has a proclivity for the heavier response. If instructions were changed such that the first weight was judged relative to the second, then the response tendency would produce positive TOE. Changing instructions in this manner has not been found to reverse TOE. On this basis, simple bias can be dismissed as a source of TOE.

John (1975) postulated what has become known as an implicit response bias. The investigator asserted that there is a "tendency for S [subjects] to respond 'louder' to loud comparison stimuli and 'softer' to soft comparison stimuli (54, op. cit.)". It was suggested that this bias might increase with ISI. The author attributed regular findings of negative TOE to individuals' greater confidence in the louder response. This theory can be evaluated in the same manner as the simple response bias. As such, it may also be dismissed as a source of TOE.

Decisional biases

Luce et al. (1963) suggested that TOE is the product of decisional bias:

‘When (x,y) is presented, the first stimulus, x, is judged (larger/smaller) than the second, y, if $X \not=> Y + c$, where c is some positive or negative number’ (p. 225).

The authors’ modification to Thurstone’s (1927) decisional rule involved the addition of a bias parameter, c, to the decision process. By instituting this bias parameter, the authors specified a mechanism for decisional bias. However, they did not actually perform comparative judgment experiments, or predict what factors might influence bias.

Allan (1977) pointed out that when TOE is found, it is often highly variable amongst participants. Leth-Steensen (1994) found extreme variability amongst participants’ TOEs for comparisons of line length. Allan insinuated that such findings might be taken as evidence that TOE is rooted in the decisional processes. The implicit assumption being that perceptual processes do not vary greatly amongst individuals.

Allan made the Thurstonian assumption that any given stimulus maps onto a subjective magnitude distribution. For example, a temporal duration may be coded as either ‘short’ or ‘long’. If the momentary, subjective magnitude of the stimulus underlies a preset criterion, then the stimulus is coded as short. If the momentary, subjective magnitude of the stimulus exceeds said criterion, then the stimulus is coded as long. Thus, every stimulus is coded as either short or long. The decision process is unbiased when the coding criterion coincides with the mean of the subjective magnitude distribution.

Given Allan's conception of TOE, it can arise if and only if two conditions are met:

1. The criterion must be biased. In essence, it cannot be equivalent to the mean of the subjective magnitude distribution.
2. The to-be-compared stimuli must receive the same coding: short-short, or long-long.

When the latter condition occurs, Allan suggested that participants use only one of the two available stimuli to make their decision (the second). If the criterion for any one stimulus is biased, then a response based solely upon said stimulus will also be biased. The author further supposed that criteria change according to the intensive level of stimuli: "Short durations are likely to be coded as short, and long durations as long" (Allan, 1977, p. 28).

According to Allan, comparative judgment is based upon the comparison of "short" and "long" codes. She does not describe these codes as "semantic". Nonetheless, this biased criterion model of TOE is clearly premised upon the semantic coding of stimuli. This model can be disproved by showing that stimuli are not semantically encoded.

Hellström (1985) criticized this model on two grounds: a) Any quantitative model involving the theory would require an inordinate number of parameters, and b) It is not intuitive that bias should be related to either stimulus magnitude, or ISI. Both of these factors are implicated in TOE effects. Although, the latter point may be moot when it comes to comparisons of temporal duration. Allan, Kristofferson, and Rice (1974) found

that increases in ISI are not necessarily associated with decrements in discriminative performance for comparisons of temporal durations. The perception of time appears to be unique in some ways.

Wave theory

Link's (1992) wave theory of comparison also attributes TOE effects to bias. Link proposed that each stimulus is internally represented by a continuous, electromagnetic wave. When two stimuli are compared, an accumulator continually sums the differences between wave amplitudes. The contents of this accumulator are compared to upper (greater) and lower (lesser) response thresholds. A response is made when the summed difference between wave amplitudes exceeds either response threshold. A bias exists when one response is favored at the expense of the alternative.

Link models the proportion of "second greater" responses in the method of paired comparisons as follows¹¹

$$P_{LG} = \frac{e^{(\Delta s/S_B)T} - e^{(\Delta s/S_B) \ln[M_B(\Delta s/S_B)]/(\Delta s/S_B)}}{e^{(\Delta s/S_B)T} - e^{-(\Delta s/S_B)T}} \quad (4)$$

P_{LG} represents the proportion of "second longer" judgments associated with a comparison pair. The base of the natural logarithm is represented by e (≈ 2.718). The proportional change between first and second stimuli is reflected by $\Delta s/S_B$. T represents the response threshold and M_B is the "moment-generating function" used to determine the bias associated with a particular Weber constant¹².

This particular wave theory model is brought to the readers' attention, because it is applicable to experiments discussed herein. Those values in the equation, which are

highlighted, should be treated as parameters when the model is evaluated according to least squares criteria. The response threshold, T , is not expected to change over ISI (Link, 1992, p. 198). However, Weber's constant and the moment-generating function for bias are expected to change.

Thurstonian scaling

Thus far, Thurstonian scaling has only received casual reference. Yet, any theorist who attributes TOE to perceptual processes is beholden to deal with issues originally brought forth by Thurstone (1927).

Unobservable, subjective differences cannot be directly measured, nor is it apparent that they should exhibit obvious numerical relations. Yet, Thurstone managed to scale these differences. He did so by focusing upon an individual's repeated judgments of the same stimulus pair. Although, the difference between two compared stimuli may be kept constant with repeated presentation, an individual's impression of this difference will vary. The existence of variability in an individual's subjective impressions is a fundamental premise of Thurstonian scaling. It is presumed that errors in judgment are normally distributed, and that there are fluctuations in the judgment processes associated with stimulus identification. A normal distribution is presumed to describe the different states of these judgment processes.

Multiple assumptions were required to derive a scale for subjective differences. These assumptions have become indispensable bulwarks in psychophysics. The assumptions are typically grouped in the manner set out by Thurstone. The author

provided five cases of assumptions, ranked according to their soundness. The following assumptions made by Thurstone are listed in order of supposed validity.

1. The subjective experience of a stimulus is not the same at different times.
2. The subjective experience of the same stimulus is not the same for different individuals (typically combined with assumption I).
3. Different subjective experiences associated with one stimulus are normally distributed.
4. The mean value of the subjective experience distribution is most highly correlated with the physical stimulus. Thus, there is a measure of central tendency that best represents a stimulus. This mean value is called the scale value (Thurstone, 1954).
5. Discriminative processes are independent of each other.
6. Distributions of subjective experience, associated with different stimuli, have the same variance. This is Thurstone's famous Case V assumption and it builds on all previous assumptions.

The renown of these assumptions does not imply that they are universally held to be true. Favreau and Everett (1996) argued that experimenters seldom find true, normal distributions. On this basis, they deemed the normality assumption to be unacceptable. If the normality assumption is valid, then one expects experiments with large sample sizes to produce data that are approximately normally distributed. According to the authors, this rarely occurs.

Hellström (1993) suggested that the distribution of subjective differences is better described by a t -distribution than by a normal distribution. In the context of the author's differential sensation weighting (DSW) model, scaling an index of subjective difference by the t -distribution resulted in better prediction for participants than that afforded by scaling these values by the normal distribution¹³. On this basis, the t -distribution with 4 degrees of freedom was found to best fit the data.

The investigator theorized that distributional deviations from normality might arise from participants' fluctuating attentions. Hellström alleged that there are distributions of subjective difference associated with each, discrete attentional level. He suggested that attentional level, itself, is a random variable. If this is true, then a distribution of subjective differences is obtained by sampling from multiple parent distributions. Each parent distribution reflects a different state of attention. The author reasoned that low states of attention might engender greater variability amongst the parent distributions of subjective difference.

Sampling from a limited number of normal distributions that have different variances is sufficient for the production of a platykurtic distribution¹⁴. Sampling from normal distributions that have different means is sufficient for the production of a skewed distribution. Hence, investigators dealing with subjective magnitude distributions would be well advised to broach this topic.

The phi-log gamma hypothesis states that points on the standardized psychometric plot are normally distributed (fit the normal ogive) when the abscissa (featuring standard normal deviates for different stimulus magnitudes) is log-transformed. This implies that

the logarithm of subjective magnitude is normally distributed (Guilford, 1954).

Unfortunately, it is difficult to discriminate between this hypothesis and the phi gamma hypothesis, which simply states that subjective magnitude, is normally distributed.

Differentiation is made difficult by the fact that distributions of both subjective magnitude and log-transformed subjective magnitude are approximately normal.

The assumption of normality is highly suspect. Nonetheless, Thurstone's least tenable assumption is that which claims equality of variances amongst subjective magnitude distributions associated with physically different stimuli. Thurstone assumed that the variability of a subjective representation is unrelated to the physical size of a stimulus. This assumption is tantamount to a null hypothesis.

Based upon scaling experiments for which Steven's power law held, Ekman (1959) suggested that subjective JNDs are proportional to subjective magnitude (i.e. Weber's law holds for subjective magnitude). The most obvious corollary of this finding being that measures of discriminative dispersion are, indeed, related to stimulus magnitude. As such, more intense stimuli should have greater subjective variability.

A statement of caution should conclude all discussions of scaling. It is sometimes presumed that only one psychophysical function relates physical stimulus values to subjective stimulus values. Currently, Steven's power law function appears best qualified to meet this "Holy Grail" status. However, it seems improbable that the brain invariably maps physical events onto mental events according to one rule. It is more likely that multiple psychophysical functions are concurrently valid. Statements akin to

this one have been made by Thurstone (1954) and Stevens (1957). Linear, logarithmic and power law functions have variable success in variable circumstances.

Differential sensation weighting

Whereas Thurstone focused upon using response proportions to predict scale values, Hellström (1977, 1979, 1985, 2000) has used scale values to predict response proportions. His DSW model describes the systematic distortion of (subjective) scale values accompanying comparative judgment. The quantitative model is applicable to those comparative judgment experiments that use two or three response categories¹⁵. There are two components to this model. Firstly, there is a reference level (RL) that affects judgment. Secondly, the relationship between stimuli and the RL is dependent upon order of stimulus presentation.

The assumptions made by Hellström are considerably fewer than those associated with Thurstone's Case V. Firstly, the investigator has assumed that physical stimuli are projected onto subjective continua, by way of the same psychophysical function. This assumption is standard. Secondly, he has assumed that the subjective impression of a physical stimulus is not the same with repeated presentations. Thirdly, Hellström has assumed that subjective magnitude distributions are symmetric. As previously mentioned, the investigator has not presumed that these distributions are normally distributed.

Nor does this investigator assume that the dispersional measures associated with physically different stimuli are equal. Although, Hellström's disavowal of this assumption is somewhat belied by his quantitative modeling practices. These practices

often involve using one comparatal dispersion value to represent multiple (supposedly different) values. Given that the DSW model is successful in such cases, one can arrive at the following conclusion: if the distributional dispersions associated with physically different stimuli are not equal, then the DSW model appears to be robust in the face of this violated (Thurstonian) assumption.

A fourth and final fundamental assumption made by Hellström is that the decisional structure is devoid of bias. This last statement deserves qualification. The quantitative DSW model incorporates a parameter that is capable of accounting for the decisional bias postulated by Luce et al. (1963)¹⁶. Hence, the extent to which Hellström has “assumed” the absence of bias is subject to debate.

Hellström has postulated decisional mechanisms that underlie comparative judgment. Within this structure, bias has an expected value of nil. It has been alleged that judgments of “lesser” and “greater” are made when subjective difference magnitudes exceed category thresholds (referred to as “T’s” by Hellström). An “equal” judgment has been associated with the failure to surpass either category threshold.

Category thresholds are not arbitrary, but meet certain criteria: i) they are identifiable with specific values found in the subjective difference distribution; ii) upper and lower category thresholds are equidistant from the mean subjective difference. These requirements hold even when one assumes (as does Hellström) that category thresholds are themselves random variables.

Greenberg (1965) showed that this type of decisional model might be adapted to situations in which equality judgments are not permitted. Essentially, one equally divides

judgments of “no difference” amongst the two remaining modes of judgment (e.g., lesser and greater) when equality judgments are not permitted. Thus, Hellström’s model for three-category judgment can be modified and applied to two-category judgment. When only two categories of judgment are available, individuals set a fixed criterion that decides the weight of evidence required for a particular judgment. This criterion is located at the subjective difference value of zero when the decision process is unbiased.

The DSW model’s testable postulate is that stimuli regress towards or away from a RL, depending upon how the stimulus-RL relationship is weighted. As such, DSW does not involve direct modification of subjective stimulus values. It is the difference between stimuli and the RL that is weighted (Hellström, 1979). In other words, differential sensation weighting determines the extent to which a stimulus’ scale value assimilates towards the RL, or contrasts with it. One should note that RL is not simply used in lieu of AL. The former term is free to take on any value, while the AL is conceptually knotted to a weighted, geometric mean.

TOE has been attributed to both assimilation and contrast effects. However, prior to the DSW model, contrast relationships were primarily postulated to exist between presented stimuli, or between presented stimuli and backgrounds. Pratt (1933) and Needham (1935b) addressed the possibility that stimuli may contrast with ‘background stimulation’, but did not discuss the issue in depth. Hellström’s major theoretical contribution to psychophysics has been the detailed description of how stimuli may contrast with an unobserved RL.

While the DSW model has successfully fit data obtained from comparative judgment experiments, a precise understanding of factors responsible for TOE has not emerged. This is the case because the weighting parameters, which are of pivotal importance to the model, are not clearly understood. According to the model, factors that affect TOE must do so by having an effect upon the RL, or the weighting parameters. Hellström (1985) has identified some of these factors in a review of TOE literature: i) level of stimulus magnitude, ii) intensity of an interpolated stimulus, iii) stimulus (presentation) duration, iv) ISI length, and v) ISI set. One might also add another factor to this list: vi) presentation background. If the DSW model is valid, then a full understanding of TOE will be accompanied by a full understanding of how these factors influence the RL and weighting parameters.

The DSW model can be simplified by assuming that one of the compared stimuli is veridically perceived (not subject to weighting). It can be shown that the empirical test associated with this “special case” of the DSW model is the same as that used to test the full model. Thus, one must primarily rely upon conceptual grounds to distinguish between the full DSW model and its embedded cousin.

The full DSW model is as follows:

$$D = k [s_1 \psi_1 + (1 - s_1) \psi_{RL1}] - [s_2 \psi_2 + (1 - s_2) \psi_{RL2}], \quad (5)$$

where s_1 and s_2 represent weighting coefficients, ψ_1 and ψ_2 denote subjective scale values for the first and second stimuli, ψ_{RL1} and ψ_{RL2} reflect RLs for the first and second stimuli. The scale constant associated with the relevant psychometric function is k and D denotes subjective difference.

It can be seen from Equation 5 that TOE will decrease as the RL approaches the value of the comparison stimuli. In fact, TOE will be absent if comparison stimuli are equal to each other and to the RLs ($\psi_1 = \psi_2 = \psi_{RL1} = \psi_{RL2}$). TOE will also be absent if both of the weighting coefficients are equal to one. Allan (1979) has suggested that this latter mechanism could be used to explain the reduction of TOE that sometimes accompanies practice (i.e., participants' weighting coefficients tend towards the value of one with practice).

Stimulus contrast with the RL occurs when a weighting coefficient is greater than one. Assimilation towards the RL occurs when a stimulus' weighting coefficient is less than one.

Hellström (2000) has simplified Equation 5 to the following form.

$$D = W_1 \psi_1 + W_2 \psi_2 + U \quad (6)$$

In this case, $W_1 = k s_1$, $W_2 = -k s_2$ and $U = W_2 \psi_{RL2} - W_1 \psi_{RL1} + k(\psi_{RL1} - \psi_{RL2})$.

The simplification brings the model visibly closer to being testable by linear regression. Scale values for ψ_1 and ψ_2 might be obtained by applying a known psychophysical function (e.g., Steven's power law) to physical stimulus values. However, D cannot be directly measured.

Hellström (1977) scaled D by the comparatal dispersion, σ_c to achieve D/σ_c , which is standard normal. He postulated that the standard normal deviates for the proportion of greater or lesser judgments made by a participant in a two-category judgment task are equal to D/σ_c . Essentially, D/σ_c values are given by the standard normal deviates associated with the proportion of 'second greater' judgments, z_{LG} .

These D/σ_c values are also given by the standard normal deviates for the proportion of “first greater” judgments, z_{LG} , because $z_{LG} = z_{GL}$.

Thurstone’s (1927) Case V equation asserts

$$(2^{-1/2}) z_{GL} = \psi_G - \psi_L \quad (7)$$

If both sides of Equation 6 are scaled by the comparatal dispersion value (σ_c), then the left side becomes a known value (either z_{GL} or z_{LG} , but not both simultaneously). Arbitrarily, we shall deal with values of z_{LG} .

$$z_{LG} = (W_1/\sigma_c) \psi_1 + (W_2/\sigma_c) \psi_2 + (U/\sigma_c) \quad (8)$$

The equation is deliberately arranged such that parenthesized values represent to-be-estimated parameters when this model is tested via linear regression. Of course, a psychophysical function, relating physical and subjective magnitude, must be adopted in order to test this model. The popularly applied form of Steven’s power law is as follows:

$$\psi = k_1 \phi^B \quad (9)$$

Herein, psi (ψ) represents subjective magnitude. Phi (ϕ) represents the physical magnitude of a stimulus. Beta (B) is an exponent that is unique to the relevant sensory dimension (e.g., visual length) and k_1 is a scale constant. When Steven’s power law is assumed, Equation 8 becomes

$$z_{LG} = (W_1/\sigma_c) \phi^B + (W_2/\sigma_c) \phi^B + (U/\sigma_c) \quad (10)$$

Highlighted terms reflect those which would be entered as parameters in a regression equation that predicted z_{LG} . Hellström (2000) treats W_1/σ_c , W_2/σ_c and U/σ_c as three parameters rather than six. The intercept of Equation 10 differs slightly from the

intercept in Equation 8. The Equation 10 intercept is given by $(U'/\sigma_c) = W_2 \phi_{r2}^B / \sigma_c - W_1 \phi_{r1}^B / \sigma_c + k (\phi_{RL1}^B - \phi_{RL2}^B) / \sigma_c$, where ϕ_{r1} and ϕ_{r2} represent the physical magnitudes associated with the RLs. The test of this model is straightforward from the regression standpoint.

$$f(x_1, x_2) = a_1 x_1^B + a_2 x_2^B + a_0 \quad (11)$$

The dependent variable is the standard normal deviate associated with the proportion of greater responses (z_{LG}). The coefficients a_1, a_2 and a_0 reflect $W_1/\sigma_c, W_2/\sigma_c$ and U'/σ_c , respectively. x_1 and x_2 indicate the physical values of the first and second to-be-compared stimuli.

The major disadvantage of regressions performed according to Equation 11 is that a separate regression must be performed for each level of ISI. Each different regression requires an additional three parameters. Although, RL is not a parameter in this particular test of the DSW model, Hellström has used an equality for its' estimation. This equality holds true when there is only one unique RL ($\psi_{RL1} = \psi_{RL2}$). It will be referenced in a forthcoming model derivation.

$$\phi_r^B (-a_1 - a_2) = a_0 \quad (12)$$

Hellström's theoretical models do not incorporate a bias parameter. Nonetheless, a bias parameter could be added to the model in Equation 10 without changing its empirical test. In order to accomplish this task, one might simply append an additive bias parameter to the equation. Such a bias would be similar to that described by Luce et al. (1963). The resulting DSW model would appear as follows:

$$z_{LG} = (W_1/\sigma_c) \phi^B + (W_2/\sigma_c) \phi^B + (U'/\sigma_c) + C \quad (13)$$

Regression Equation 11 would be used to test this model. In this event, a_0 would represent $U'/\sigma_c + C$ as opposed to U'/σ_c . If C represents bias, then the expected value of C is zero for an unbiased model of decisional processes. One might statistically test for this type of bias by performing a one-sample t-test upon participants' C values¹⁷. If this test were to be significant, then there would be some evidence for the existence of a simple decisional bias.

The DSW model may be formulated with Fechner's law as the psychophysical function.

$$\psi = k_2 \log (\phi) \quad (14)$$

In this case, k_2 is a scale constant, and ψ is subjective magnitude. The base of the log transformation is irrelevant. Although esteem for Fechner's law has declined in recent decades, there is reason to prefer its usage when modeling TOE effects: this psychophysical function involves one less parameter than Steven's power law. If Fechner's law aptly describes the scaling of stimuli, then its utilization may be desirable in cases where experimental conditions (and z_{LG} data points) are few. In cases where a data set is small, overfitting complicates model validation. Eliminating superfluous parameters can mitigate the overfitting hazard.

Nevertheless, it has been contended "the logarithmic transform of Fechner's Law arises because, for the number of degrees of freedom typically required in the *chi squared* model, the logarithm of a chi squared variable is, to a good approximation, normal" (Laming, 2001, p. 36). If this is a valid claim, then it certainly denigrates the value of

Fechner's law. However, one must consider the possibility that scales are artificial constructs whose merit should be evaluated primarily on the basis of functionality.

Markley, Ayers, and Rule (1969) performed both magnitude estimation and similarity experiments involving line length. They found support for Fechner's law. Parker, Schneider, and Kanow (1975) similarly found support for Fechner's law with respect to line length. Albeit, Stevens' power law provided a slightly better fit to their data.

Hellström (1985, 2000) has generally found that RLs for first and second stimuli, ϕ_{RL1}^B and ϕ_{RL2}^B , tend to be the same. This finding implies that the first comparison stimulus does not alter the RL. Ergo, a RL is not greatly influenced by any single stimulus. When RLs for first and second stimuli are the same, the DSW model holds that TOE effects are caused solely by a difference amongst weighting parameters.

Two sub-mechanisms appear to be at work within the weighting of stimuli. Both the absolute magnitude of the weighting coefficients and the difference between coefficients are salient. *Coefficient magnitude reflects the extent to which a stimulus moves towards (or away from) the RL, irrespective of the other stimulus¹⁸*. This aspect of the weighting mechanism explains why response asymmetry should exist for absolute judgments. Comparison is not necessary for the distortion of subjective magnitude. A stimulus that is presented for absolute judgment will move either towards or away from the RL.

Coefficient differences reflect varying relationships between the RL and the first and second stimuli. Thus, a difference between weighting coefficients indicates an

interaction between the RL and presentation order. This interaction can scarcely be described as anything but a pairing effect.

In the context of the DSW model, pairing effects have been said to occur when first and second stimuli are differentially weighted. Most theorists have attributed pairing effects to proactive and/or retroactive contrast, which may be conceived as an interaction between presentation order and stimulus magnitude. The DSW model can approximate results that would be obtained if stimuli directly assimilated to or contrasted from one another. However, there is a large conceptual gap between direct assimilation/contrast effects and those that use the RL as a proxy.

Derivation of a new DSW model

Studies involving the reproduction of single, or paired stimuli do not provide strong evidence for veridical perception. However, it is possible that reproduction paradigms do not accurately capture the perception process. This is not an argument, but a caveat. If veridical perception does take place, it may well be limited to scenarios in which judgment is made during stimulus presentation. The majority of comparative judgment experiments have required that judgments be made when the second stimulus is no longer available for consultation. In contrast, the experiments described herein require that judgments be made while the second stimulus is still presented. In this context, it seems appropriate to evaluate the feasibility of a DSW model that presumes veridical perception of one stimulus.

If both comparison stimuli are veridically perceived on a systematic basis, then there cannot be TOE. However, TOE is possible when one stimulus is veridically

perceived and the other is not. According to the DSW model, a stimulus is accurately perceived when its weighting coefficient assumes the value of one. In this case, the stimulus will neither assimilate to, nor contrast with the RL. Hence, a veridical perception model is embedded in the DSW model. This embedded model involves only one weighting coefficient and only one RL. Equation 5 is simplified by fixing s_2 at the value of one.

$$D = k \{s_1 \psi_1 + (1 - s_1) \psi_{RL} - \psi_2\} \quad (15)$$

We next arrive upon z_{LG} by scaling D by the comparatal dispersion.

$$z_{LG} = k \{s_1 \psi_1 + (1 - s_1) \psi_{RL} - \psi_2\} / \sigma_c \quad (16)$$

By assuming Steven's power law and rearranging k ,

$$z_{LG} = \{s_1 \phi_1^B + (1 - s_1) \phi_{RL}^B - \phi_2^B\} k / \sigma_c \quad (17)$$

Next, k / σ_c is distributed to obtain a structure more suitable for linear regression.

$$z_{LG} = \frac{k}{\sigma_c} s_1 \phi_1^B + \frac{k}{\sigma_c} (1 - s_1) \phi_{RL}^B - \frac{k}{\sigma_c} \phi_2^B \quad (18)$$

This model lends itself to linear regression. Those values, which are highlighted, can be treated as parameters. Direct estimates of the RL can be obtained if ϕ_{RL}^B is retained as a parameter. To minimize the number of estimated parameters and attain the simplest structure, the regression equation for the model in Equation 18 should assume the form,

$$g(x_1, x_2) = b_1 * x_1^B + b_2 * x_2^B + b_0 \quad (19)$$

$$b_0 = (1 - s_1) * k / \sigma_c * \phi_r^B$$

$$b_1 = s_1 * k / \sigma_c$$

$$b_2 = - k / \sigma_c$$

It is evident that Equation 19 is equivalent to Equation 11. Hence, the embedded model has predictive capability equal to that of the full DSW model. Therefore, one must distinguish between the two models on theoretical grounds.

The RL can be obtained by fitting Equation 19 and manipulating the obtained coefficients. One finds the RL for Equation 19 by making use of the same equality that can be used for Equation 11. This proof is provided in Appendix B. If there are multiple ISIs, then the RL (ϕ_{RL}^B) is found by regressing b_0 on $[-(b_1 + b_2)]$ (Hellström 2000). The resultant coefficient represents the RL.

To determine if two RLs fit the experimental data better than one, Hellström performed (1979) linear regressions involving the parameter estimates of Equation 12. This issue will not be explored in depth. The proposed, embedded DSW model requires only that the weighting coefficient of a stimulus take on the value of one. The veridical perception of one stimulus is technically harmonious with the existence of two RLs¹⁹.

Until now, we have dealt only with the proportion of ‘second greater’ responses as a dependent variable. It may also be desirable to collapse across presentation order and predict the average deviation from the symmetrical response proportion of .50. The first step in finding a subjective analogue to such a measure entails averaging the subjective differences for the two presentation orders.

$$\overline{D} = (D_{LG} + D_{GL})/2 \quad (20)$$

D_{LG} = subjective difference in the lesser, greater presentation order

D_{GL} = subjective difference in the greater, lesser presentation order

Equation 5 can be reformulated into,

$$\begin{aligned} \overline{D} = & \{k [s_1 \psi_1 + (1 - s_1) \psi_{RL1}] - [s_2 \psi_2 + (1 - s_2) \psi_{RL2}] \\ & + k [s_1 \psi_2 + (1 - s_1) \psi_{RL2}] - [s_2 \psi_1 + (1 - s_2) \psi_{RL1}]\} / 2 \end{aligned} \quad (21)$$

If one assumes that $\psi_{RL1} = \psi_{RL2} = \psi_{RL}$, then Equation 21 simplifies algebraically,

$$\overline{D} = k [(s_1 - s_2) (\psi_1 + \psi_2) + 2 (1 - s_1) \psi_{RL} - 2(1 - s_2) \psi_{RL}] / 2 \quad (22)$$

If one further assumes veridical perception of the second comparison stimuli occurs, then $s_2 = 1$, and Equation 22 simplifies to the following form:

$$\overline{D} = k[(s_1 - 1)(\psi_1 + \psi_2) + 2(1 - s_1) \psi_{RL}] / 2 \quad (23)$$

Equation 23 cannot be tested in its current form, because $\overline{D}$ does not correspond to any observable, participant response. However, Hellström (1977, 1978, 2003) has pointed out that when the distribution of subjective differences, D values, is uniform (a.k.a. rectangular), one can scale $\overline{D}$ by the reciprocal range of this distribution to arrive at Jamieson and Petrusic's T_e index. Thus, the DSW model is also compatible with the response index T_e . An excursus on this topic is provided in Appendix C.

$$T_e = \overline{D} / R_D \quad (24)$$

Assuming Steven's power law, Equation 23 is scaled by R_D to become

$$T_e = \{ k/R_D (s_1 - 1)(\phi_1^B + \phi_2^B) + k/R_D 2(1 - s_1)\phi_{RL}^B \} \quad (25)$$

This model can be evaluated by the following linear regression.

$$h(x_1, x_2) = c_1 * (x_1^B + x_2^B) + c_0 \quad (26)$$

$$c_0 = k / R_D * 2 * (1 - s_1) * \phi_{RL}^B$$

$$c_1 = k / R_D * (s_1 - 1)$$

In this case, the RL can be found by dividing the intercept (c_0) by -2 times the regression coefficient (c_1). This proof is furnished in Appendix B.

It is convenient to assume that subjective difference distributions have a rectangular form. The simplified version of the DSW model would not perform well in the event of a normal distribution. However, the possibility that the distribution is not normally distributed leads to some intriguing possibilities. If Hellström (1993) supposed correctly, then the manipulation of attentional factors could lead to changes in the form of a subjective difference distribution. Characteristically low attention levels might then grant greater predictive power to the DSW models involving T_e .

The current investigation

Several investigators have documented TOE in comparisons of visual extent (McClelland, 1943; Inomata, 1959; Hellström, 2003). Nonetheless, TOE effects associated with comparisons in this sensory dimension have not been extensively documented. Only two firm conclusions can be drawn about the manifestation of TOE in comparisons of line length: i) TOE tends to be negative and ii) the magnitude of a remembered extent increases with stimulus-duration (e.g., McClelland, 1943).

The current investigation aimed to explore the relationship between stimulus magnitude, ISI, and TOE with respect to comparisons of visual extent. The method of paired comparisons was chosen so that category effects could be discounted as a source

of TOE. Instructions were varied between experimental trials so that simple bias could be identified.

It was expected that comparisons of visual extent would be characterized by negative TOE and that this error would be sensitive to both stimulus magnitude and ISI. It was also hypothesized that TOE would decrease as participants became more practiced. Traditional data analyses were supplemented by evaluation of Hellström's DSW model and Link's wave theory model.

Experiment 1

Methods

Participants.

Twenty undergraduate students participated in two 1-hour sessions to fulfil a course requirement.

Apparatus.

SuperLab Pro V. 2.0, a software program, regulated instruction presentation, stimulus presentation, timing, event sequencing, and the recording of responses. This program was run on a Pentium(r) II Intel NMX computer with 511 MB of RAM. Timing was accurate to within 1 ms. Visual display was provided by a 19 inch ViewSonic PF790 Super VGA monitor that had a 118 Hz refresh rate, and was set to 1280 X 1024 pixel resolution. Participant responses were issued by way of a Microsoft serial mouse (from which the tracking device had been removed).

The stimuli were black, horizontal, visual extents measuring 2 pixels in thickness. Three pairs of extents were used. These pairs were 50-51, 200-204, and 400-408 pixels

in length. Visual angles for the shortest members of each pair, at a viewing distance of 40 cm, were 1.2 °, 4.7 °, and 9.3 °. Stimuli always appeared and disappeared, instantaneously, in the vertical center of the viewing screen. Either the first or second extent appeared in the horizontal center of the screen. The other line appeared in a location horizontally offset from the center by a distance of ± 5 pixels. This horizontal offset was counterbalanced against order.

Procedure.

Participants were tested in isolation. The testing room was illuminated and participants were allowed to adjust their distance from the viewing screen as they saw fit. Written instructions informed the participants that their task was to identify either the longer, or the shorter member of a line pair. The instructions indicated that the word ‘SHORTER’ would appear on the viewing screen when the shortest line was to be selected. Similarly, the appearance of the word ‘LONGER’ was indicative that the longest line was to be selected. Clicking the left mouse button corresponded to selection of the first presented line. Clicking the right mouse button corresponded to selection of the second line. Written instructions were verbally reinforced. After instructions were administered, three practice trials were performed in the experimenter’s presence.

For any given trial, a blank ISI separated the presentation of two, different line lengths. Each trial began with the appearance of a ‘SHORTER’ or ‘LONGER’ instruction. These words appeared above the screen’s center. Instructions were presented for 1000 ms before the first line appeared. This line remained on the screen for

2000 ms before disappearing. The instruction remained on the screen throughout the subsequent ISI. A second line followed the ISI and remained on the screen until participants made a response by right- or left-clicking the mouse. The inter-trial interval was fixed at 2000 ms. The ISIs were 500, 3000, and 6000 ms.

All participants performed 36 practice trials at the beginning of each session. These practice trials were excluded from the main analyses. After the practice block, participants completed two blocks of 72 trials (in each session). Over the two sessions, participants performed eight repetitions of each experimental condition.

Results and Discussion

Analysis of variance.

A four-way within subjects analysis of variance was performed to evaluate the effects of pair length, presentation order, instruction type, and ISI using the 2 X arcsine transformed proportion of correct responses as the dependent measure²⁰. The independent variables consisted of pair length with three levels (50-51, 200-204, and 400-408 pixels), presentation order with two levels (short-long, or long-short), instruction type with two levels (shorter, or longer), and ISI with three levels (500, 3000, or 6000 ms). The sphericity assumption was evaluated using Mauchly's test with an alpha level of .05. Unless otherwise stated, the sphericity assumption was retained, and the significance level, α , was set at .05 for all of the following tests.

The order X instruction X ISI interaction was significant, $F(2,38) = 5.76, p < .01$, partial $\eta^2 = .23$. Testing of the three-way interaction was followed by testing of the simple order X instruction interactions at each level of ISI. Of the three simple

interactions tested, only the order X instruction interaction at the 3000 ms level was significant, $F(1, 19) = 4.73, p < .05$. When a Bonferroni adjustment was made to α , none of the simple simple main effects of instruction within the 3000 ms level of ISI were significant. The three-way order X instruction X ISI interaction is presented in Figure 1.

 Insert Figure 1 about here

The TOE did not change direction when the instruction was altered. Thus, one can discard the notion that TOE is a product of simple response bias. It is clear that instruction type influenced TOE to the greatest extent at the 3000 ms ISI level. However, even in this condition, changing instruction failed to reverse the TOE's direction.

The pair X order X ISI interaction was also significant, $F(4, 76) = 5.037, p < .01$, partial $\eta^2 = .21$. The significant three-way interaction was followed by tests of the simple pair X order interactions at each level of ISI. These tests revealed simple pair X order interactions at the 3000 ms ISI, $F(2,38) = 8.97, p < .01$, and at the 6000 ms ISI, $F(2,38) = 11.04, p < .01$. When a Bonferroni adjustment was made to α , two simple simple main effects were significant. A significant simple main effect of order was found for the longest line pair (400-408 pixels) when the ISI was 3000 ms, $F(1,19) = 13.41, p < .01$. A significant simple simple main effect of order was also found for the longest line pair when the ISI was 6000 ms, $F(1,19) = 16.37, p < .01$. The pair X order X ISI interaction is shown in Figure 2.

 Insert Figure 2 about here

TOE became increasingly positive with ISI for the shortest line pair. It remained uniformly negative across ISI for the mid-range line pair. Finally, TOE became increasingly negative with ISI for the longest line pair. It seems improbable that a bias theory of TOE could account for the given interaction. If one attempts to explain the results in terms of Allan's (1977) bias paradigm, then several conclusions are inescapable: a) criterion values differ amongst line pairs; b) criterion values shift with ISI, and c) the direction of criterion value shift is of opposite direction in short and long pairs. One cannot entirely dismiss the possibility that systematically shifting bias is responsible for the observed TOE effects. Nonetheless, it is not apparent why this should be the case. Hence, the significant pair X order X ISI interaction may be viewed as strong evidence that the observed TOE is not primarily a product of bias.

A significant effect of pair magnitude was not found, $F(2,38) = 1.88, p = .17$. Neither was a significant effect of ISI found, $F(2,38) = 1.36, p = 0.37$. Overall accuracy was .54. Thus, participants had great difficulty discriminating between the comparison lines. TOE effects occurred when discriminative success was scarcely better than chance.

A planned, paired samples *t*-test was performed to determine if T_e differed between first and second trial blocks for either of the two experimental sessions. A significant difference between T_e values for the first block of trials ($M = -.09, SD = 0.13$) and the second block of trials ($M = -.06, SD = 0.16$) was found for a one-tailed *t*-test, $t(19) = -1.77, p < .05$. Thus, TOE effects did diminish with practice. This finding suggests that the mitigating influence of practice is not limited to comparisons performed with sound intensities (Needham, 1934a) and durations (Woodrow, 1935).

A binomial test, with a test proportion of .5, was used to evaluate the predominance of baseline positive or negative T_e values. Fourteen out of 20 participants demonstrated negative T_e values, $p = .12$. The result is not significant, but is still reasonably congruent with previous studies, which have found negative TOE to be prevalent when participants compare line lengths.

DSW model analysis.

In accordance with Hellström's DSW model, a multivariate linear regression was performed at each ISI. In this regression, the standard normal deviate for the proportion of 'longer' responses (z_{LG}) was a function of first, and second line lengths²¹. The data were collapsed across participants. Equation 19, which incorporates Steven's power law, was empirically fit to the data according to least squares criteria. This equation was then modified so that the DSW model, which incorporates Fechner's law, could also be tested. The modifications consisted of omitting the power parameters and logarithmically transforming both first and second stimuli.

The predicted values for the two alternate forms of the DSW model were obtained, and used in the calculation of change statistics. The two sets of predicted values, themselves, became predictor variables in the comparative analysis of models. The increment in explained variance associated with entering the power law variable after the logarithmic variable was not significant, $F(1,15) = 0.00$, $p = 1.00$. Fechner's law became the preferred psychophysical function, because it requires one less parameter than the power law. The data were overfit by the DSW model, which incorporated the power law. Issues of validity plague models that are overfit. Furthermore, overfitting

increases the variability of parameter estimates. Subsequent results pertain to the DSW model that incorporates Fechner's law²².

Since the data sets were particularly small, R_{press}^2 was chosen as a measure of association. Information about the R_{press}^2 statistic is provided in Appendix F. The DSW model was separately fit to the data at each level of ISI. The model significantly fit data at the 500 ms level, $F(2,3) = 10.43$, $p < .05$, $R = .94$, $R_{\text{press}}^2 = .55$. It also fit at the 3000 ms level, $F(2,3) = 152.85$, $p < .01$, $R = .995$, $R_{\text{press}}^2 = .94$ and at the 6000 ms level, $F(2,3) = 152.85$, $p < .01$, $R = .94$, $R_{\text{press}}^2 = .53$.

R_{press}^2 values of .55, .94, and .53 all indicate that a strong relationship exists between the dependent variable and the independent variables. The DSW model successfully predicted the TOE. The model most successfully described the variance in TOE at the 3000 ms ISI.

Coefficient values, t -test statistics, and the associated p -values are provided in Table 1.

 Insert Table 1 about here

There was congruence amongst the coefficients obtained for regressions performed at the 3000 ms, and 6000 ms ISI levels. The coefficients associated with the 500 ms ISI level were markedly different from those obtained at the other two levels. In most cases, the t -tests were significant, or approached significance. This significance is impressive given the extreme multicollinearity of the independent variables. In multiple regression, the confidence intervals for the coefficients expand with multicollinearity.

The variance inflation factor (VIF) associated with the regressions was 1906.36^{23} . The conspectus is that VIFs, which exceed 10, indicate the existence multicollinearity.

Multicollinearity's effect upon power coefficients may differ from its' effect upon standard, polynomial coefficients. Equation 12 was fit to the group collapsed data, and power law exponents were obtained for each level of ISI. The obtained power exponents for the 500 ms, 3000 ms and 6000 ms ISIs were 0.0006, 0.20, and 0.71, respectively. Only the last of these exponents can be conceived as plausible. It is not evident that ISI should affect the values of power law exponents, as documented. Unfortunately, little is known about the role played by ISI in scaling tasks.

A group RL of 293.76 pixels was calculated for the fitted DSW model that incorporated the logarithmic psychophysical function. The data set that was used to obtain the RL can be found in Appendix D. The SPSS command syntax used in the determination of this RL can be found in Appendix E. RL values were also obtained for each individual participant. Figure 3 is a box plot of participant RLs.

 Insert Figure 3 about here

Given that the distribution's *SD* is greater than its mean, one can conclude that the data are highly variable. Significance of the Kolmogorov-Smirnov statistic with Lilliefors correction for significance indicates that the distribution was not normally distributed, $K-S(20) = 0.314$, $p < .01$. As can be seen from the box plot, the RL distribution was positively skewed [skewness = 2.10, *SE* skewness = 0.51]. It was also leptokurtic

(kurtosis = 3.48, *SE* kurtosis = 0.992). The median RL was 250.72 pixels. A *z*-criterion of ± 2 *SD* was used to identify two outlying RL values of 2394.16 and 2983.43 pixels.

The extreme variability of RL values is problematic to the DSW model. The obtained RLs could certainly not be mistaken for PSEs obtained from a line length discrimination task based upon the method of constant stimuli. The calculated RLs were larger than both the arithmetic and geometric means of the stimulus set. The most plausible RL estimate was provided by the median RL value.

According to the DSW model, stimulus weights are the primary determinants of TOE. The empirically obtained stimulus weights can be used to answer two questions. Firstly, do stimuli assimilate to the RL, or contrast with the RL? Secondly, is the second-presented line perceived veridically?

One can obtain a ratio of the two stimulus weights (s_1/s_2) by taking the coefficients from the empirically fit model and dividing the coefficient for first stimuli by -1 times the coefficient for second stimuli. Equation 6 can be used to authenticate this procedure. The s_1/s_2 values obtained for the DSW model fits to the 500, 3000, and 6000 ms levels of ISI (using the collapsed data set) were .996, .998, and .953, respectively. These results suggest that the second stimuli were weighted more heavily than the first. Alternatively, it is possible that second stimuli were veridically perceived and that the first stimuli, with weights less than 1.00, assimilated toward the RL.

An analysis of the s_1/s_2 values obtained for each participant was also performed. The sample of s_1/s_2 values for each ISI was obtained. Unfortunately, the measures of central tendency obtained for the sample distributions did not help clarify the relationship

between stimulus weighting and ISI. Distributional analyses revealed that the assumption of normality is hazardous when dealing with s_1/s_2 . Statistics pertaining to the sample distributions of s_1/s_2 are provided in Table 2.

 Insert Table 2 about here

The normality assumption was rejected for the 500 ms ISI level, $K-S(20) = .431$, $p < .01$ and for the 6000 ms level, $K-S(20) = 0.34$, $p < .01$. In both cases, the distributions were negatively skewed and leptokurtic. The range of s_1/s_2 values was far greater than one would expect. The s_1/s_2 value of 1.00 is most closely associated with veridical perception. Comparison can be grossly inaccurate if the ratio of DSW coefficients, s_1/s_2 , greatly differs from 1.00. The distribution of s_1/s_2 values at the 500 ms level of ISI featured three outliers (defined by $|z| > 2$) whose unstandardized values all equaled -1.00. A fourth outlier of -1.00 was found in the sample distribution for the 6000 ms level.

If one of the DSW coefficients, s_1 or s_2 , assumes a value approaching -1.00, then the RL will disproportionately influence comparative judgment. A s_1 value of less than -1.00 could, depending upon the RL, lead to drastically inaccurate judgment. The negative weighting coefficient scenario might best be described as a theoretical morass.

The presence of extreme multicollinearity means that one cannot place certainty in the sign of any coefficient (ergo, one cannot be certain of the signs given to s_1 and s_2). Nonetheless, an analysis of weighting coefficients did afford some insight into the observed TOE effects.

Participants 9 and 13 had RL values of 136.88 and 156.44 pixels, respectively. The calculated s_1/s_2 values for participant 9 tended to be greater than 1.00 (1.29, 1.06 and 0.99 in order of increasing ISI level), and the s_1/s_2 values for participant 13 tended to be less than one (1.10, 0.92, 0.79 for increasing ISI level). Recall that s_1 and s_2 indicate whether a stimulus assimilates to the RL, or contrasts with the RL.

In Appendix G, the pair X ISI interaction is plotted in terms of T_e for all participants. When one looks at the absolute magnitude of the TOEs for each participant, then it is evident that participant 9 exhibits the greatest, negative TOE.

The weighting coefficient ratios obtained for participants 9 and 13 are markedly different. Nonetheless, the s_1/s_2 values for both participants decrease with ISI. TOE for the shortest pair moves in the positive direction as ISI lengthens. TOE for the longest pair moves in the negative direction as ISI lengthens. If perception of the second stimulus is veridical, then participant 9's strange result can be explained in terms of the first stimulus contrasting with the RL.

The derived DSW model that predicts T_e could not satisfactorily be tested, because only nine conditions existed once the data were collapsed across presentation order. Instead of testing the derived model at each ISI, it was fit to the entire set of data. This data set consisted of the group means for the pair X ISI interaction. The model was tested by Equation 26, which was modified to incorporate Fechner's law as the psychophysical function. Eliminating the power coefficients and adding a logarithmic transformation of the variables accomplished this modification. The model provided a

significant fit to the data, $F(1,7) = 13.01, p < .01, R = .81, R_{\text{press}}^2 = .34$. The moderate R_{press}^2 value indicates that this particular model merits further testing and consideration.

It was shown in Equation 15 that the DSW model could account for simple decisional bias. In order to determine if simple decisional bias occurred, the intercept from the regression used to calculate RL was determined for each participant²⁶. These values have an expected value of zero if there is no simple decisional bias of the type previously specified. Consequently, a one-sample t -test with an expected value of zero served to evaluate the presence, or absence of bias. The test was not significant, $t(19) = -.923, p = .37$.

Wave theory analysis.

Equation 4 was empirically fit to the data according to least squares criteria. The precise equation that was input into SigmaPlot 5's regression library is provided in Appendix E. The regression performed at the ISI level of 500 ms was significant, $F(1,4) = 21.97, p < .01, R = .92, R_{\text{press}}^2 = .65$. Meanwhile, the model did not fit the data associated with the 3000 ms ISI, $F(1,4) = 1.11, p = .35, R = .47, R_{\text{press}}^2 < 0$, or the 6000 ms ISI, $F(1,4) = 0.80, p = .42, R = .41, R_{\text{press}}^2 < 0$ ²⁴.

A R_{press}^2 value of .65 reflects a strong relationship. Wave theory did describe TOE effects at the shortest ISI, but it did not describe those effects found at the longer ISIs.

Experiment 2

A second experiment was performed to determine if the previous results were a product of the particular stimuli, and method of stimulus presentation employed. Two alterations to the design of experiment 1 were made so that reliability might be assessed

through systematic replication. Firstly, members of each comparison pair were made to differ by 3% rather than the 2% so as to improve discriminative accuracy. Secondly, the horizontal offset between first and second stimuli was changed from a fixed distance to a distance proportional to the lengths of the first stimuli.

It was hypothesized that these changes would have little affect upon the experimental findings. Namely, it was predicted that a pair X order X ISI interaction would once again be found. It was also theorized that DSW and wave theory models would fit the data in a manner similar to that observed in the first experiment.

Methods

Participants.

Eighteen undergraduate students participated in two 1-hour sessions to fulfil a course requirement

Apparatus.

Equipment and software were identical to those used in the first experiment. The stimuli were black, horizontal, visual extents measuring 2 pixels in thickness. Three pairs of extents were used. These pairs were 33-34, 200-206, and 400-408 pixels in length. Visual angles for the shortest members of each pair, at a viewing distance of 40 cm, were 0.8 °, 4.7 °, and 9.3 °. Stimuli always appeared and disappeared, instantaneously, in the vertical center of the viewing screen. Either the first or second extent appeared in the horizontal center of the screen. The remaining line appeared in a location horizontally offset from the screen center by a distance equal to 12% of the first extent. The offset of first and second extents was counterbalanced against order.

Procedure.

The procedure was identical to that of the first experiment.

Results

Analysis of variance.

In the same manner as the first experiment, a four-way within subjects analysis of variance was conducted to determine the effects of pair length, presentation order, instruction type, and ISI upon the transformed proportion of correct responses. There were three levels of line pair (33-34, 200-206, and 400-412), two levels of presentation order (less-greater, or greater-less), two levels of instruction (shorter, or longer) and three levels of ISI (500, 3000, or 6000 ms).

The sphericity assumption had to be rejected for the pair X order X ISI interaction. Thus, the Huynh-Feldt correction to degrees of freedom was used when determining the three-way interaction to be significant, $F(3.34, 56.70) = 5.27, p < .01$, partial $\eta^2 = .24$. This interaction is displayed in Figure 4. As in the previous experiment, TOE for the shortest line pair was negative at the shortest ISI, and became positive as ISI lengthened. TOE for the intermediate pair remained uniformly negative across ISI. TOE for the longest line pair became increasingly negative as ISI lengthened.

 Insert Figure 4 about here

The simple pair X order interactions at each level of ISI were tested. Simple pair X order interactions were found for the 3000 ms ISI, $F(2,34) = 8.15, p < .01$, and at the 6000 ms ISI, $F(2,34) = 6.09, p < .01$. Simple main effects of order were tested. A

simple main effect of order was found for the shortest line pair (33,34 pixels) when the ISI was 6000 ms, $F(1,17) = 5.61, p < .05$. A second simple main effect of order was found for the longest line pair (400-408 pixels) when the ISI was 3000 ms, $F(1,17) = 5.16, p < .05$.

A planned, paired samples, one-tailed t-test was performed to determine if T_e differed between first and second trial blocks. The test was not significant, $t(17) = .03, p = .96$. A binomial test, with a test proportion of .5, was used to evaluate the preponderance of positive or negative T_e values. Ten out of 18 participants demonstrated negative T_e values, $p = .82$.

DSW model analysis.

The data were analyzed in the same manner as described in the previous results section. Once again, the alternative psychophysical functions were evaluated. Stevens' power law function and Fechner's logarithmic function could not be differentiated on the basis of superior fitting capability. Consequently, the logarithmic function was incorporated into the DSW model.

The DSW model was fit, separately, to each level of ISI. The model successfully fit data at the 500 ms level of ISI, $F(2,3) = 24.64, p < .05, R = .97, R_{\text{press}}^2 = .71$. It also fit data the 3000 ms level, $F(2,3) = 116.40, p < .01, R = .99, R_{\text{press}}^2 = .94$ and at the 6000 ms level, $F(2,3) = 92.50, p < .01, R = .99, R_{\text{press}}^2 = .91$.

The R_{press}^2 values indicate that the DSW model fit the data exceptionally well. Coefficient values, coefficient test statistics, and the associated p -values for the regressions are provided in Table 3.

 Insert Table 3 about here

With only one exception, the coefficients were significant. These significant coefficient tests came, notwithstanding severe multicollinearity. The VIF for the three regressions performed in the second experiment was 1257.52. This VIF value is considerably lower than that obtained in the previous experiment. However, it is still exceedingly large. Equation 12 was also fit to the data so that Stevens' power law exponents might be obtained for the ISI levels of 500, 3000, and 6000 ms. The exponents were 0.0001, 0.001, and 0.0003, respectively. It is puzzling that these exponents should be even more implausible than those obtained for the first experiment, where multicollinearity was a greater problem.

An RL of 178.44 pixels was obtained. The data set and SPSS syntax that were used to calculate this RL are presented in Appendices C and D, respectively. The DSW s_1/s_2 values obtained for the 500, 3000, and 6000 ms regressions were 1.00, 0.964, and 0.965, respectively.

RL and DSW weighting coefficient values were also obtained for each individual participant. A box plot of the participant RLs is shown in Figure 5. The assumption of normality could be retained according to the Kolmogorov-Smirnov test, $K-S(18) = .18, p = .15$. The distribution RLs is evidently more homogeneous than the one obtained in the previous experiment. It is possible that this fact has bearing upon the superior fits provided to this set of data by the DSW model.

 Insert Figure 5 about here

The distribution of s_1/s_2 values at each ISI was also examined. At the 500 ms ISI, the assumption of normality was retained, $K-S(18) = .195$, $p = .07$. Meanwhile, at the 3000 ms ISI, the assumption of normality was rejected, $K-S(18) = .209$, $p < .05$. The distribution was positively skewed and leptokurtic. The assumption of normality was also rejected at the 6000 ms ISI, $K-S(18) = .220$, $p < .05$. The distribution was platykurtic. There was one outlier in each sample of s_1/s_2 values associated with an ISI. None of the values were less than zero. Table 4 provides specific statistics pertaining to the sample distribution of s_1/s_2 values.

 Insert Table 4 about here

The derived DSW model was empirically fit to the T_e data points for the pair X ISI interaction. model provided a significant fit to the data, $F(1,7) = 12.01$, $p = .01$, $R = .79$, $R_{\text{press}}^2 = .28$. The moderate strength of association lends credence to the model's validity.

The presence of simple decisional bias was evaluated using a one-sample t -test. The non-significant result, $t(17) = .20$, $p = .84$ suggests that simple decisional bias did not contribute to the observed TOEs.

Wave theory analysis.

Once again, Equation 4 was empirically fit to the data. The regression performed at the 500 ms ISI level was significant, $F(1,4) = 30.30$, $p < .01$, $R = .94$, $R_{\text{press}}^2 = .78$. The

model did not fit the 3000 ms level, $F(1,4) = 2.85$, $p = .17$, $R = .65$, $R_{\text{press}}^2 < 0$, or the 6000 ms level, $F(1,4) = 2.31$, $p = .20$, $R = .61$, $R_{\text{press}}^2 < 0$.

Experiment 3

AL theory and the DSW model both postulate that remote anchors play key roles in comparative judgment. Meanwhile, contrast explanations of TOE hold that asymmetric error is due to an interaction between series stimuli. On the one hand, TOE is a product of human calibration. On the other hand, TOE is a product of differential carryover effects, or sequential dependencies.

Interpolation experiments have traditionally been employed to study contrast effects. When inserted into the time interval separating two to-be-compared stimuli, an extraneous stimulus tends to proactively contrast with the second stimulus (Needham, 1935a). The degree and type of contrast between the interpolated stimulus, and the second comparison stimuli, determines the direction and magnitude of the TOE.

Interpolated stimuli might primarily affect series stimuli, or remote anchors. Experiment 3 was performed to differentiate between these two possibilities.

If an interpolated stimulus primarily influences a remote anchor, then it is not evident that presentation sequence should greatly affect TOE. However, if an interpolated stimulus proactively contrasts with subsequently presented stimuli, then TOE should be affected by the temporal proximity between the interpolated anchors and the comparison stimuli.

In a third experiment, small and large stimuli were interpolated into varying positions within a fixed ISI. Participants were told to ignore the interpolated lengths, and

focus upon the comparison lines. It was predicted that TOE would be influenced by magnitude of the interpolated stimuli. It was also predicted that there would be an interaction between the effects interpolative magnitude and temporal location of the interpolation (within the fixed ISI).

Methods

Participants.

Twenty-one undergraduate students participated in a single, 1-hour session to fulfil a course requirement.

Apparatus.

Equipment and software were identical to those used in the first experiment. The stimuli were black, horizontal visual extents measuring 2 pixels in thickness. One pair of extents served as comparison stimuli, while two other extents alternately served as an interpolated stimulus. The comparison extents were 200 and 206 pixels in length. The interpolated extents were 47 and 850 pixels in length. These lengths correspond to visual angles of 4.7° for the smallest comparison extent and 1.1° and 19.0° for the interpolated extents. Stimuli always appeared and disappeared, instantaneously, in the vertical center of the viewing screen. Interpolated lines always appeared in the horizontal center of the screen. Meanwhile, either the first or the second extent was horizontally centered. The remaining extent appeared in a location horizontally offset from the screen center by a distance of 24 pixels. The offset of first and second extents was counterbalanced against order.

Procedure.

Instructions were similar to those of the first two experiments. Participants were informed as to the nature of the “SHORTER” and “LONGER” instructions. They were also told how responses might be issued via the mouse. Furthermore, participants were instructed that three lines would appear on the viewing screen and that they should focus solely upon comparing the first and last lines.

Each trial began when an instruction appeared above the center of the viewing screen. This instruction remained on the screen for the duration of a trial. The first stimulus appeared 1000 ms after this instruction and remained on the screen for 1000 ms before disappearing. An ISI of 3000 ms followed. The interpolated stimulus was presented either 500, 1250, or 2000 ms into this ISI. This interpolated line was presented for 500 ms. A third visual extent appeared at the end of the 3000 ms ISI and remained on the screen until a response was provided.

All participants performed 24 practice trials at the beginning of each session. These practice trials were excluded from the main analyses. Participants completed 6 repetitions of each condition.

Results

A four-way within subjects analysis of variance was performed to determine the effects of presentation order, instruction type, location of an interpolated extent within an ISI, and interpolated stimulus magnitude upon the 2 X arcsine transformed proportion of correct responses.

The order X interpolated magnitude interaction was marginally significant, $F(1,20) = 4.28$, $p = .05$, partial $\eta^2 = .18$. This interaction is shown in figure 6. When the interpolated line was shorter, the TOE was most negative. The TOE was minimally negative when the interpolated line was longer. These results provide some support for the contrast theory of TOE. Albeit, it is not clear why negative TOE should be associated with a large interpolated line. In this circumstance, the contrast theory predicts positive TOE.

 Insert Figure 6 about here

It is also unclear why the order X interpolated magnitude X time-to-interpolation interaction did not approach significance, $F(2,40) = 0.15$, $p = .86$, partial $\eta^2 = .01$. While interpolative magnitude appeared to influence TOE, it did not matter when the interpolated lengths appeared within the ISI. This non-significant interaction is shown in figure 7.

 Insert Figure 7 about here

The order X instruction X interpolated magnitude interaction was significant, $F(1,20) = 5.06$, $p < .05$, partial $\eta^2 = .20$. Figure 8 illustrates this three-way interaction. The simple interaction of order X instruction was significant for the larger (850 pixels) interpolated extent, $F(1,20) = 6.55$, $p < .05$. When a large line was interpolated into the ISI, TOE changed from positive to negative as instruction was shifted from “choose the shorter line” to “choose the longer line”. It is possible that this result arose because

participants were judging the interpolated lines rather than the comparison stimuli. Participants had been instructed to ignore the interpolated line so that this possibility would be avoided.

 Insert Figure 8 about here

General discussion

In the first two experiments, an interaction between stimulus level, presentation order and ISI was observed. In all cases, Hellström's DSW model fit the data well. Link's wave theory model provided a good fit to the data obtained for the shortest level of ISI. However, the wave theory model did not fit the TOEs associated with the longer ISIs. In general, the perceptually-based DSW model provided a superior account of the data. Typically, bias theories of TOE do not make provision for ISI effects. Thus, it is not wholly surprising that wave theory failed in this respect.

The DSW model was best applied in conjunction with Fechner's logarithmic psychophysical function. Stevens' power law overfit the data.. Furthermore, the power law exponents were not plausible. Nonetheless, Fechner's law was not conclusively superior. Multicollinearity may have adversely affected the empirical fits for the DSW model, which incorporated Stevens' power law.

The derived DSW model, which accounts for the dependent variable T_e , provided encouraging results. Unfortunately, a large number of stimulus pairs are required to properly evaluate the new model. The modicum of success that was achieved by the

model provides some corroborating evidence for the validity of Hellström's full DSW model.

The DSW model was not wholly supported. Negative TOE was found for the shortest line pair, at the shortest ISI, in each experiment. This occurrence cannot be explained via the DSW model²⁵. The results can be explained by supposing that a simple decisional bias acted in conjunction with the DSW mechanisms. It is possible that the statistical tests for this type of bias produced non-significant results, because of extreme variability in the coefficient data sets.

The results showed that response preference, and John's (1975) implicit response bias are not responsible for presentation order effects. In only one circumstance did TOE reverse direction with a change in instruction type. This reversal was observed in the third experiment, when a large line was interpolated into the ISI. In this case, the 'longer' instruction led to more 'second line longer' responses. It seems plausible that participants became confused about which line they should judge.

In the first two experiments, the TOEs for short and long pairs moved in opposite directions as ISI lengthened. Other researchers studying the comparison of visual length have obtained equivalent results (Hellström, 2003; Inomata, 1959). Although TOEs with duration comparisons also exhibit a levels effect as in line length comparisons, with positive TOEs for brief durations (e.g., 200, 225 ms) and negative TOEs for long durations (e.g., 5000, 5500 ms), ISI effects differ from those of the present line-length comparisons. Jamieson and Petrusic (1975a, 1975b) found that TOEs decrease with increasing ISI for both brief and long durations.

Varying the location of an interpolated line within a fixed ISI did not significantly affect the TOE. This result suggests that contrast effects were not responsible for the observed TOE. An alternative explanation is that the interpolated stimuli altered participants' RLs.

Several issues have not been resolved in this disquisition. Firstly, it is not clear why Link's (1992) wave theory should exclusively fit data associated with a brief ISI. Secondly, it seems that the DSW model is capable of accounting for simple decisional bias, and that this type of bias is present when individuals compare visual extents. Thirdly, the evidence against proactive contrast was inconclusive, because it relied upon a null result.

Future investigations into TOE might explore the relationship between TOE and the Signal Detection measure of sensitivity, d^1 . It also seems plausible that Link's (1992) Wave theory might be adapted to account for the effects of a subjective reference level. The DSW model is able to empirically capture TOE effects. Nonetheless, a thorough, conceptual understanding of TOE remains on the horizon.

References

- Allan, L. G., Kristofferson, A. B. & Rice, M. E. (1974). Some aspects of perceptual coding of duration in visual duration discrimination. *Perception & Psychophysics*, 15(1), 83-88.
- Allan, L. G. (1977). The time-order error in judgments of duration. *Canadian Journal of Psychology*, 31(1), 24-31.
- Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, 26(5), 340-354.
- Baranski, J. V. & Petrusic, W. M. (1992). The discriminability of remembered magnitudes. *Memory & Cognition*, 20(3), 254-270.
- Ekman, G. (1959). Weber's law and related functions. *The Journal of Psychology*, 47, 343-352.
- Ellis, H. D. (1973a). An examination of the adaptation-level theory account of time errors. *Journal of Experimental Psychology*, 98(1), 160-163.
- Ellis, H. D. (1973b). Proactive effects of interpolated anchors. *Journal of Experimental Psychology*, 98(2), 233-238.
- Favreau, O. E., & Everett, J. C. (1996). A tale of two tales. *American Psychologist*, 51(3), 268-269.
- Fechner, G. (1966). *Elements of psychophysics*. (Transl. By H. E. Adler). New York: Holt, Rinehart and Winston.
- Fernberger, S. W. (1931). On absolute and relative judgments in lifted weight experiments. *American Journal of Psychology*, 43, 560-578.
- Geertsma, R. H. (1958). Time-order errors in comparative judgments of hurtfulness. *Journal of Experimental Psychology*, 55(3), 284-288.
- Gleitman, H. (1957). Proactive and retroactive assimilation in the successive comparison of loudnesses. *American Journal of Psychology*, 70, 117-119.
- Greenberg, M. G. (1965). A modification of Thurstone's law of comparative judgment to accommodate a judgment category of "equal" or "no difference". *Psychological Bulletin*, 64(2), 108-112.

- Guilford, J. P. & Park, D. G. (1931). The effect of interpolated weights upon comparative judgments. *American Journal of Psychology*, 43, 589-599.
- Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Hellström, Å. (1977). Time errors are perceptual. *Psychological Research*, 39, 345-388.
- Hellström, Å. (1978). Factors producing and factors not producing time errors: an experiment with loudness comparisons. *Perception & Psychophysics*, 23(5), 433-444.
- Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance*, 5(3), 460-477.
- Hellström, Å. (1985). The time-order error and its relatives: mirrors of cognitive processes in comparing. *Psychological Bulletin*, 97(1), 35-61.
- Hellström, Å. (1993). The normal distribution in scaling subjective differences: less "normal" than we think? *Perception & Psychophysics*, 54(1), 82-92.
- Hellström, Å. (2000). Sensation weighting in comparison and discrimination of heaviness. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 6-17.
- Hellström, Å. (2001). Time-order effects for aesthetic preference. In Sommerfeld, E., Kompass, R & Lachmann, T. (Eds.), *Fechner Day 2001: Proceedings of the Seventh Annual Meeting of the International Society of Psychophysics* (pp.421 - 426). Lengerich: Pabst Science Publishers.
- Hellström, Å. (2003). Comparison is not just subtraction: effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics*, 65(7), 1161-1177.
- Helson, H. & Mawr, B. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology*, 47(1), 1-28.
- Helson, H., Michels, W. C. & Sturgeon, A. (1954). The use of comparative rating scales for the evaluation of psychophysical data. *American Journal of Psychology*, 67, 321-326.
- Helson, H. (1964). *Adaptation-Level Theory*. New York: Harper & Row.
- Inomata, S. (1959). The analysis of conditions behind the time-error in the successive comparison of visual length. *Japanese Psychological Research*, 7, 36-53.

- Inomata, S. (1964). Comparison of time-error phenomena in different sense-modalities. *Japanese Psychological Research*, 6(1), 10-20.
- Jamieson, D. G. & Petrusic, W. M. (1975a). The dependence of time-order error direction on stimulus range. *Canadian Journal of Psychology*, 29(3).
- Jamieson, D. G. & Petrusic, W. M. (1975b). Presentation order effects in duration discrimination. *Perception & Psychophysics*, 17(2), 197-202.
- Jamieson, D. G. & Petrusic, W. M. (1975c). Pairing effects and time-order errors in duration discrimination. *Perception & Psychophysics*, 18(2), 107-113.
- John, D. (1975). A common mechanism mediating the time-order error and the cross-over effect in comparative judgments of loudness. *Australian Journal of Psychology*, 27(1), 51-60.
- Leth-Steensen, C. (1994). *Memory load and the rate of forgetting of unidimensional perceptual stimuli*. Unpublished master's thesis, Carleton University, Ottawa.
- Link, S. W. (1992). *The Wave Theory of Difference and Similarity*. Hillsdale: Lawrence Erlbaum associates.
- Luce, D. R., Bush, R. R. & Galanter, E. (1963). *Handbook of Mathematical Psychology*. New York: John Wiley & Sons.
- Markley, R.P., Ayers, D. & Rule, S. J. (1969). Similarity judgments of line length. *Perception & Psychophysics*, 6(1), 58-60.
- McClelland, D. C. (1943). Factors influencing the time error in judgments of visual extent. *Journal of Experimental Psychology*, 33(2), 81-95.
- Michels, W. C. & Helson, H. (1954). A quantitative theory of time-order effects. *American Journal of Psychology*, 67, 323-334.
- Needham, J. G (1934a). The time-error as a function of continued experimentation. *American Journal of Psychology*, 46, 558-567.
- Needham, J. G. (1934b). The time-error in comparison judgments. *Psychological Review*, 31(4), 229-243.
- Needham, J. G. (1935a). Interpolation effects with different time intervals. *Journal of Experimental Psychology*, 18, p. 767-773.

- Needham, J. G. (1935b). The effect of the time interval upon the time-error at different intensive intervals. *Journal of Experimental Psychology*, 18, 530-543.
- Parker, S., Schneider, B. & Kanow, G. (1975). Ratio scale measurement of the perceived lengths of lines. *Journal of Experimental Psychology: Human Perception and Performance*, 104(2), 195-204.
- Peak, H. (1940). The time order error in successive judgments and in reflexes: time error theories. *Psychological Review*, 1940, 47.
- Petrusic, W. M., Baranski, J. V. & Kennedy, R. (1998). Similarity comparisons with remembered and perceived magnitudes: memory psychophysics and fundamental measurement. *Memory & Cognition*, 26(5), 1041-1055.
- Petrusic, W. M., Harrison, D. H. & Baranski, J. V. (2004). Long-term memory for elementary visual percepts: memory psychophysics of context and acquisition effects. *Perception & Psychophysics*, 66(3), 430-445.
- Poulton, E.C. (1989). *Bias in Quantifying Judgments*, New Jersey: Lawrence Erlbaum Associates.
- Pratt, C. C. (1933). The time-error in psychophysical judgments. *American Journal of Psychology*, 45, 292-297.
- Rambo, W. W. & Johnson, E. L. (1964). Practice-effects and the estimation of adaption-level. *American Journal of Psychology*, 77(1), 106-110.
- Schab, F. R. & Crowder, R. G. (1988). The role of succession in temporal cognition: is the time-order error a recency effect of memory? *Perception & Psychophysics*, 44(3), 233-242.
- Stevens, S. S. (1957). On the psychophysical law. *The Psychological Review*, 64(3), 153-181.
- Stevens, S. S. & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 54(6), 377-409.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah: Lawrence Erlbaum Associates.
- Thurstone, L. L. (1927). A law of comparative judgment. *The Psychological Review*, 34, 273-286.
- Thurstone, L.L. (1954). The measurement of values. *Psychological Review*, 61(1), 47-58.

- Tresselt, M. E. (1948). Time errors in successive comparison of simple visual objects. *American Journal of Psychology*, 57, 555-558.
- Tresselt, M. E. & Mayzner, M. S. (1967). Anchoring and time-error. *Psychonomic Monograph Supplements*, 2(3, Whole No. 19), 39-42.
- Wickelgren, W. A. (1966). Consolidation and retroactive interference in short-term recognition memory for pitch, *Journal of Experimental Psychology*, 72(2), 250-259.
- Woodrow, H. (1935). The effects of practice upon time-order errors in the comparison of temporal intervals. *The Psychological Review*, 42(2), 127-152.

Appendix A

TOE in the Context of Classical Psychophysics

TOE and the just noticeable difference

Suppose one performs an experiment according to the method of constant stimuli, wherein a standard stimulus alternately precedes and follows the variable stimuli. In this scenario, a JND can be obtained for each presentation order (standard-variable and variable-standard). One might achieve an index of TOE by taking the difference between these two JNDs. Researchers tend not to use such a representation. However, the plausibility of such a measure has been established (e.g., Luce et al., 1963). A JND index of TOE would be appropriate, because greater discriminative performance in one of two possible stimulus orders necessitates a lower JND for that order. Accuracy is not the same across presentation orders when TOE is found²⁶. Ergo, one cannot assume that difference thresholds are the same for the two presentation orders when TOE is found. When there is a large amount of TOE, an investigator might consider reporting two JNDs rather than one.

The relationship between the JND and absolute magnitude of the TOE tends to be positive and monotonic (e.g., Baranski & Petrusic, 1992). TOE is a type of comparative error and greater error is associated with a greater JND. Currently, one can only speculate as to whether the JND is linearly or curvilinearly related to TOE. The curve-fitting procedures needed to diagnose linearity (and non-linearity) require a large number of data points. Consequently, an experiment featuring a large number of standard stimuli must be designed and executed if we are to determine the precise nature of the TOE-JND

relationship. The creator of such an experiment would have to address the issue of power.

Since difference thresholds are affected by TOE, so too should be the confusion scales, which are built from these empirically obtained limens²⁷. This corollary has been largely ignored. Fechnerian scaling²⁸ received diminished attention since Stevens (1957) issued his famous polemic for the use of direct scaling measures. Poulton's (1989) apt exposé of the flaws inherent in direct scaling methodologies has not changed this status quo.

TOE and the PSE

Findings of negative TOE prevail, but are not the rule, for comparative judgment experiments dealing with the sensory dimensions of weight, sound intensity, temporal duration, and line length²⁹. Typically, TOE is positive when less intense stimuli are compared and negative when moderate or intense stimuli are compared. A tolerable generalization would be that PSEs tend to be less than standard stimuli.

The PSE obtained from a comparative judgment experiment also tends to change with ISI. Needham (1934a) coined the term 'pause -interval function', or 'p -function' to express changes in TOE as a function of ISI. Using the method of constant stimuli, Needham (1934a, 1935a) performed experiments involving comparisons of sound intensity and found that PSEs changed over ISI. Inomata (1959) also observed p-functions for line length comparisons performed according to the method of constant stimuli. Unfortunately, many comparative judgment experiments performed in the first half of the 20th century kept ISI fixed. Early investigators' inattention to time course

seems to have deleteriously affected subsequent theory. Several investigators have advanced theories of TOE, which fail to address inter-stimulus interval effects.

Moreover, the results of any scaling experiments are suspect if the PSE is calculated via the traditional method. Woodrow (1935) found evidence that upper and lower difference thresholds are unequal. This result was taken as evidence that it would be inappropriate to portray the PSE as a midpoint between upper and lower difference limens when TOE occurs.

Woodrow performed a duration discrimination experiment³⁰ involving the comparison of unfilled, temporal intervals. Five variable stimuli were compared to one standard stimulus and judgments of “shorter”, “longer” or “equal” were elicited. Participants performed 120 comparisons, with standard preceding variable in half of the trials and with variable preceding standard in the remaining trials. The TOE was negative. However, the notable finding was that the upper difference threshold exceeded the lower difference threshold. An overly large PSE would have been calculated had the experimenters used the difference limens to arrive at their value. Fortunately, a variety of methods for determining the PSE exist.

TOE and the psychometric function

Data obtained from comparative judgment experiments based upon the method of constant stimuli can be displayed in different manners. Typically, the proportion of times a variable stimulus is judged greater than a standard stimulus is plotted against the physical difference between stimuli and the standard. Values along either axes may be

standardized. This method of display necessitates the provision of different plots for each experimental standard.

The functions that best fit the observed data are called psychometric functions. They are typically sigmoidal (S-shaped). JNDs are reflected in the slopes of the function. One cannot observe the PSE directly. However, one can observe if a PSE is less than, equal to, or greater than a standard stimulus by finding the difference associated with the response proportion of .50. This difference will be zero if the PSE and standard are of equal magnitude.

Given that TOE is tied to the JND, one expects the error to be visually identifiable within the psychometric function. This expectation is partly realized. The zero difference is not associated with the proportion of .50 when TOE exists. This is not the only method of identifying the presence of TOE. Left and right portions of the function reflect the two different presentation orders: greater-lesser and lesser-greater, respectively. Each half of the psychometric function has a different slope when one jnd fails to describe both presentation orders. Unfortunately, slope interpretation is difficult when the psychometric function is sigmoidal.

It is also possible to identify TOE by comparing psychometric functions for different standard stimuli. Baranski and Petrusic (1992) performed a comparative judgment experiment involving line lengths. They found that those standard stimuli, which were characterized by greater TOE, tended to have shallower psychometric functions. Link (1992) insinuated that such a finding should be commonplace.

Hence, any procedure, which calculates one JND from the slope of the psychometric function is suspect. Historically, this point has been of little concern to psychologists. The mistreatment³¹ of TOE is exemplified in a psychometric function example provided in the Handbook of Mathematical Psychology (Luce et al., 1963, p. 200). The span of two JNDs is visually demarcated. This span straddles the midpoint of the psychometric function. Thus, the two encompassed JNDs represent different presentation orders. The presence of TOE dictates that the JNDs for each order are different. The author's facilitated the belief that JNDs for each presentation order are equal by referring to two unequal JNDs as 'two JNDs'.

Luce et al. (1963) asserted that the empirical³² and generalized³³ forms of Weber's law are validated by different psychophysical functions. The authors concluded that when Weber's empirical law describes the psychometric function, his general law does not; the psychophysical functions that validate the two forms of Weber's law are related by an asymmetrical 'constant error'³⁴. The term error is appropriate insofar as the commonly used, empirical form of Weber's law does not accurately represent the general law.

A formulaic description of this constant, asymmetrical error associated with the empirical form of Weber's law was provided. A detailed exposition of this formula would necessarily be lengthy. Principally, asymmetrical error was found to be a linear function of stimulus magnitude.

It should be noted that Luce et al. did not attempt to explain experimentally observed TOE by way of their discovery. They only sought to introduce a more suitable

method of validating Weber's law. This method consisted of an alteration in the empirical Weber function. Nonetheless, their finding is important because it draws attention to the connection between the psychometric function and theory. The JND and PSE are implicated in this discussion by way of their association with the form of the psychometric function.

Given that the PSE, JND, and difference limens are related to presentation order, the utility of the psychometric function must be questioned. Does the graphed function accurately depict response patterns? Is there some alternative figure that might succinctly and accurately portray participant responses, while allowing for the ready identification of TOE? Lacking such a desideratum, we must accept that close scrutiny of the psychometric function is required. Differences between the JNDs for each presentation order cannot readily be discerned.

Appendix B

Proofs

1. Proof that the linear regression of c_0 upon $-(c_1 + c_2)$ will yield the relational level (ϕ_r^B) value for Equation (21).

i) Prove

$$\phi_r^B * (-b_1 + -b_2) = b_0$$

ii) Given

$$g(x_1, x_2) = b_0 + b_1 * x_1^B + b_2 * x_2^B$$

where

$$b_0 = (1 - s_1) * k/\sigma_c * \phi_r^B$$

$$b_1 = s_1 * k/\sigma_c$$

$$b_2 = -k/\sigma_c$$

iii) Proof

$$\phi_r^B * -(b_1 + b_2) = \phi_r^B * -[(s_1 * k/\sigma_c) - k/\sigma_c]$$

$$= \phi_r^B * -s_1 * k/\sigma_c + k/\sigma_c$$

$$= \phi_r^B * k * (-s_1 * 1/\sigma_c + 1/\sigma_c)$$

$$= \phi_r^B * k/\sigma_c * (1 - s_1)$$

$$= (1 - s_1) * k/\sigma_c * \phi_r^B$$

$$= b_0$$

2. Division of the intercept, q_0 , by the -2 times the coefficient, c_1 , will yield the relational level (ϕ_r^B) value for model (27).

i) Prove

$$c_0 / (-2 * c_1) = \phi_r^B$$

ii) Given

$$h(x_1, x_2) = c_1 * (x_1^B + x_2^B) + c_0$$

where

$$c_0 = k / R_d * 2 * (1 - s_1) * \phi_r^B$$

$$c_1 = k / R_d * (s_1 - 1)$$

iii) Proof

$$c_0 / (-2 * c_1) = \{ k / R_d * 2 * (1 - s_1) * \phi_r^B \} / \{ -2 * k / R_d * (s_1 - 1) \}$$

$$= \{ k / R_d * (1) * (1 - s_1) * \phi_r^B \} / \{ (-1) * k / R_d * (s_1 - 1) \}$$

$$= \{ k / R_d * (1 - s_1) * \phi_r^B \} / \{ (-1) * k / R_d * (s_1 - 1) \}$$

$$= \{ k / R_d * (1 - s_1) * \phi_r^B \} / \{ k / R_d * (-1) * (s_1 - 1) \}$$

$$= \{ k / R_d * (1 - s_1) * \phi_r^B \} / \{ k / R_d * [-s_1 + 1] \}$$

$$= \{ k / R_d * (1 - s_1) * \phi_r^B \} / \{ k / R_d * (1 - s_1) \}$$

$$= \phi_r^B$$

Appendix C

A Rationale for $T_e = \bar{D} / R_D$

Jamieson and Petrusic's T_e index is based upon response proportions for the short-long and long-short presentation orders. Hellström (1977,1978) proposed that this response index can be used to infer subjective difference. Recall equation 26:

$$T_e = \bar{D} / R_D$$

$\bar{D}$ is the mean subjective difference between two lines that have been compared in both presentation orders. R_D is the range of the distribution of $\bar{D}$ values.

The validity of equation 26 is not self-evident. In fact, the equation will only be valid if $\bar{D}$ is uniformly distributed.

The uniform distribution is characterized by equal probability density for all values bounded by points 'a' and 'b'. Values falling outside this range have a probability of zero.

Uniform distribution

The cumulative density function $[D(x)]$ for a uniform distribution is as follows:

$$D(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x > b \end{cases}$$

Note that ' $b - a$ ' is equivalent to the range of the distribution.

If $\bar{D}$ is normally distributed, then the $\bar{D}$ distribution will appear as follows:

Given this distribution, one can calculate the probability of obtaining a $\bar{D}$ value between

0 and $\bar{D}_1$. This probability is determined by taking the difference between the

cumulative densities for $|\bar{D}_1|$ and zero. The absolute value of $\bar{D}_1$ must be used in order for the probability to be guaranteed a positive sign.

$$p(0 \leq x \leq |\bar{D}_1|) = \frac{|\bar{D}_1| + a}{b - a} - \frac{(0 + a)}{b - a} = \frac{|\bar{D}_1|}{b - a} = \frac{|\bar{D}_1|}{R_D}$$

$R_D = b - a = \text{range of the uniform distribution}$

According to the decisional mechanism postulated by Hellström, the unbiased criterion (for a lesser or greater judgment) is located at the subjective difference value of zero. If perceptually rooted TOE occurs, then $\bar{D}$ values for a particular condition will be nonzero when, averaged across presentation order, the two to-be-compared lines are equal.

The probability of obtaining a nonzero $\bar{D}$ value is also the probability that the response proportion for either presentation order will deviate from .5 (when lines are presented in both orders, equally often). When $\bar{D}$ is uniformly distributed, $\bar{D} / R_D$ gives the probability of a nonzero value. Jamieson and Petrusic's TOE index provides the probability that the proportion of longer responses differs from the proportion of shorter responses.

$$T_e = \frac{1}{2} [p(\text{correct}|\text{GL}) - p(\text{correct}|\text{LG})]$$

The probability of a correct response in the greater-lesser presentation order is equivalent to the probabi

Appendix D

Data Sets

1. Data set for experiment 1 DSW regressions.

ISI	1 st line	2 nd line	Proportion correct
500.00	50.00	51.00	.6540
500.00	51.00	50.00	.4840
500.00	200.00	204.00	.6880
500.00	204.00	200.00	.4030
500.00	400.00	408.00	.6580
500.00	408.00	400.00	.4780
3000.00	50.00	51.00	.5000
3000.00	51.00	50.00	.5840
3000.00	200.00	204.00	.6340
3000.00	204.00	200.00	.4690
3000.00	400.00	408.00	.7190
3000.00	408.00	400.00	.3970
6000.00	50.00	51.00	.4980
6000.00	51.00	50.00	.5480
6000.00	200.00	204.00	.5800
6000.00	204.00	200.00	.4510
6000.00	400.00	408.00	.7350
6000.00	408.00	400.00	.4150

2. Data set for experiment 2 DSW regressions.

ISI	1 st line	2 nd line	Proportion correct
500.00	33.00	34.00	.6880
500.00	34.00	33.00	.5850
500.00	200.00	206.00	.6710
500.00	206.00	200.00	.5260
500.00	400.00	412.00	.6160
500.00	412.00	400.00	.5620
3000.00	33.00	34.00	.5060
3000.00	34.00	33.00	.6440
3000.00	200.00	206.00	.6320
3000.00	206.00	200.00	.4670
3000.00	400.00	412.00	.6870
3000.00	412.00	400.00	.4410
6000.00	33.00	34.00	.5040
6000.00	34.00	33.00	.6970
6000.00	200.00	206.00	.6310
6000.00	206.00	200.00	.5020
6000.00	400.00	412.00	.7180
6000.00	412.00	400.00	.4230

3. Coefficient data set used for RL determination in experiment 1. Each row features the coefficients from a different regression.

a (intercept)	b1	b2
0.3748	78.5022	-78.8335
5.3690	52.9611	-55.2984
5.3822	48.6103	-50.9874

4. Coefficient data set used for RL determination in experiment 2. Each row features the coefficients from a different regression.

a (intercept)	b1	b2
-0.2755	70.5573	-70.4507
3.2496	40.2508	-41.7672
4.0442	50.7526	-52.6117

Appendix E

Syntax and Equations used in Computer Analyses

1. SPSS syntax used to fit DSW model 12, or model 20 when Fechner's law is presumed to describe the psychophysical function.

IF (stim1 < stim2) pcor = 1 -pcor .

EXECUTE .

DESCRIPTIVES

VARIABLES=pcor /SAVE

/STATISTICS=MEAN STDDEV .

COMPUTE logs1 = LG10(stim1) .

EXECUTE .

COMPUTE logs2 = LG10(stim2) .

EXECUTE .

SORT CASES BY isi .

SPLIT FILE

SEPARATE BY isi .

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT zpcor

/METHOD=ENTER logs1 logs2.

2. SPSS syntax used to determine an RL value based upon the coefficient values obtained from the fit of DSW model 12, or model 20.

COMPUTE bdiff = -(b1 + b2) .

EXECUTE .

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT a

/METHOD=ENTER bdiff.

3. The SigmaPlot regression library equation used to test Link's wave theory model is given by

$$f = (2.71828^{**} (x*a) - 2.71828^{**} -(x*b))/((2.71828^{**} (x*a)) - (2.71828^{**} -(x*a)))$$

fit f to pcor

where x represents the proportional difference between 1st and 2nd lines. a and b are parameters. The dependent variable, pcor, is probability of a correct response in the long-short presentation order.

Appendix F

Using R_{press}^2 for model validation

Stevens (2002) asserts that the PRESS statistic is important when attempting to validate a model based upon a small data set.

The PRESS approach is important in that one has n validations, each based on $(n - 1)$ observations. Thus, each validation is based on essentially the entire sample. This is very important when one does not have large n . . . Let us denote the predicted value for subject i , where the subject was *not* used in developing the prediction equation, by $\hat{y}_{(-i)}$. Then the PRESS residual for each subject is given by

$$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)}$$

and the PRESS sum of squared residuals is given by

$$PRESS = \sum \hat{e}_{(-i)}^2$$

Therefore, one might prefer the model with the smallest PRESS value. The preceding PRESS value can be used to calculate an R^2 -- like statistic that more accurately reflects the generalizability of the model. It is given by

$$R_{Press}^2 = 1 - (PRESS) / \sum (y_i - \bar{y})^2 \text{ (Stevens, 2002, pp. 118-119).}$$

Appendix G

Experiment 1 participant plots of the pair by ISI interaction for T_e

Experiment 2 participant plots of the pair by ISI interaction for T_e

Line Pair (pixels)	
33-34	
200-206	————
400-412	----

¹ The method of constant stimuli is an experimental procedure used in comparative judgment research. It involves standard and comparison stimuli. A standard stimulus is the reference stimulus when difference judgments are to be made. Typically, the standard does not change across experimental trials. In the vast majority of experiments, trial presentation format is such that a standard stimulus precedes the (variable) comparison stimuli.

² The PSE is the physical magnitude of a comparison stimulus that would be subjectively equivalent to a standard stimulus.

³ Stimulus range refers to the range of physical stimulus values, which are presented to experimental participants.

⁴ Inter-stimulus interval is the period of time separating the presentation of two stimuli.

⁵ The primacy effect involves the superior recall of first-presented list items in a memory task. The recency effect consists of superior recall for recently presented list items.

⁶ Space-order error becomes an issue when compared stimuli are simultaneously presented.

⁷ A review of the literature indicates that researchers seldom bother to justify transference of knowledge between absolute and comparative judgment experiments.

⁸ The equipment used by Needham did not allow for precise estimates of sound intensity.

⁹ The absolute indifference interval is the physical standard for which TOE effects are not present. It is relevant to the method of constant stimuli.

There may be more than one absolute indifference interval.

¹⁰ The authors treated judgment category as a ratio variable.

¹¹ This particular model applies to those experimental designs in which the two presentation orders occur equally often.

¹² Link (1992) provides the following logistic response function for the Poisson distribution:

$P_{LG} = 1 / (1 + e^{-\theta^* A})$. This is algebraically equivalent to

$P_{LG} = (e^{\theta^* A} - 1) / (e^{\theta^* A} - e^{-\theta^* A})$. This latter equivalency is typically referenced by Link.

¹³ The author scaled his D/T index of TOE. D represents subjective difference. T is a scalar unique to each participant.

¹⁴ The Central Limit Theorem does not apply to this circumstance.

¹⁵ Experiments involving two response categories require that judgments of the ‘lesser’ and ‘greater’ type be made. Experiments involving three response categories allow for judgments of lesser, equal, or greater.

¹⁶ The possibility that the DSW model handles simple response bias is elaborated upon in later sections of this document.

¹⁷ To find C it is necessary that regression coefficients for the empirical fit of Equation 13 be obtained. These coefficients are obtained by collapsing the data across participants, and performing a separate regression at each ISI. The

coefficients a_0 , a_1 , and a_2 correspond to the intercept and parameters of these regressions. Once these coefficients have been obtained, a_0 is regressed on $(-a_1 - a_2)$, and the intercept of the resultant equation equals C .

¹⁸ This statement pertains exclusively to the mathematical relations expressed within the DSW model.

¹⁹ In the case where s_2 assumes the value of 1, two RLs may exist. Although, the RL associated with the second stimulus will serve no function.

²⁰ Howell (2002) suggested the 2 X arcsine transformation for data that are proportions.

²¹ Line magnitude was entered in pixels. The unit of measurement affected the fitting of the DSW model that incorporated Steven's power law function. The exponent was sensitive to changes in measurement scale.

²² There are additional factors that lead one to favour Fechner's law over the power law. Only regressions involving the former psychophysical function produced significant coefficient tests, and plausible RL values.

²³ The regressions performed at each level of ISI have the same VIF, because they share the same set of independent variables.

²⁴ The R_{press}^2 statistic can assume values less than one. In such cases, the strength

of association is functionally zero.

²⁵ Analyses did not reveal that the RL changed substantially over ISI.

²⁶ It is possible that discriminative sensitivity is the same in both presentation orders, but that accuracy differs.

²⁷ Fechner (1966) asserted that a subjective scale based upon equal JNDs could be built, regardless of the validity of Weber's law.

²⁸ Fechnerian scaling uses the JND as the fundamental scale unit. JNDs are presumed to be subjectively equal.

²⁹ Some individuals may readily dispute the claim that negative TOE prevails in certain sensory dimensions. Admittedly, the generalization is somewhat gross. Subsequent exposition of experimental findings should absolve the statement.

³⁰ Duration discrimination experiments study time perception via the comparison of temporal intervals.

³¹ The distinction between mistreatment and neglect is crucial. The presence of TOE challenges certain psychophysical assumptions. As such, the failure to interpret findings in the context of TOE constitutes mistreatment.

³² $JND(s) = K(s)$, where K is an empirically derived constant unique to a stimulus dimension (e.g., sound intensity, line length) and s = stimulus magnitude.

³³ $JND(s) = K(s) + C$, where C is an additive constant unique to a stimulus dimension.

³⁴ The error is only constant insofar as it is described by one equation. It is proportional to stimulus magnitude.

Table 1

Coefficients for the empirically fit DSW model in experiment 1.

ISI level	Parameter	Parameter value	<i>t</i> statistic	<i>p</i> -value
500 ms	a_0	0.37	.42	.70
500 ms	a_1	78.50	4.48	.02
500 ms	a_2	-78.83	-4.50	.02
3000 ms	a_0	5.37	15.86	.001
3000 ms	a_1	52.96	8.02	.004
3000 ms	a_2	-55.30	-8.37	.004
6000 ms	a_0	5.38	4.46	.02
6000 ms	a_1	48.61	2.07	.13
6000 ms	a_2	-50.99	-2.17	.12

Table 2

Distributional characteristics of s_1/s_2 values at each ISI in experiment 1.

ISI level	Median	Mean	Range	Skewness	<i>SE</i> Skew	Kurtosis	<i>SE</i> Kurtosis
500 ms	0.99	0.70	2.29	-2.06	0.51	2.64	0.99
3000 ms	0.96	0.97	0.22	-0.56	0.51	1.68	0.99
6000 ms	0.98	0.86	2.12	-4.15	0.51	18.00	0.99

Table 3

Coefficients for the empirically fit DSW model in experiment 2.

ISI level	Parameter	Parameter value	t statistic	p -value
500 ms	a_0	-.028	-0.44	.69
500 ms	a_1	70.56	7.02	.01
500 ms	a_2	-70.45	-7.00	.01
3000 ms	a_0	3.25	11.84	.001
3000 ms	a_1	40.25	9.08	.003
3000 ms	a_2	-41.77	-9.42	.003
6000 ms	a_0	4.04	10.61	.002
6000 ms	a_1	50.75	8.24	.004
6000 ms	a_2	-52.61	-8.54	.003

Table 4

Distributional characteristics of s_1/s_2 values at each ISI in experiment 2.

ISI level	Median	Mean	Range	Skewness	<i>SE</i> Skew	Kurtosis	<i>SE</i> Kurtosis
500 ms	1.00	1.00	0.12	0.77	0.54	0.22	1.04
3000 ms	1.00	0.997	0.51	-0.94	0.54	6.05	1.04
6000 ms	1.00	0.96	0.18	-4.15	0.54	-0.11	1.04