

# Extracting magnitude estimations of loudness from pairwise judgments<sup>1</sup>.

Eugene Galanter  
Columbia University

## *Introduction*

A primary goal of psychophysical scaling is to allow the numerical assessment and comparison of perceptual events at super-threshold magnitudes. The successful application of such scaling procedures would let us describe similarities and differences in human experience over a wide range of psychological domains. Further, and perhaps most important, these procedures would serve to connect the constraints on perception imposed by the physical world to the proximal stimulus.

The experimental nature of these judgments of magnitude comprise what I call “numerical introspection,” only marginally distinct from the efforts of Titchener. The distinction rests on demonstrated coherence with behaviorally based experiments (2), (6). When such behavioral procedures are confirmatory, direct scaling methods can be used safely to shorten data collection and provide information across sensory spans that are unavailable to threshold methods.

At least three problems limit the widespread application of numerical introspection: **1)** the operational meaning of these judgments (3). **2)** their reliability when drawn from a single observer (5). **3)** the amalgamation methods used to strike averages from several observers (7). We shall attend to the first and second problem here, and make suggestions about the third.

Aggregated numerical judgments of stimulus magnitudes by groups of individuals are highly reliable regardless of the methods used to obtain the judgments (1, 9). But problems arise when these methods are used to scale stimuli from individual subjects. For example, Green and Luce (5) report psychophysical functions for individuals that reveal cusps and singularities. They make the observers out to be highly idiosyncratic. Indeed, Green (4) remarked, “For practical applications of this (magnitude estimation) technique there is no alternative but to use large numbers of subjects.”

My own view of these data is that the subject tries too hard. A subject may use the same remembered numerical response for a stimulus he believes was previously presented even when the perceptual effect is different. If in the blur of possible numbers that might be attached to a particular stimulus, a value from the tail of the response distribution is chosen, clinging to this response will lead to an oddly placed but consistent estimate of the subjective magnitude of the stimulus. I need hardly mention the “round number” tendency; but this effect may not be quite as common with keyboard responses.

---

<sup>1</sup> The original programming was based on work by Dr. Thomas E. von Wiegand, and revised by Danial Fitousi.

To counter these sources of individual bias, we developed a scaling method that requires subjects to base their judgments of repeated stimulus presentations on a numerical modulus that changes from one trial to the next within an experimental run. This technique defends against the use of the same “nomenclature” even when the same stimulus is presented twice in a row. This is because the comparative judgment is made against a shifting standard. The technique is an experimental realization of what Krantz (7) and Shepard (8) refer to as “relation theory.” The results show that, empirically, the desire to be consistent generates the inconsistency found in procedures first used by Stevens & Galanter (9).

### *Method*

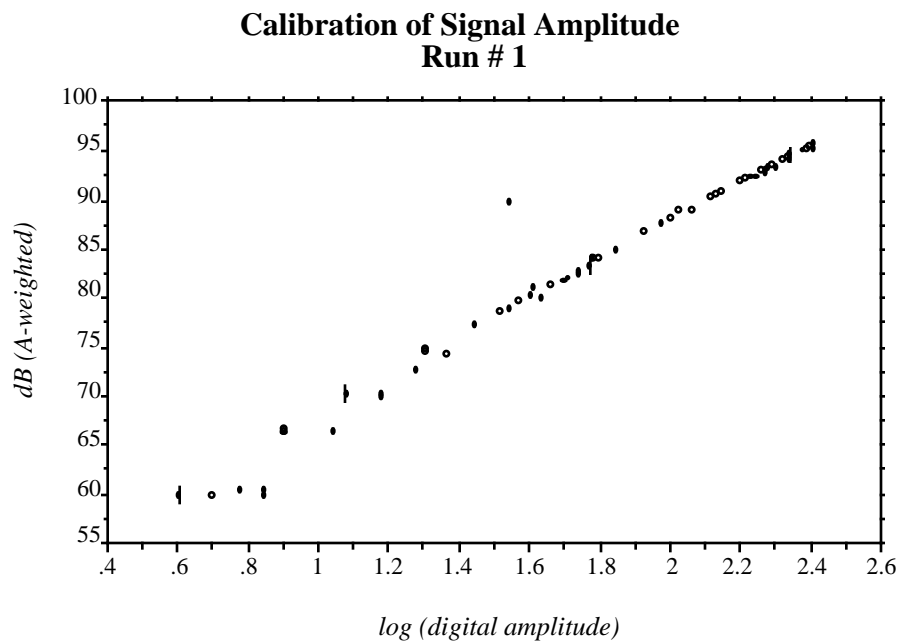
#### *Subjects*

Nine subjects (5 male and 4 female) were visitors and students in the Psychophysics Laboratory. They were invited to participate in a short experiment, and were instructed in the safeguards of the Columbia University Human Subject’s Review Procedures.

#### *Apparatus*

The subjects were seated at a console consisting of a CRT and keyboard connected to a microcomputer. The computer was in turn connected to auxiliary equipment arranged to present auditory tones of 1000 Hz through an array of speakers at signal levels at the subject’s ears of ca. 55 to 95 dB SPL. The acoustic system was calibrated daily, and every experimental run was monitored remotely.

Figure 1 shows a calibration curve. Among these 450-500 records, several outliers are most likely keyboard response errors. The slope is close to 1. Because the tone amplitudes are generated digitally, their analogue representations in deciBels at low levels are somewhat discrete as can also be seen.



*Figure 1*

## *Procedure*

After some initial activity by the experimenter to enter the name of the data file and initialize the system, the subject (alone now) was shown by the machine how to make magnitude estimations by displaying pairs of lines of different (or occasionally identical) length. The subject was tested from time to time by requiring him or her to enter a number estimating the length of one of the lines relative to an arbitrary number assigned by the machine to the other<sup>2</sup>.

After eight minutes of line length estimation, the subject heard sounds of varying amplitude and was asked to estimate how loud the second tone was relative to the first which was assigned an arbitrary numerical value just as the lines had been. After two minutes of training on this task, thirty numerical introspections were obtained from five of the nine subjects. Four subjects agreed to participate in a slightly longer experiment that yielded 100 judgments, the first thirty of which were used in some of the analyses.

The computer program selected a standard and comparison tone more or less at random for each trial. The selection was constrained to avoid too many fractional responses. Various numerical tags, sometimes called “moduli,” were assigned arbitrarily to the standard tone. Each modulus was limited to integral values that ranged from 1 to 800. If the modulus was larger than one digit it was rounded to zero in the least significant digit. The subject keyed in his numerical response to the comparison tone<sup>3</sup>.

## *Results*

We generate 256 digital amplitude levels, but analog output changes only by steps of four. The computer quasi-randomly generated two tones from these 64 levels for successive presentations on each trial. The first tone was the “standard” stimulus from a set designated  $S_j$ . The second tone, the “comparison,” was a nominal stimulus from the same set of  $i = j = 64$ , designated  $C_i$ .

The machine assigned a numerical tag to  $S_j$ , designated  $\psi(S_j)$ , and the subject keyed in a numerical response to represent the comparison stimulus magnitude relative to the modulus assigned to the standard. That response number is designated  $\psi(C_i)$  re  $S_j$  which lets us define an experimental trial as  $R_{i,j}$ , and

$$\psi(R_{i,j}) = \psi(C_i) / \psi(S_j),$$

which is an abstract representation of the instructions.

In some contexts stimuli are often used that also have a physical representation, for example the dB scale of acoustic amplitude which we represent as  $\phi( )$ , filling in the parentheses with the appropriate stimulus name. In the present experiment the stimuli  $\phi(C_i)$  and  $\phi(S_j)$  can be

---

<sup>2</sup> See Appendix 1 for an explanation of these operational assumptions

<sup>3</sup> See Appendix 2 for an explanation of these operational assumptions

represented as

$$\phi_{\text{dB}}(C_i) = 20 \log_{10} (\phi(C_i) / \alpha)$$

where the quantity  $\alpha$  is a parameter unique to the D to A tone generator. When such a physical representation exists, the relation between the  $\psi$  and the  $\phi$  values constitutes the psychophysical law, which we may write as

$$\psi(R_{i,j}) / \psi(R_{k,l}) = \phi(R_{i,j}) / \phi(R_{k,l})$$

which leads to the psychophysical power law.

$$\psi(R_{i,j}) = \phi(R_{i,j})^\beta$$

We can derive the more usual representation in terms of individual stimuli  $C_i$  from this form of the law and their physical metrics  $\phi(C_i)$  by taking the ratios that make up  $\psi(R_{i,j})$  and  $\phi(R_{i,j})$  and rearrange the terms to get

$$\psi(C_i) = \phi(C_i)^\beta \left[ \frac{\psi(S_j)}{\phi(S_j)^\beta} \right]$$

Let the bracketed term equal  $k$  gives the law in its standard form

$$\psi'(C_i) = k \phi(C_i)^\beta$$

The experimental data are the ratios of the computer generated number assigned to the standard stimulus, and the judgments made to the comparison stimulus. The physical stimuli are values of  $\phi(C_i)$  and  $\phi(S_j)$ . This makes the ratio of the stimuli

$$\phi_{\text{dB}}(R_{i,j}) = \phi_{\text{dB}}(C_i) - \phi_{\text{dB}}(S_j)$$

The numerical values of  $\psi(C_i)$  and  $\psi(S_j)$  yield the ratio  $\psi(R_{i,j})$ , which expressed in the same logarithmic framework as the physical stimuli gives  $\psi_{\text{dB}}(R_{i,j})$  as:

$$\psi_{\text{dB}}(R_{i,j}) = 20 \log \psi(R_{i,j}).$$

These numbers,  $\phi_{\text{dB}}(R_{i,j})$  and  $\psi_{\text{dB}}(R_{i,j})$  provide the data points for a representative subject plotted in Figure 2.

“Holt” (ratio Judgments) N = 100

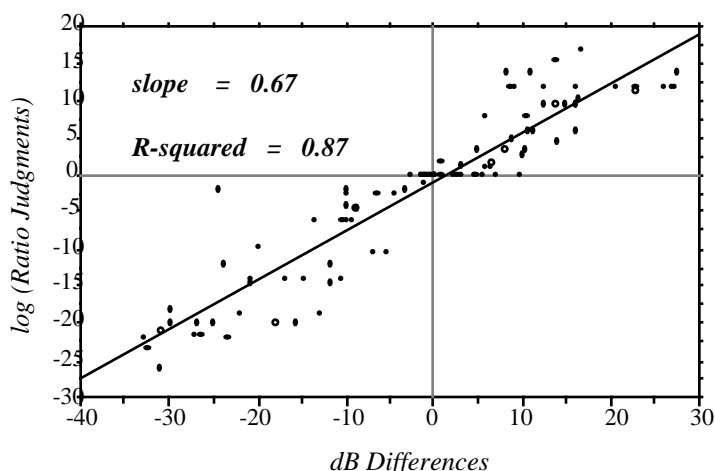


Figure 2

With a slope estimate from these ratios we can construct the psychophysical function in its usual form. The slope parameter of each subject is used to normalize the ratio judgments by transforming the reported values into an equivalent value referenced to the comparison stimulus projected onto the regression line. This equation is

$$\Psi'_{dB}(C_i) = \Psi_{dB}(R_{i,j}) + \beta \phi_{dB}(S_j).$$

The value,  $\Psi'_{dB}(C_i)$  is the logarithmic transformation of the ordinate that is used to

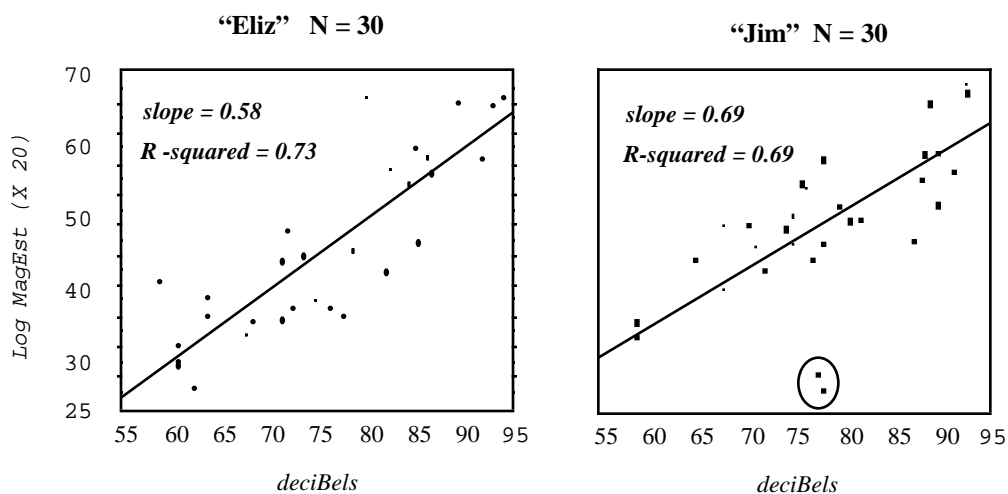


Figure 3

make the axes equivalent. Here it takes the form  $\Psi_{dB}( ) = 20 \log_{10}(\Psi( ))$ . Graphs of transformed data are shown in Figure 3. Circled points are outliers removed from the data analysis. Least squares estimates of the slopes of these functions are reported in Tables 1 and 2.

**Table 1**

Subjects	N	Slope	R <sup>2</sup>
Chas	100*	0.54	0.61
David	100*	0.49	0.78
Holt	100*	0.68	0.76
Pat	100*	0.63	0.81
Eliz	30	0.58	0.73
Susy	30	0.72	0.58
Jim	30	0.69	0.69
Vic	30	0.57	0.66
Ali	30	0.45	0.69
mean		0.594	
S. D.		0.10	

\*Statistics are estimated from the first 30 responses

**Table 2**

Subjects	Slopes	
	1st Half	2nd Half
Chas	.554	.536
David	.523	.466
Holt	.684	.749
Pat	.648	.645
Eliz	.583	.606
Susy	.668	.567
Jim	.690	.706
Vic	.544	.668
Ali	.476	.424

$r_{1,2} = 0.782$   
Standard error = 0.07

### *Discussion*

A new method to construct individual psychophysical functions we call the *shifty modulus* demonstrates that people can report their numerical introspections reliably. The importance of this method resides in the data coherence—“smoothness”—that results from constraints on the temporal depth of judgment. In classical magnitude estimation subjects may carry along some mental representation of the “modulus” throughout the experiment. Our method expects that the span of comparison be at a single memorial ply.

The processing of input on which to base a relational comparison is very restricted temporally, and perhaps generalizing more than these data warrant, may be a bound on any comparison. Judgments based on long term or spatially distant events suggests that we probably see “single stimulus” reports or response bias if a comparison is called for. The relevance of such a view for the practical control of behavior beyond the first ply is clear: We must provide stimulus standards at every point in any control process where we expect accurate response tracking.

## References

- (1) Braida, L. D. and Durlach, N. I. (1972) Intensity perception II: Resolution in one-interval paradigms *Journal of the Acoustical Society of America*, **51**, 483-502.
- (2) Galanter, E and Hochberg, J. (1983) Behavioral indicators of pilot workload. In: *Second Symposium on Aviation Psychology*, Columbus Ohio, 243-252
- (3) Graham, C., and Ratoosh, P. (1962) Notes on some inter-relations of sensory psychology, perception, and behavior. In: *Psychology: A study of a science*, Ed. S. Koch. New York:McGraw-Hill, **4**, 483-514.
- (4) Green, D. M. (1978) Variability in magnitude estimation, *Journal of the Acoustical Society of America*, **63**, 1, S18.
- (5) Green, D., and Luce, R. D. (1974) Variability of magnitude estimates: A timing theory analysis. *Perception and Psychophysics*, **15**, 2, 291-300.
- (6) Kornbrot, D., Donnelly, M., and Galanter, (1981) E. Estimates of utility function parameters from signal detection experiments. *Journal of Experimental Psychology*, **7**, 441-458.
- (7) Krantz, D. (1972) A theory of magnitude estimation and cross-modality matching. *Journal of mathematical psychology*, **9**, 168-199. (cf p. 197).
- (8) Shepard, R. (1981) Psychological relations and psychophysical scales: On the status of “direct” psychophysical measurement, *Journal of Mathematical Psychology*, **24**, 21-57.
- (9) Stevens, S. S. and Galanter, E. (1957) Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, **54**, 377-411.

## Appendix 1

We make several assumptions about a naive subject. These assumptions concern the form of his or her numerical introspections of line length judgments. They are:

- 1) Naive subjects can report the ratio of two line-length segments.
- 2) A subject's judgments will be a power function of line length with an exponent close to one.
- 3) If a subject does not meet these conditions he or she is unsuitable for this task.

In our experience the proportion of college students who fail criterion 3 is less than five percent. The result of these conditions is that a shifty modulus experiment always begins with a training sequence that permits the experimenter to test the initial assumptions. In the current version of the experiment, the subject is also taught to enter judgment integers and rational fractions. Decimal entries are not permitted and error detection and training subroutines are included to explain these procedures to the subject.

## Appendix 2

The shifty modulus design requires that an experimental trial consist of the following sequence of experimental events:

- 1) A warning that a stimulus is to occur.
- 2) A stimulus presentation  $F_1$ , drawn at random from the set of available stimuli.
- 3) A statement of the numerical magnitude of  $F_1$  drawn at random from a defined span of positive integers, called arbitrarily  $Y_1$ .
- 4) A stimulus presentation  $F_2$ , drawn at random from the set of available stimuli.
- 5) A request that the subject supply a rational numerical magnitude ( $Y_2$ ) that makes a judged numerical ratio equal to the perceived stimulus ratio.
- 6) An invisible (to the subject) test of the relation between the numerical ratios.
- 7) Possible error messages, feedback messages, or repetitions of the trial events contingent on well-defined criteria concerning these ratios.

The messages in item 7) provide a vehicle for introducing a payoff function. There are three conditions in which we can be reasonably sure that we understand the subject's internal state. These conditions include judgmental reversals of stimulus magnitude, judgments of equality when (fairly disparate) unequal stimuli are presented, and judgments that reflect a large disparity between response and stimulus relative to previous data. The last "error" may, in fact, reflect keyboard entry error. We can use these three conditions to control the response bias of the subject. This control is achieved by querying the subject contingent on four criteria:

- 1) Query on a random proportion of trials, (usually fewer as experience grows).
- 2) Query on monotone reversals of judgments to stimuli.
- 3) Query when the subject responds with a ratio of 1 to unequal stimuli. The range of stimuli construed as equal can be adjusted to match the subject's experience.
- 4) Query when the subject makes an incorrect entry, e.g., letters for numbers, or decimals for fractions.

Feedback in the form of messages that suggest that the selected response exceeds bounds set by the experimenter for the current stage of experience may also be triggered. These messages serve to keep the subject on track and awake to the task. All data, regardless of the feedback messages that are generated are included in the data set for analysis.