

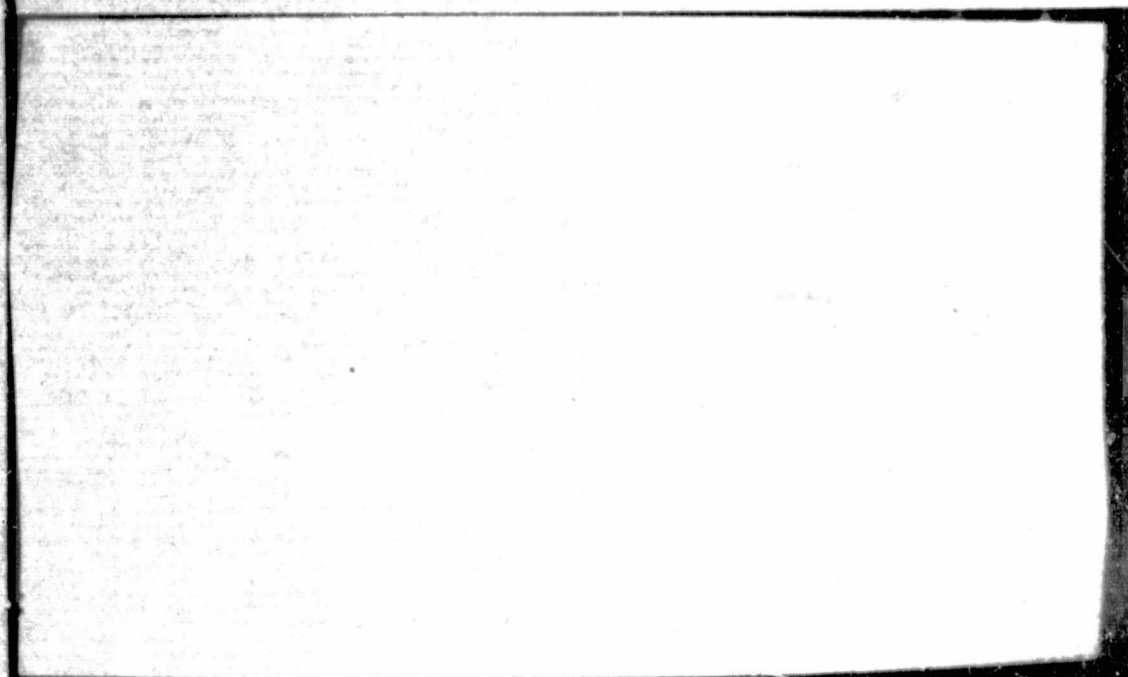
General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.
- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.
- This document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

AD629678

UNIVERSITY OF WASHINGTON

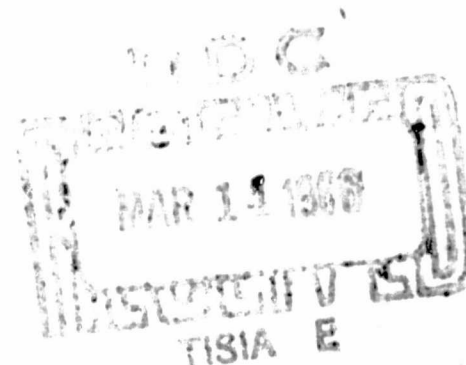


CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION

Hardcopy	Microfilm		
\$ 4.60	\$ 0.50	41	2P as

ARCHIVE COPY

Code 1



This report has been prepared under contract NONR 477(34) between the Office of Naval Research and the University of Washington. Inquiries concerning the contents of this report should be made to the Principal Investigator, Eugene Galanter, University of Washington, Seattle, Washington 98105.

Reproduction in whole or in part is permitted for any purpose of the United States Government.

THIS REPORT SHOULD NOT BE REFERENCED OR QUOTED

X

PSYCHOPHYSICS LABORATORY REPORT

PLR - 14N

Studies of the Constant Error

by

Don A. Ronken and Eugene Galanter

PLR-14N

Abstract

Studies of the Constant Error

by

Don A. Ronken and Eugene Galanter

University of Washington

A series of experiments are reported in which attempts are made to study the role of stimulus variables that may affect the magnitude of the constant error. The results indicate that in addition to bias effects that cannot rightly be ascribed to the stimulus, there are certain forms of potential sequential biases that may or may not be stimulus induced. The classical experiment to investigate the phenomenon is redesigned, and some preliminary results are presented to show that the technique is feasible.

Studies of the Constant Error¹

Don A. Ronken and Eugene Galanter
University of Washington

Experimentation on the constant error, or time-order error, was at the height of its popularity 30 years ago. During that period, the effect was examined as a phenomenon in its own right, rather than a nuisance to be eliminated from experimental data. A review by Needham (1934) summarizes the theories of Fechner, Köhler and Lauenstein, all of which attempt to account for the usual negative CE.

Fechner's theory was that the "memory image" of the first stimulus fades, so that the second stimulus seems larger by comparison. Since Fechner's experiments involved successive presentations of stimuli, it was obvious that the "fading" took place in time, and Fechner was the first to show that the longer the delay between stimuli, the greater was the overestimation. This effect was attributed to the increased fading of the image of the first stimulus.

By the turn of the century, the memory image theory had been discredited by the phenomenological criticism that no conscious image of the stimulus could be demonstrated by introspection. Köhler's theory, proposed in 1923, gave Fechner's concept renewed status by phrasing it in terms of physiological traces for which no conscious correlates were supposed to exist. Köhler's well-known relation is reproduced in Fig. 1 (top) (from Needham, 1935a). A seemingly minor variant of Köhler's theory by Lauenstein gave rise to a flurry of experimentation and criticism regarding the interaction of the physiological traces. Another alternative explanation of the CE, designed to handle the similar bias in absolute judgments also, does not bear the name of any single investigator (Pratt, 1933). This variant

¹Portions of this research were supported by Contract NONR 477(34) between the University of Washington and the Office of Naval Research. This work was conducted during the first author's pre-doctoral traineeship provided by the National Aeronautics and Space Administration.

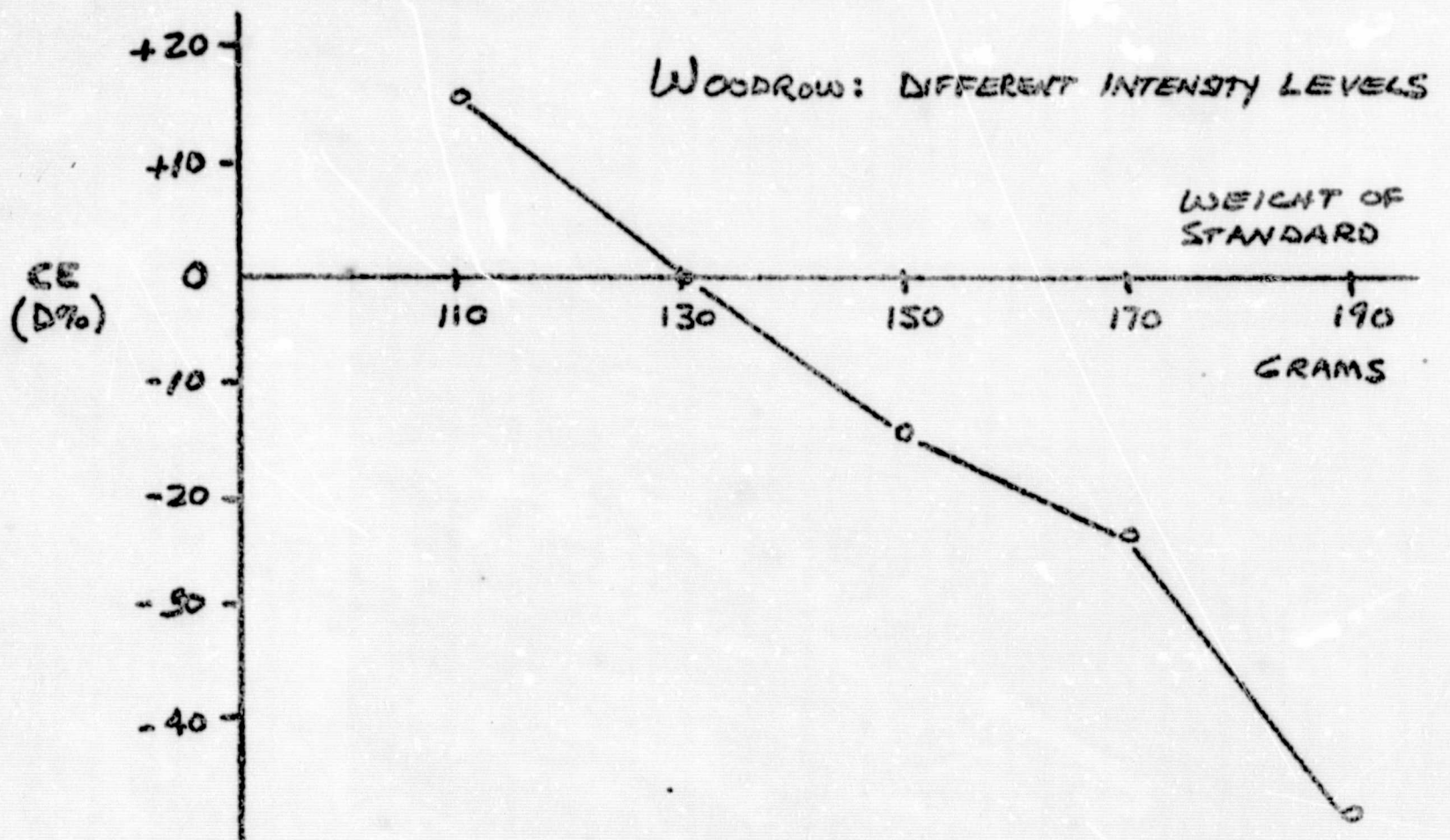
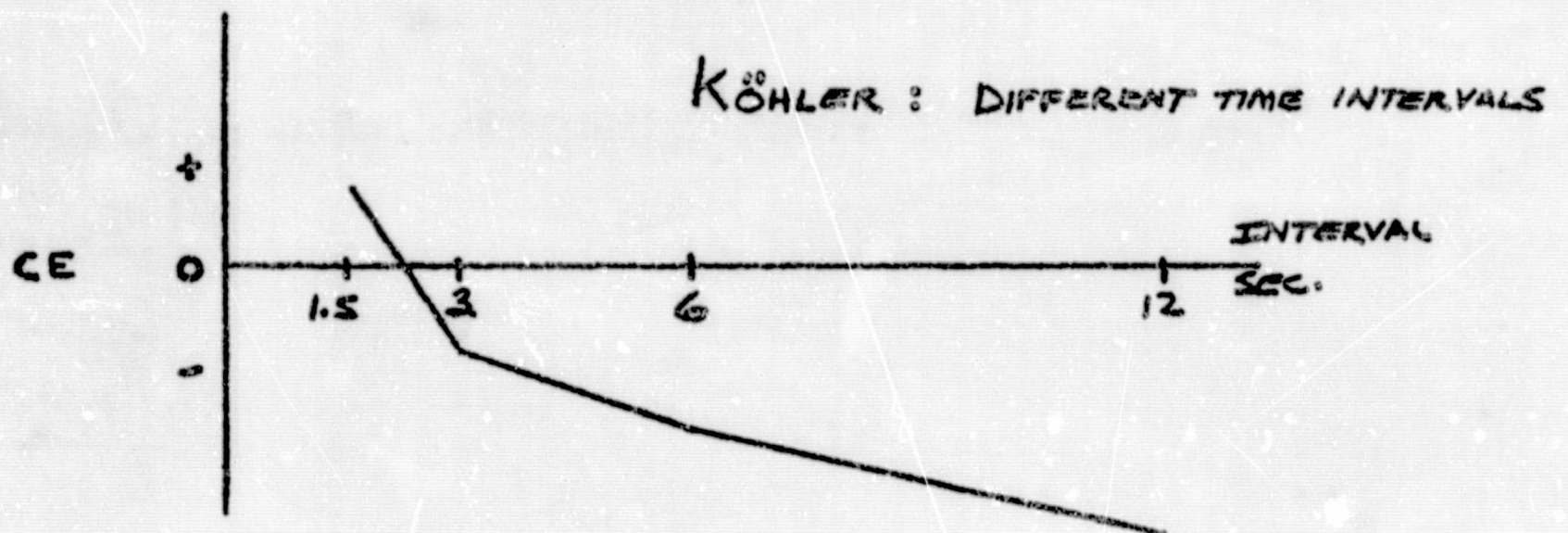


FIGURE 1

assumes the entire series of stimuli establishes an average reference level trace which fades during the experimental session. The theory accounts for the successive comparison data by assuming the series trace is made ineffective when the standard stimulus is repeatedly presented.

The analysis of the CE problem failed, and interest in the problem waned. Modern theory (Luce & Galanter, 1963) implies that the CE is a bias which should be subject to manipulation by response parameters, not, as was thought in the 1930s, by stimulus effects.

Recent experiments performed in this laboratory have demonstrated an inadequacy of early experimental techniques used to investigate the CE. In fact, it appears that most modern procedures are probably also inappropriate. The difficulty is that the early work lacked control over an important aspect of the experimental situation. The majority of these experiments used classical discrimination paradigms in which it is assumed that the subject's response is based upon the relation between the stimuli presented by the experimenter on each trial. Evidence will be presented in this report that under these standard conditions, subjects often respond on an absolute basis rather than a relative basis. Because the 1930 era analysis of the CE presupposed an experimental condition which was probably not attained, it may be worthwhile to repeat the question of that period. What are the variables and relations which determine the constant error?

The early investigators did not emphasize asymptotic performance, indeed, most of the results were based upon observations taken from unpracticed observers, and employed substantially larger stimulus differences than those commonly used today. Much of this early work centered on the determination of the function relating the CE to the temporal interval between the stimuli (Needham, 1934), and upon manipulations which perturbed this function. The term p-function (p for period) has been suggested (Needham, 1935a) as a description of the function relating the CE to the temporal interval.

Experiment 1

The favorite technique used by experimenters to collect discrimination data has been the method of constant stimuli. It was not until a variation

of the technique was used in this laboratory that its inadequacies became apparent. The method was used in its barest essential form: one standard stimulus and two comparison stimuli selected to provide the most information (Bush, 1963). The objective of the experiment was to see whether the CE would follow Weber's Law, an hypothesis proposed by Galanter (1962) and theoretically justified by Luce and Galanter (1963). The stimulus presentations with this procedure assumed the form $(s, s \pm k)$. That is, all the presentations on a given day were of a single standard s followed by one of two comparison stimuli which always differed from the standard by the amount k .

The stimuli were pure tones, differing only in amplitude, and were produced by interrupting an oscillator-attenuator set with a simple cam timer. The tones, presented binaurally through a pair of PDR-8 earphones, were responded to by Ss depression of one of two switches that signified "louder" or "softer." During the course of the experiment, the intensity of the standard was varied on different days over a wide range in 10 decibel steps. The value of k always remained a constant in decibels.

Unhappily, the result of this experiment was less than conclusive. Even neglecting the sign of the CE and considering only its absolute value, the constant error determinations fluctuated irregularly within the margin of experimental error. As a check on this abbreviated form of the method of constant stimuli, the same stimuli were used in another experiment.

Experiment 2

The presentations heard on separate days in Experiment 1 were now combined so that all the intensive levels were encountered on the same day. Thus the possible presentations were:

$$(s_1, s_1 \pm k); (s_2, s_2 \pm k); (s_3, s_3 \pm k); (s_4, s_4 \pm k).$$

Each of these eight pairs were presented equally often in an experimental session. The open circles of Fig. 2 display the CE bias obtained with this procedure. The variable on the ordinate represents the subject's tendency to use the "louder" response alternative as opposed to "softer."

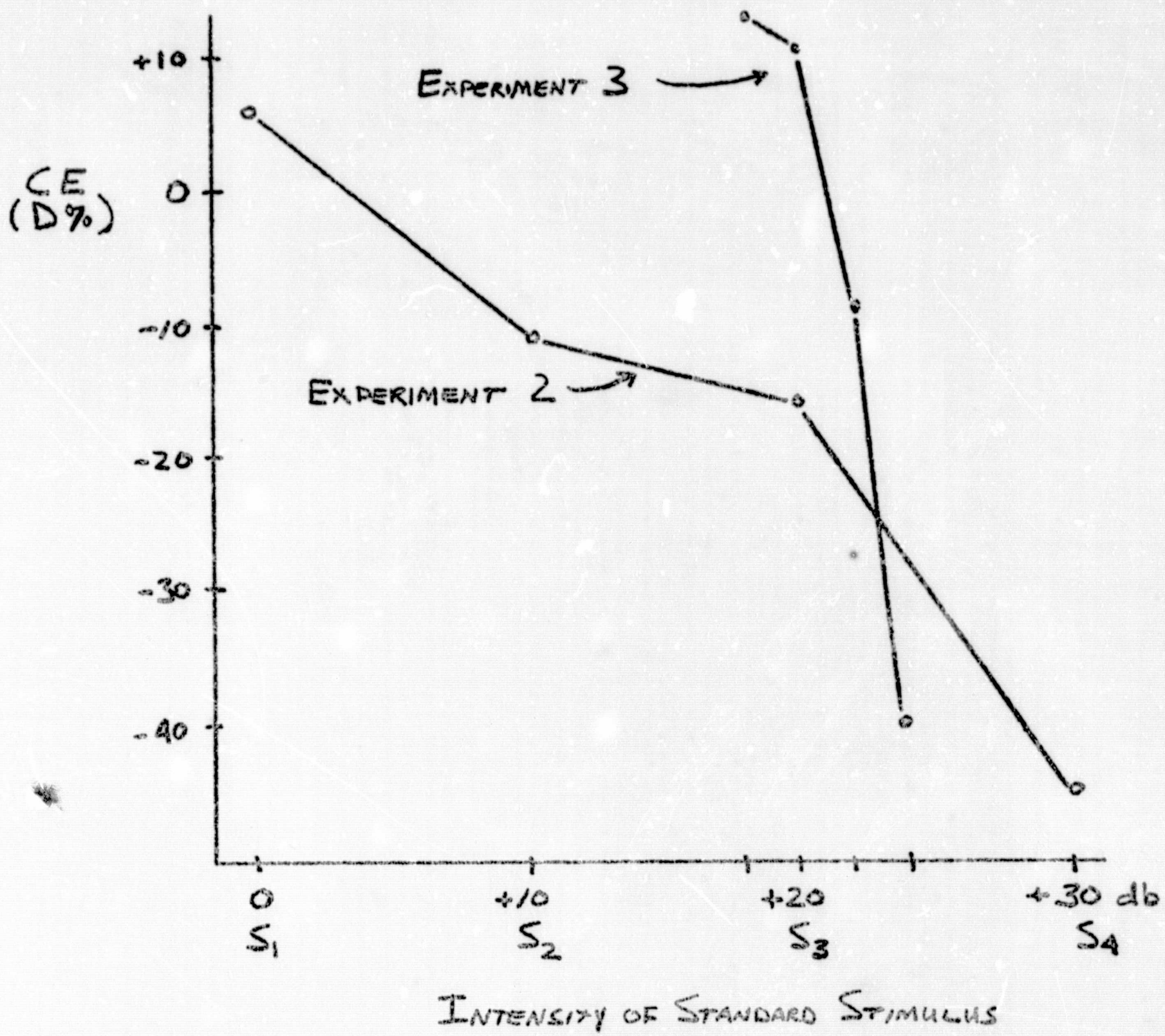


FIGURE 2

The measure of the CE is

$$D = \frac{S-L}{N}$$

where S = the number of softer responses, L = the number of louder and N = the total number of responses. Negative values of D thus represent a preponderance of louder responses, in keeping with the usual convention. This particular measure of the CE is used to facilitate comparison with other experiments which are described later.

At first glance, Fig. 2 suggests that the CE changes from positive to negative in a regular fashion with increases in the intensity of the standard. Closer inspection reveals that this cannot be the result of a Weber's Law effect on the CE. The bias obtained at the more intense standards is enormous by comparison to what is usually obtained. The largest bias represents a case in which the subject responded louder on nearly 75% of the trials, a value which seems unreasonable compared to the usual range for these stimuli of perhaps 50 to 60%. Such a binary response partitioning of 0.75, 0.25 is the classical value required for the determination of the JND, so that these data would imply the CE has an effect equal to the difference limen!

If the results of Experiment 2 cannot be accepted as data representing the CE, what is producing the response bias? Other experiments using the same stimuli in separate sessions failed to produce such results, which would imply that the procedure of presenting all levels of the stimuli seriatim is responsible for the difference.

Fernberger (1920) was perhaps the first to question the validity of assuming that successive responses are independent of one another in psychophysical experiments. Using lifted weights, he found that the particular sequence in which the subject encounters the weights affects the location of the PSE, because the same comparison weights were judged light or heavy, depending upon which comparison weight was used in the previous trial.

Irwin and Preston (1937) demonstrated that subjects tend to avoid repeating judgments of the same category, and that such avoidance is not dependent upon the stimulus. They presented the same weight and the same intensity click on alternate trials for intensity judgments. Although

different verbal responses were given to different sensory modalities, the subject's weight judgments affected the loudness judgments, and vice versa. Irwin and Preston concluded that the response dependence could not be attributed to avoidance of the same words in responding because the dependency was observed between different words (louder - heavier), nor could it be accounted for by the stimulation, since the stimuli were identical from trial to trial. A similar inter-trial dependency holds for absolute visual threshold determinations (Verplanck, Collier & Cotton, 1952). Analysis of successive responses to the same flash of light in this experiment showed that responses separated by as many as 10 trials are significantly dependent.

In Experiment 2 there are undoubtedly response dependencies which are not related to the stimulus, but an inspection of Fig. 2 shows that there is also a large bias which increases as a function of the stimulus. Results very similar to Experiment 2 have been obtained using lifted weights (Woodrow, 1933). By using standard stimuli at varying intensive levels within the same presentation series, Woodrow found that the CE changed from positive to negative as the weight of the standard increased. Fig. 1 (bottom) has been constructed from Woodrow's tables to illustrate the similarity of his results. As in Experiment 2, the bias effect was large enough to equal the effect of one JND at the highest intensity of the standard.

Experiment 3

To establish the exact role of the physical stimulus in producing the atypical results of Experiment 2, the difference in intensity between each of the four standards was reduced from 10 to 2 decibels and the experiment was repeated. Four standard-comparison combinations near the intensity of s_3 were presented, making the total intensive range covered one fifth of that used in Experiment 2, as shown in Fig. 2. If the wide separation of the stimuli was the determining factor of the bias, it was reasoned that reducing the range of stimulation presented ought to reduce the bias correspondingly. The filled data points of Fig. 2 indicate the result; the changes produced in response bias were almost identical to those

obtained with the larger range of stimuli.

A similar invariance of response bias, independent of the range of the stimulus, was obtained by Needham (1935a), who combined the experimental variations of Woodrow (1933) and Köhler (1923) into a single experiment. The CE data obtained by Woodrow and Köhler in their separate experiments are illustrated in Fig. 1. Needham used pure tones at four intensity levels of the standard stimulus, each separated by 10 decibels. Inter-stimulus intervals of 1, 3 and 6 seconds were used at each of the four standard intensities to provide the data reproduced in Fig. 3 (Needham, 1935a).

As anticipated, the response bias was dependent upon the intensity of the standard. What was not expected was the divergence of the curves for the different intensities. The fading trace theory would predict parallel curves, because the p-function of Köhler (Fig. 1) should apply the same bias to all the intensities of the standard. Needham attributed the increased bias at the longer inter-stimulus intervals to the dominance of absolute judgments over relative ones. In a supplementary part of the same experiment (Needham, 1935a), five stimuli one decibel apart were used in a paired-comparison replication of the same experiment. The "standards" under these conditions were then effectively one decibel apart. The same variations of response bias, that of Fig. 3, was found under these conditions, in agreement with Experiment 3.

The results of Experiments 2 and 3, as well as others (Needham, 1935a; Pollack, 1955; Woodrow, 1933) suggest the operation of what have been called contrast effects. Contrast is assumed to be operating when the subject appears not to base his judgment upon the stimulus presented by the experimenter on a single trial, but rather on some sort of impression formed from the entire sequence of stimuli experienced during the session. This phenomenon may be related to the results observed many years ago when the variable stimulus was presented first in the method of constant stimuli. Under these conditions, observers often made their responses before exposure to the standard, suggesting they had another reference available.

Contrast effects are not removed by elimination of the standard stimulus (Needham, 1935b; Pratt, 1933). Observers using the method of

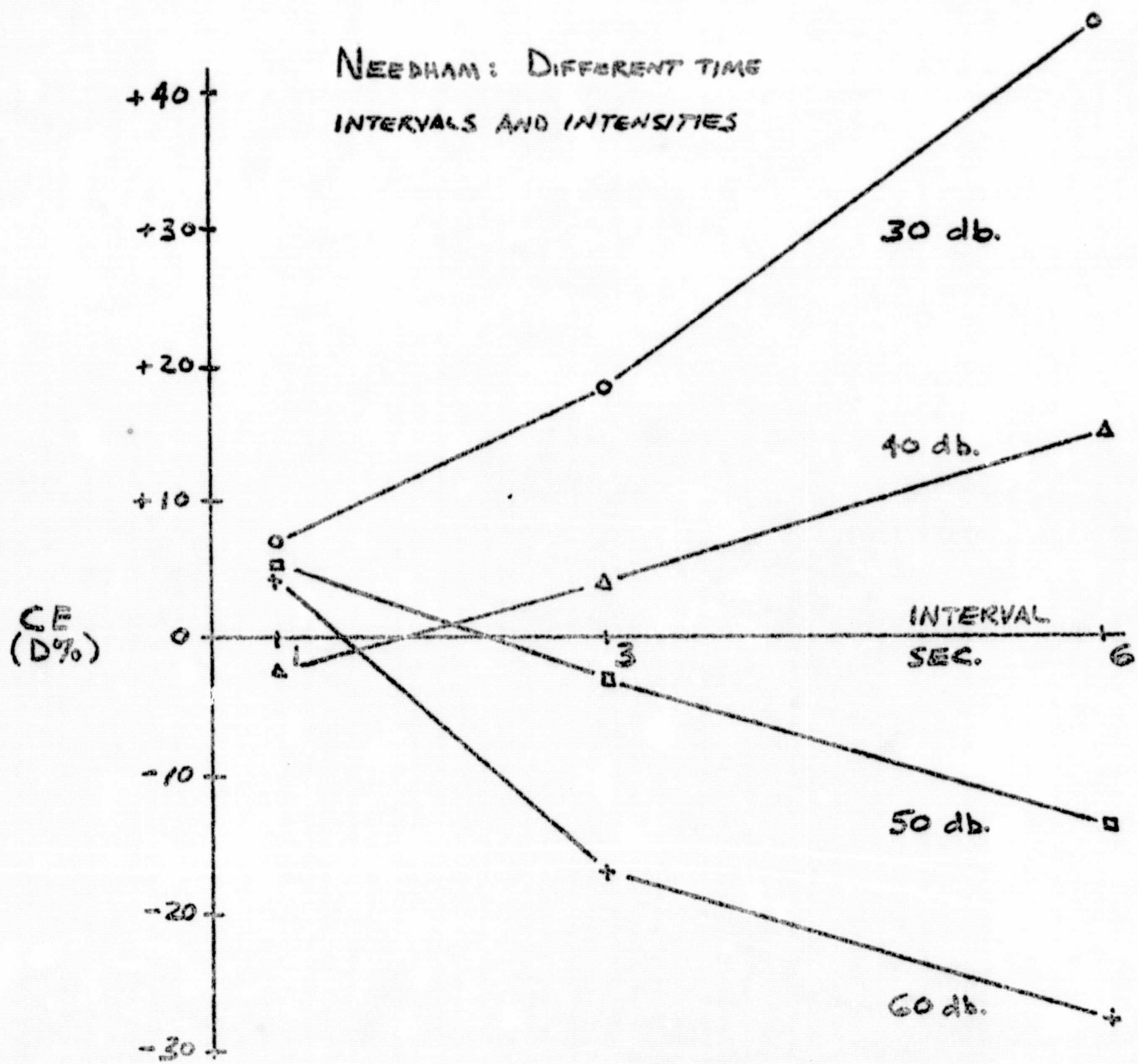


FIGURE 3

single stimuli can make accurate judgments when the stimuli are only one decibel apart, but adding one more stimulus to the extreme of the stimulus series, without the subject's knowledge, induces a large context shift in the responses.

The theory developed by Helson (Helson, 1954; Michaels & Helson, 1954) is an attempt to account for phenomenon such as the contrast effect. These authors, using the concept of an adaptation level, conclude that "...the PSE is found to be a weighted mean of all stimuli, past or present, which affect the judgment." This curve-fitting procedure apparently can predict "averages of the average judgments of 12 Os," but this power is of marginal interest compared to theories which account for a wide variety of data with a few potentially meaningful parameters (Luce and Galanter, 1963).

Evidently, relative judgments unconfounded by absolute or contextual bias are rarely obtained. Furthermore, it is obvious that theories of the CE which assumed a recent memory mechanism for the first stimulus must have been derived from and evaluated against confounded data. Clearly the experiment must include more constraint.

Experiment 4

Although our suspicions have been raised, we have, so far, encountered response biases that may have been induced by the relative difference between stimulus pairs. Earlier experiments, as we have noted, demonstrated biases of a similar kind when sets of stimuli were presented singly. If some or all of this bias is independent of stimulus differences, one may be able to observe it when only a single stimulus is used. If we were convinced of that point, we would then be in a position to design an experiment in which relative judgment alone could occur. To examine this question we conducted an experiment in which subjects were asked to heft weights and make absolute judgments about them. Although the subject believed that he was lifting a variety of different physical weights, the experimenter presented only a single weight to him on each trial.

The conduct of this series of experiments followed a similar form. The subject, in this case a University of Pennsylvania graduate student,

was led into an experimental chamber containing two chairs and a table. On the table were arrayed a series of 14 weights. The subject was instructed along the following lines, "You are going to make judgments about how heavy these weights feel. This will be done on an equal interval 7 point (in other experiments 3 point) scale. This weight, the first that you will judge, will be called 1. (The subject was then allowed to heft a weight of 100 grams.) The second weight you will judge will be called 7. (The subject was then allowed to heft a weight of 200 grams.) From then on you will lift each weight when I say lift, judge it, and give it a number between 1 and 7 that represents its relative heaviness. Now put on this blindfold. Once again here is the first weight. Remember it is called 1. (Present the 100 gram weight). Now here is the second weight, remember it is called 7. (Present the 200 gram weight.)"

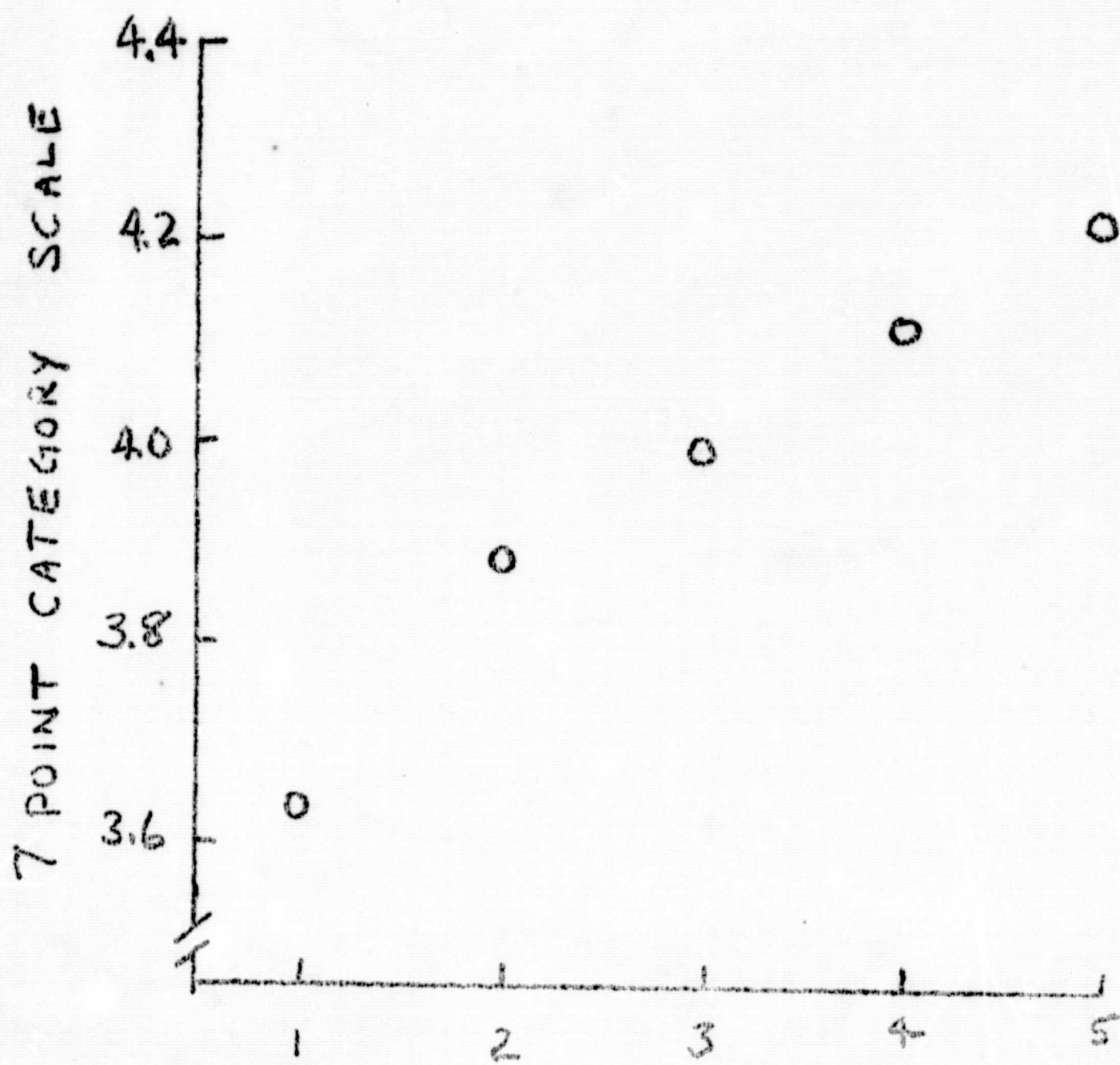
From this point onward each of ten subjects were given the 100 gram weight to heft for the remaining trials. Figure 4 shows a subject by blocks of trials average of the numerical values assigned to this weight.

Variations on the experiment were conducted. Some involved presentations of the heavy weight first followed by the light weight followed by hefting the light weight, and other variations that are described in Fig. 5. In addition, the judgment scale in the later experiments was reduced from 7 to 3 points.

The results of these variations were uniform regardless of the numerical scale that was used. As Figure 5 shows, only one condition, that in which the light weight is presented first, the heavy weight second, and light weights are presented from that point forward, gave rise to the increase in judgment of the magnitude of the weight being hefted.

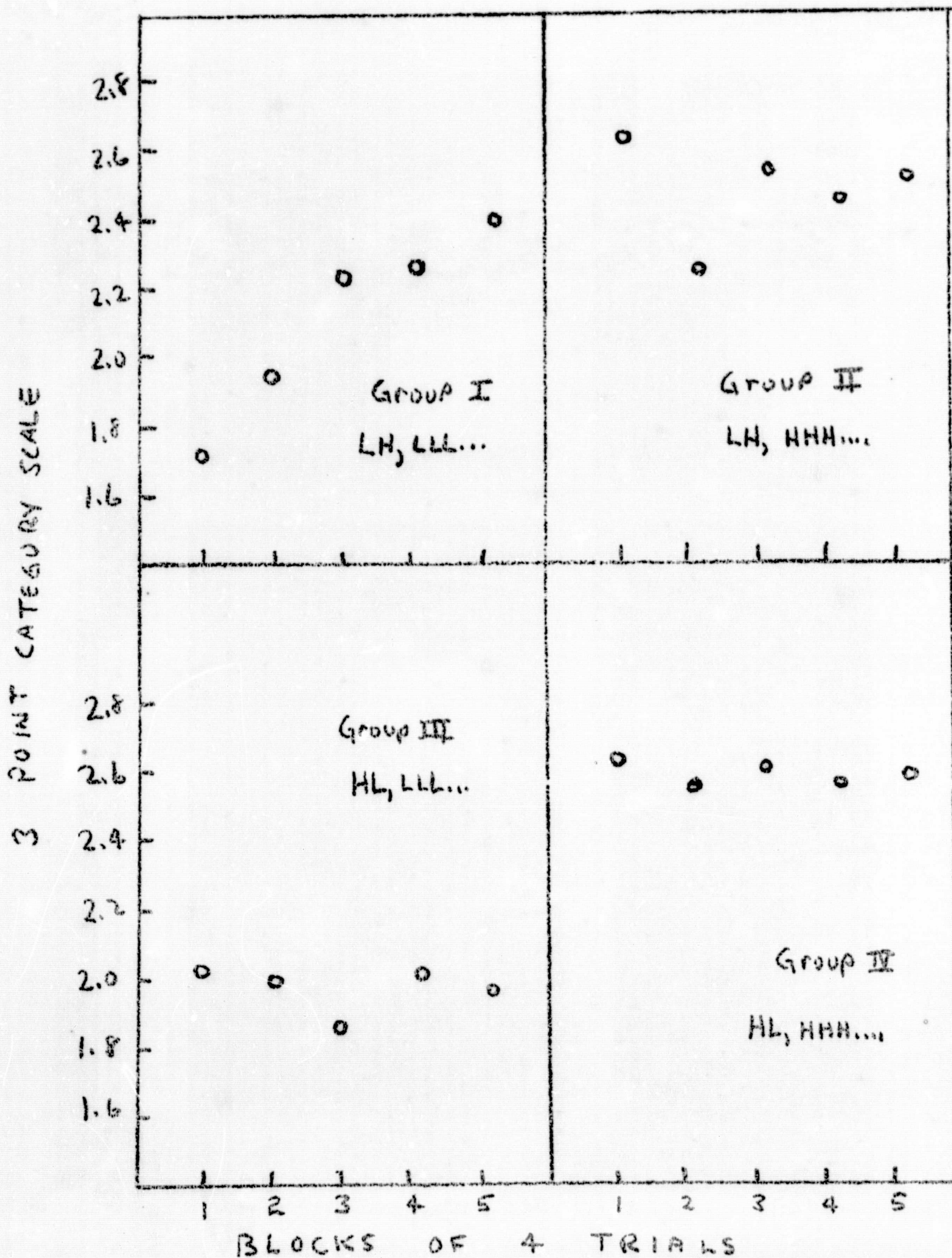
The results are fairly straightforward. Under the first procedure, that is, the procedure in which there is room to climb the scale and in which a marked contrast exists between the second weight to be judged and the remainder of the series, there is strong tendency for the subject to accumulate his bias and to report more and more frequently that the weight feels heavier and heavier. The change in the appearance of the weight as reported by the subject extends, at least in the case of the three point scale, over one-third the range of the judgments available to the subject. The time course of the change suggests that the bias grows in magnitude

Fig. 4



Blocks of Two Trials
10 S's / point weights presented
L, H, LLL...

Fig. 5



from the early trials. There is, of course, no evidence that the bias effect extends back more than a single trial.

Just as the bias can be eliminated by altering the sequence to be judged and the relative order of presentations of the standards, it can also be eliminated by reintroducing both standards before each trial. When this was done in one experiment in which the standards were presented in the order that induced the bias originally, the bias disappeared.

These results simply reinforce some of the conclusions of our earlier experiments, but they do suggest that the biases may be eliminated. Let us turn then to the description of the design for a discrimination experiment that constrains the subject to judge the relation between a pair of stimuli on each trial independent of the stimulus pairs that have been presented earlier. In what follows we shall describe the idea of the experiment and some of the preliminary results. These results lead to the need for an improvement over our instrumentation which is currently in progress.

Experiment 5

Consider the case of the usual method of constant stimuli where the two comparison stimuli closest to the standard are separated from the standard by a small amount, say one half decibel. Under these conditions, the comparison stimuli themselves are separated by a full decibel, or more, a value which has been shown to be adequate for discrimination under the single stimulus paradigm (Needham, 1935b). The subject seems to be able to perform two ways: by using his own noisy stored reference to test alternatives separated by one decibel or by using a less noisy reference to test alternatives separated by one half decibel.

For true relative judgments, the desired situation is one where the only information available to the subject is conveyed by the relation between the stimuli to be judged, not from their absolute values. Consider now a situation where the stimuli to be used are available in discrete energy values: 1, 2, 3, ... as indicated in Fig. 6. Represent the pair presented, (s_1, s_2) , as a point on the grid by locating the first stimulus along the abscissa and the second along the ordinate. On this plot, pairs

which are above the line $x = y$ are those values for which $s_1 < s_2$.

Now, if the discrete values of stimuli are selected so that the intervals between them are appropriate, i.e., on the order of $\frac{1}{2}$ decibel, it is possible to select sets of pairs such that the effect of an absolute judgment will result only in a chance level performance. One such possible set is indicated in Figure 6. If each of these pairs are presented in a random sequence with equal probability, then the pairs (3,4) and (4,3) can reasonably be said to be controlled for absolute judgment effects. Under these conditions, if a subject were to identify the second stimulus correctly as being a value of 3, and classify this stimulus as "softer" on an absolute basis, his response would only be correct half the time in the long run, for the pairs (2,3) and (4,3) occur with equal probability. The same inadequacy of absolute judgments applies to signal values 3 and 4 in both the first and the second interval. Under these conditions, the subject ought to be constrained to store a representation of the first signal in order that it will be available when the second is presented. Given this arrangement, two questions arise; first will a bias occur? Second: if such a bias occurs how will it depend upon the temporal interval during which the subject must store the first signal.

With this new experimental paradigm the subject's response set could be the classic binary pair "louder-softer," but recent experimentation has shown (Egan, Schulman & Greenberg, 1959) that subjects can partition their responses into more than two alternatives. This expanded design is based upon the observer's ability to maintain several criteria simultaneously. The experimenter infers the existence of separate criteria from the subject's use of confidence rating responses.

There are two principal advantages to using a rating method as opposed to a binary choice technique. First, a considerable economy is realized in the number of observations needed to estimate response probabilities, because the subject has the potential to convey up to three bits of information to the experimenter per trial instead of the usual one bit (for the case of 8 response alternatives as opposed to 2). Second, the use of multiple criteria allows constant monitoring of one of the fundamental assumptions of signal detection theory: the assumption of equal variance of the two distributions upon which the subject is presumed to be basing his responses.

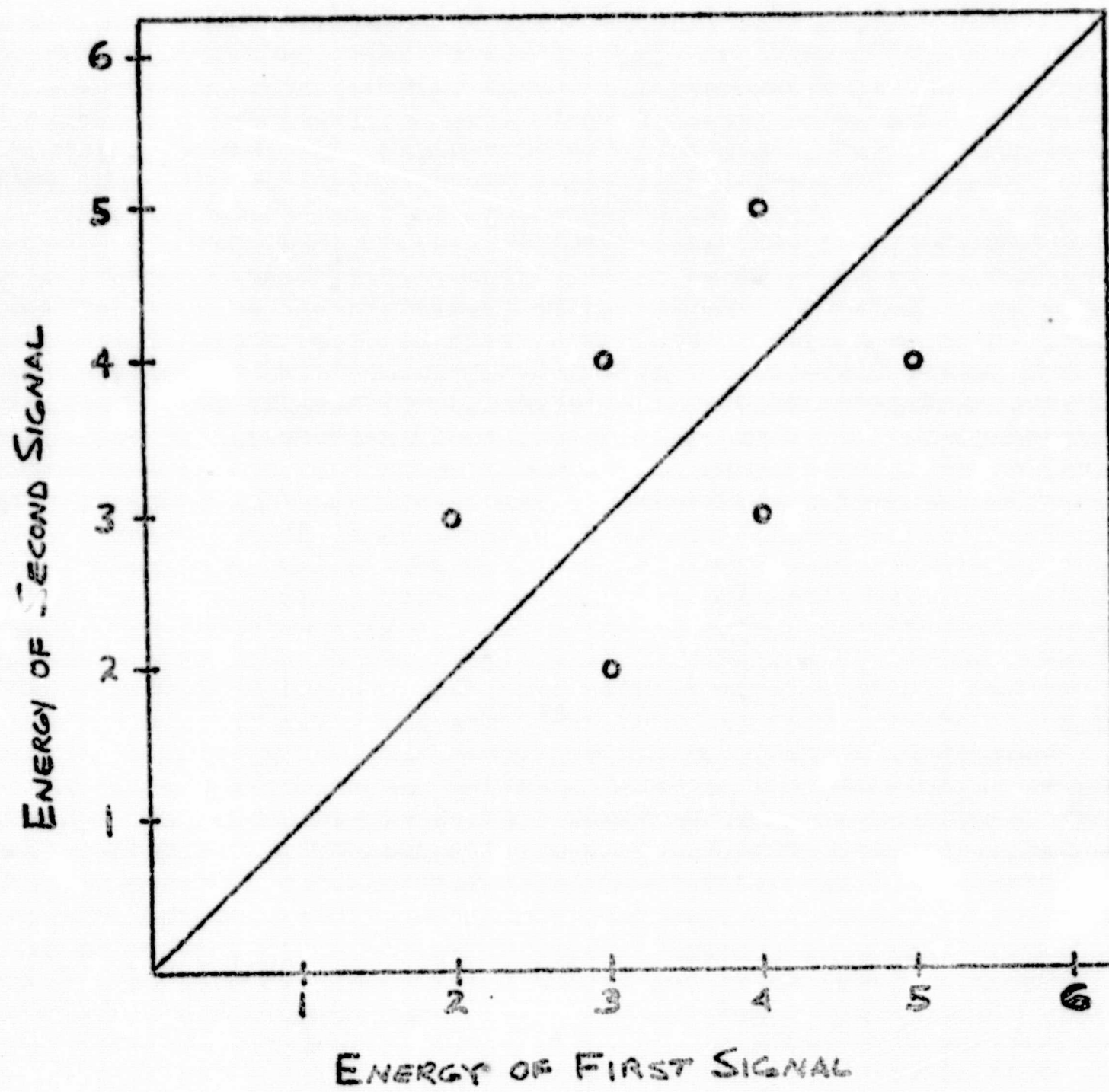


FIGURE 6

Except in those psychophysical experiments where an outcome structure is inappropriate, (Bush, Galanter & Luce, 1963) immediate feedback to the subject following a response is often used. The feedback usually indicates the correctness of the response, and according to signal detection theory, is used by the subject in acquiring knowledge of the signal distributions so that he may establish his response criteria (Luce & Galanter, 1963).

Another large body of experimental work, that connected with mathematical learning models, has indicated how powerful information feedback can be in influencing response proportions. The potential of such effects upon psychophysical tasks has been convincingly shown by Friedman and Carterette (1964). Their experiments show large detection differences for stimulus sequences having different trial-to-trial transition probabilities, even when the overall probability of a signal is held fixed. That is, subjects can effectively use even information which is only probabilistically dependent upon the events of the previous trial. Some learning models for psychophysics which have touched upon the problem of information feedback (Atkinson, Carterette & Kinchla, 1962; Atkinson, Carterette & Kinchla, 1964) demonstrate how different models will probably be required to describe the data obtained with and without feedback.

Previous unpublished experimentation in this laboratory has established that the size of the CE is reduced when feedback is used, presumably by training out the bias of interest. It is an interesting and currently unresolved question whether the CE can be reduced to zero in this fashion. For this reason, and because of its complicating influence through learning mechanisms, immediate feedback was not used in the following experiments. However, for training purposes in the use of confidence ratings, feedback was presented after blocks of responses as described in the procedure section. This method has already been described as being effective in the use of ratings (Clarke, 1960; Pollack & Dicker, 1958).

Apparatus

The signal source was a Hewlett-Packard 200 AB Oscillator which operated at 1 kc. and a fixed gain throughout the experiments. Different signal levels were obtained by switching to any one of 8 continuously adjustable T-pad attenuators. The output of the attenuators was calibrated relative to the output of a standard precision attenuator by the use of a sensitive Fluke voltmeter as a null instrument. From the attenuators, the signal was led to a pair of PDR8 receivers.

The stimulus and temporal intervals used on each trial were determined by a perforated tape reader which advanced during the inter-trial interval. Signal duration was held constant at 250 msec. and was obtained by activating a zero-crossing switch, which followed the oscillator. The signal was amplified by a stable-gain Dynakit amplifier before passing to the T-pad attenuators. The paper tape could be programmed to select on any trial one of 8 signal levels in the first interval, any of 4 inter-stimulus intervals, any of 8 signals in the second interval and any of 4 inter-trial intervals. This information plus one of 8 responses by the subject was recorded on a digital printer.

A signal light warned the subject of signals which were to follow. The minimal length of inter-trial interval varied from 2.4 to 5 seconds as determined by the program, but the subject was permitted to take as long as he wanted to respond, so the effective length varied.

Procedure

The subject was first given practice in discrimination using a binary louder-softer response until stimulus pairs to be used in the experiment were correctly discriminated about 75% of the time. The instructions were to judge the second signal as louder or softer than the first. Following this training, the subject was introduced to the confidence rating method, which required the use of 8 push button type microswitches mounted in a horizontal row. The buttons were arranged into two groups of 4, half for louder responses and half for softer. The buttons were assigned the designations L4, L3, L2, L1; S1, S2, S3, S4, where the numbers correspond to the 4 criteria which were used.

After experience with the binary choice situation, the subject agreed that upon some occasions of responding he felt more certain of the correctness of his response than at other times. This distribution of confidence was to be the basis on which the subject made his rating responses, and was to provide the data for determination of the separate criteria used.

The subject was instructed that when he felt very certain of his decision, he should push the "4" buttons; when very uncertain he ought to use the "1" buttons, and so on. To insure stability of the criteria, the subjects use of the categories was correlated with his proportion of correct responses, in the following way. The 4 categories were assigned arbitrary values with which they were to be used correctly. Category 1 was to be used when the response (either softer or louder) was correct 50-65% of the time. That is, on those occasions when button 1 was pushed the subject was to be correct 50-65% of the time. For category 2, the values were 65-80%, category 3, 80-90% and category 4, 90-100%. Evaluations of the use of the categories were given verbally by the experimenter at the end of blocks of 100 trials, until category usage was stabilized in the correct range. For instance, if at the end of a block of 100 trials, the subject was correct 80% of the time when he used category 1, he was informed that he was using the categories too conservatively. As the performance more nearly approximated the specifications, the feedback was made more specific to the category boundaries. The subjects developed this skill surprisingly quickly, and they did not feel that the task was particularly difficult. At least once during every session, the data were examined to insure the correct distribution of responses.

The actual partitioning of the category responses is of course arbitrary, a sufficient condition is that an ordinal arrangement of confidence is made (Swets, Tanner & Birdsall, 1961). In the experiments to follow, the determination of the CE was based upon the reduced binary data, and the role of the category responses is only to aid in the construction of iso-sensitivity curves. For this reason, results of the experiment are not particularly sensitive to moderate shifts in use of the categories. The reason for including confidence judgments is to obtain a detailed analysis of the discrimination process with as few observations as possible.

The stimulus pairs used to evaluate the discrimination were selected as shown in Fig. 6. In addition to these pairs, which occurred equally often, the first four trials in every block of 100 and 10% of the following trials were actually practice pairs. These were pairs for which the voltage of the difference signal was increased 2, 4 or 8 times. These "easy" trials seem to be required for a subject to retain the ability to discriminate under these no-feedback conditions, and were not included in evaluation of the data. The classical case of the method of constant stimuli was also used without feedback and included such supra-liminal pairs.

Unlike the classical experiments however, Experiment 5 used a payoff function to determine the wages of the subject. The use of payoffs when outcomes for each response are not included is somewhat irregular according to the classification of choice experiments by Bush, Galanter and Luce (1963). This unusual situation is the result of combining a partial identification design and a form of category scaling experiment.

In the partial identification part of the experiment, the possible responses are "louder" and "softer." This binary pair could be assigned an outcome function because the physical relations between the stimuli, $(s_1 < s_2)$ and $(s_1 > s_2)$ are unambiguous criteria for deciding when a response is correct. As mentioned earlier, we choose not to use immediate feedback so that no outcomes are actually presented. The subject is instructed in the existence of the identification function for the louder-softer responses, as is ordinarily done in discrimination experiments. He is made to understand that each of his binary responses is unequivocally correct or incorrect, and that his wages are determined by a symmetric payoff matrix based upon this relation, even though the apparatus does not reveal the outcome of each specific response.

On the other hand, for the confidence rating responses which have possible values 1, 2, 3 or 4, no identification function exists. Consequently neither an outcome function nor a payoff function could be used with these responses. However, at one point in the experiment, it may have appeared that an identification function was assumed. During the training period, when the subject was learning to distribute his use of the categories over the "percentage correct" scale, the experimenter provided delayed verbal feedback after blocks of trials. This procedure does not actually use an

identification function, for the experimenter had no way of deciding when any single response was correct. It was only by examining the distributions of large numbers of rating responses that the feedback could be constructed. At no time was a payoff associated with the rating responses.

Data Analysis

the
The data obtained by/rating procedure may be represented in a 2×8 celled table, as table 1 indicates. The large table collapses to the 2×2 binary table if only the louder-softer criterion is considered. The estimates for the conditional response probabilities are obtained from the 2×8 table following a procedure which is justified in the Appendix. Using the designation L_i for the criteria, $i = 4, 3, 2, 1$, we calculate

$$P(L|s_1 < s_2) = \sum_{i=4}^C P(L_i|s_1 < s_2)$$

and

$$P(L|s_1 > s_2) = \sum_{i=4}^C P(L_i|s_1 > s_2)$$

These values are the probabilities that the response exceeded the criterion c . The sum begins at 4 and goes to 1 since the "4" on the subject's response panel indicated a strict criterion, i.e., when he was to be correct 90-100% of the time.

The four pairs of values obtained from this procedure generate the operating characteristic. Unlike the binary choice case, these rating data are not completely specified by giving only the louder response probabilities. There are also four pairs of softer conditional probabilities which are in general different, but these differences appear small and display no significant pattern, so that these data will not be examined here.

The data are most conveniently displayed on the unit square, where the ordinate corresponds to correct discriminations and the abscissa gives the false alarm rate. A typical experiment, the data from table 1, is displayed in Fig. 7. The smooth curve faired through the four data points, which represent the response proportions generated by the subject adopting

TABLE 1

STIMULUS PRESENTATION		RESPONSE							
		LOUDER				SOFTER			
		L4	L3	L2	L1	S1	S2	S3	S4
$S_1 < S_2$		20	19	36	10	3	2	4	0
$S_1 > S_2$		2	5	16	8	12	33	21	3
		22	24	52	18	15	41	25	3
									200

RATING
DATA

	L	S	
$S_1 < S_2$	85	15	100
$S_1 > S_2$	31	69	100
	116	89	200

COLLAPSED
BINARY
FORM

CRITERION EXCEEDED

	c4	c3	c2	c1
$P(L/S_1 < S_2)$	.20	.39	.75	.85
$P(L/S_1 > S_2)$	.02	.07	.23	.31

CONDITIONAL
PROBABILITY
ESTIMATES

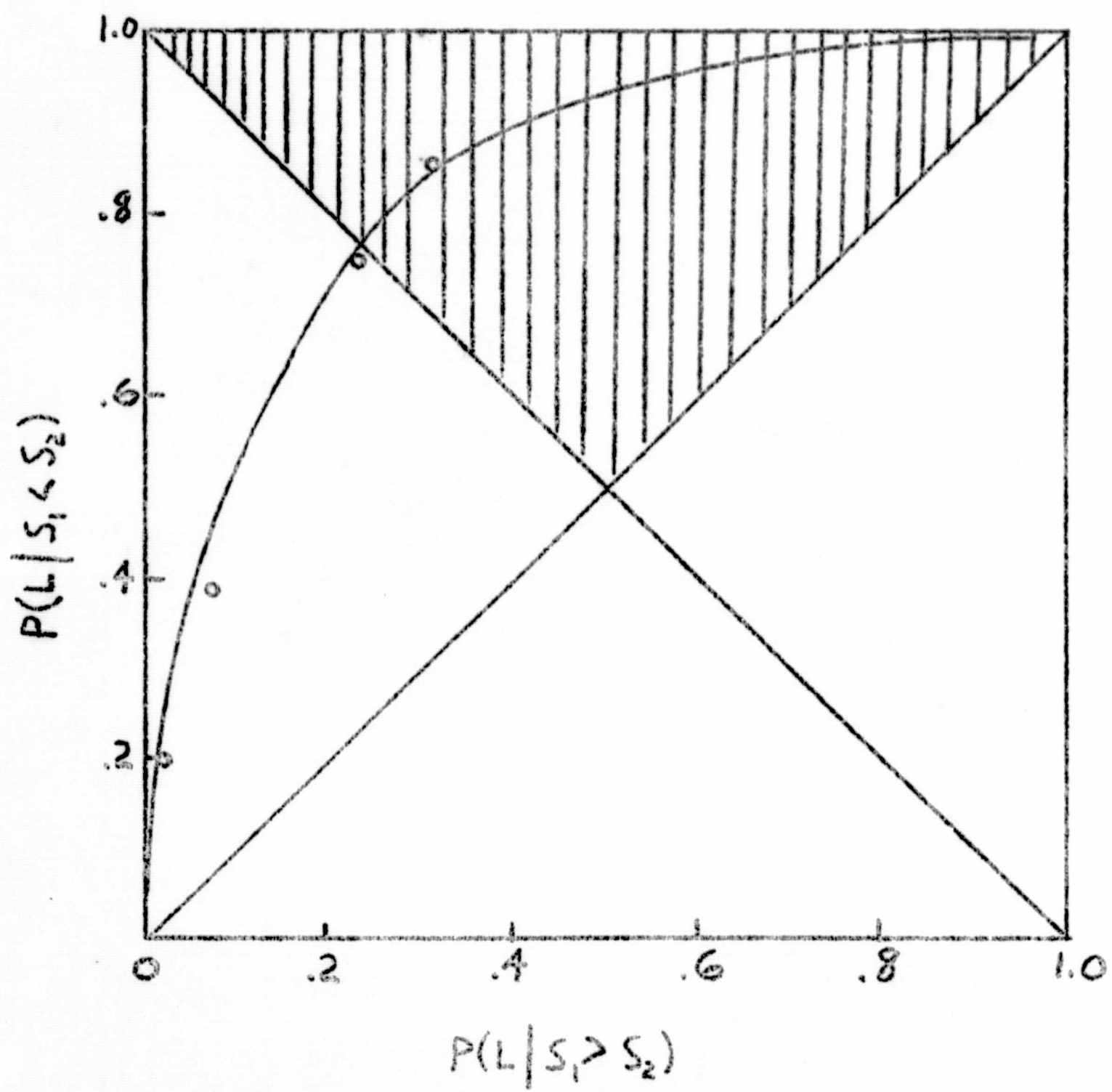


Figure 7

successively less strict criterion, is the collection of all data points which could be obtained when the energy of the difference signal is held constant. Displacement of individual points along this curve represent different criterion points. This is perhaps clearest in the case where the energy of the difference signal is zero. Then all of the obtained points ought to fall on the diagonal running from lower left to upper right, and the particular location of any point is only a reflection of how often the subject says "louder" relative to "softer". Classical results of the CE experiment would fall within the shaded area, for this is the region of a negative time error, where the subject supposedly underestimates the value of the first stimulus. The diagonal from upper left to lower right is the location of all points where there is no bias for one response, i.e., where the CE is zero. The displacement of the data from this diagonal is a dimensionless measure of the CE.

The data of Experiments 2 and 3 were actually obtained with the confidence rating procedure just described. The following detailed description of Experiment 3 will illustrate the power of the rating method and demonstrate the appearance of context effects in the method of constant stimuli.

There were 8 possible stimulus configurations presented to the subject with equal probability:

$$(s_1, s_1 \pm k); (s_2, s_2 \pm k); (s_3, s_3 \pm k); (s_4, s_4 \pm k)$$

The difference between adjacent standards was two decibels, while k was constant at 0.4 decibels, so that the entire stimulus set can be represented ~~on the energy scale as~~ *schematically as*

$$\begin{array}{cccccccccccc} s_1^{-k} & \hat{s}_1 & s_1^{+k} & s_2^{-k} & \hat{s}_2 & s_2^{+k} & s_3^{-k} & \hat{s}_3 & s_3^{+k} & s_4^{-k} & \hat{s}_4 & s_4^{+k} \end{array}$$

If there were no context effects, the CE should be approximately the same for each standard-comparison pair.

Fig. 2 shows that the four standards produce binary data points which are displaced from the minor diagonal which represents the location of all points for which the CE is zero. The binary points are represented by the last symbol of a given kind on the curves of Fig. 2. Furthermore, the magnitude of the displacement increases with the intensity of the standard. (Except for s_1 which has a value closer to zero than s_2 . The small number of observations, 100 for each standard, makes this inversion not unlikely, given the margin of experimental error.)

An interesting aspect of these data is the apparent separation into individual iso-sensitivity curves. Without the use of the rating data, which produced these curves, another experiment would have been necessary to determine whether such pairs of points as s_3 and s_4 were actually on separate operating characteristics. Separate curves are not what would be expected from signal detection theory, for the ~~energy of the difference signal~~ ^{S/N} is constant across these four conditions. The meaning of this last result has yet to be explored, but the inadequacy of the method of constant stimuli for the examination of the CE is apparent.

Figure 3 displays these same data in a form which allows a check on the equality of variance assumption referred to earlier. The coordinate system of this plot is one in which the probability scale is transformed so that normal deviates are linearly spaced. On this square, if the data were indeed drawn from two normal probability density functions of equal variance, the data would lie in a straight line having a slope of unity. It is seen that even for such small samples, these variance assumptions are probably valid.

Conclusions

From the foregoing set of observations we may conclude that a bias in the subject's response structure exists relative to his location of the point of subjective equality. Classically this bias, the constant error, was attributed to the signal processing characteristics of the receptor system. At other times, it was thought to be the result of spatial or temporal transformations on stored information prior to the subject's response. It is fairly evident now, that this bias may be attributed to a

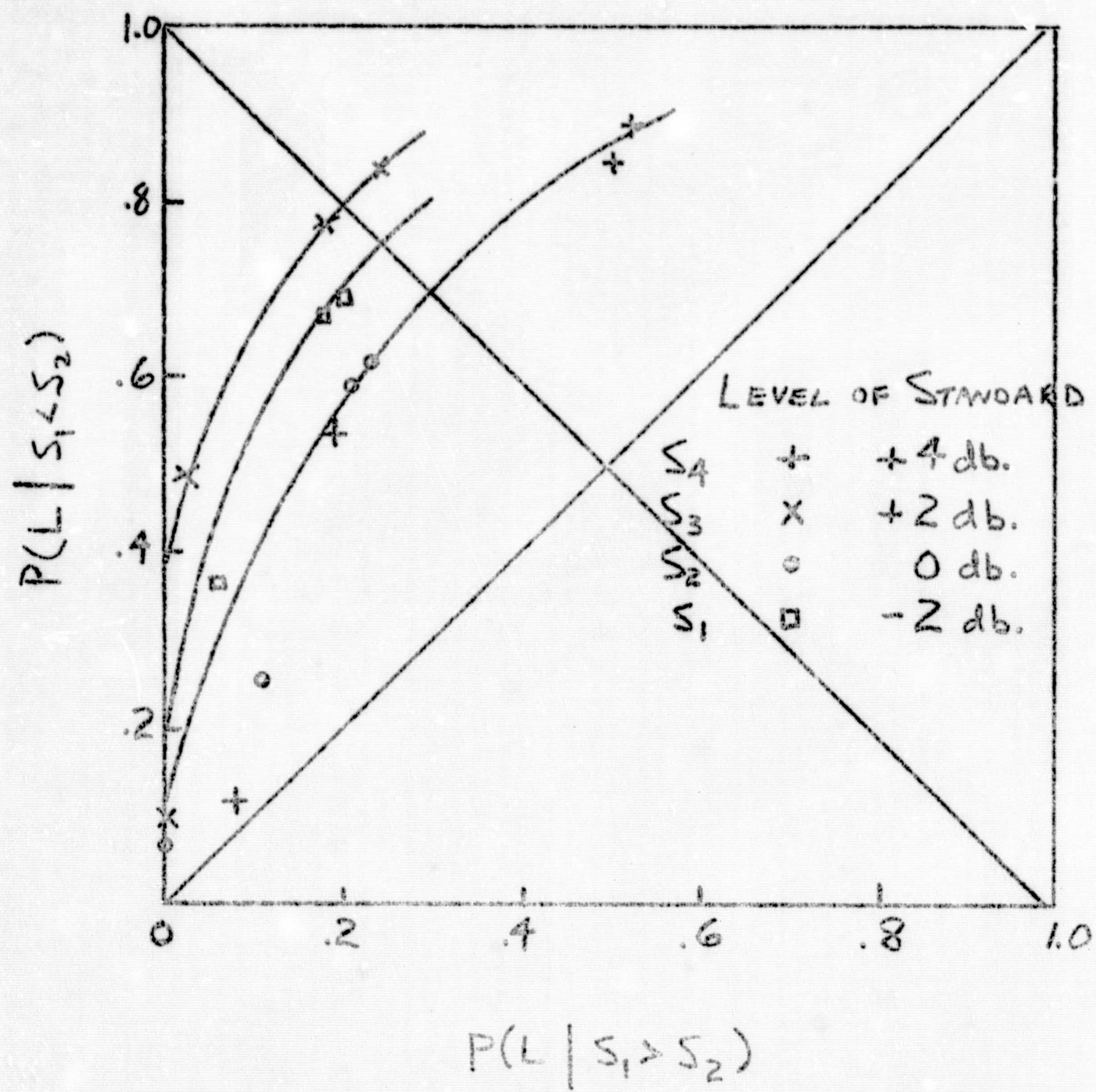
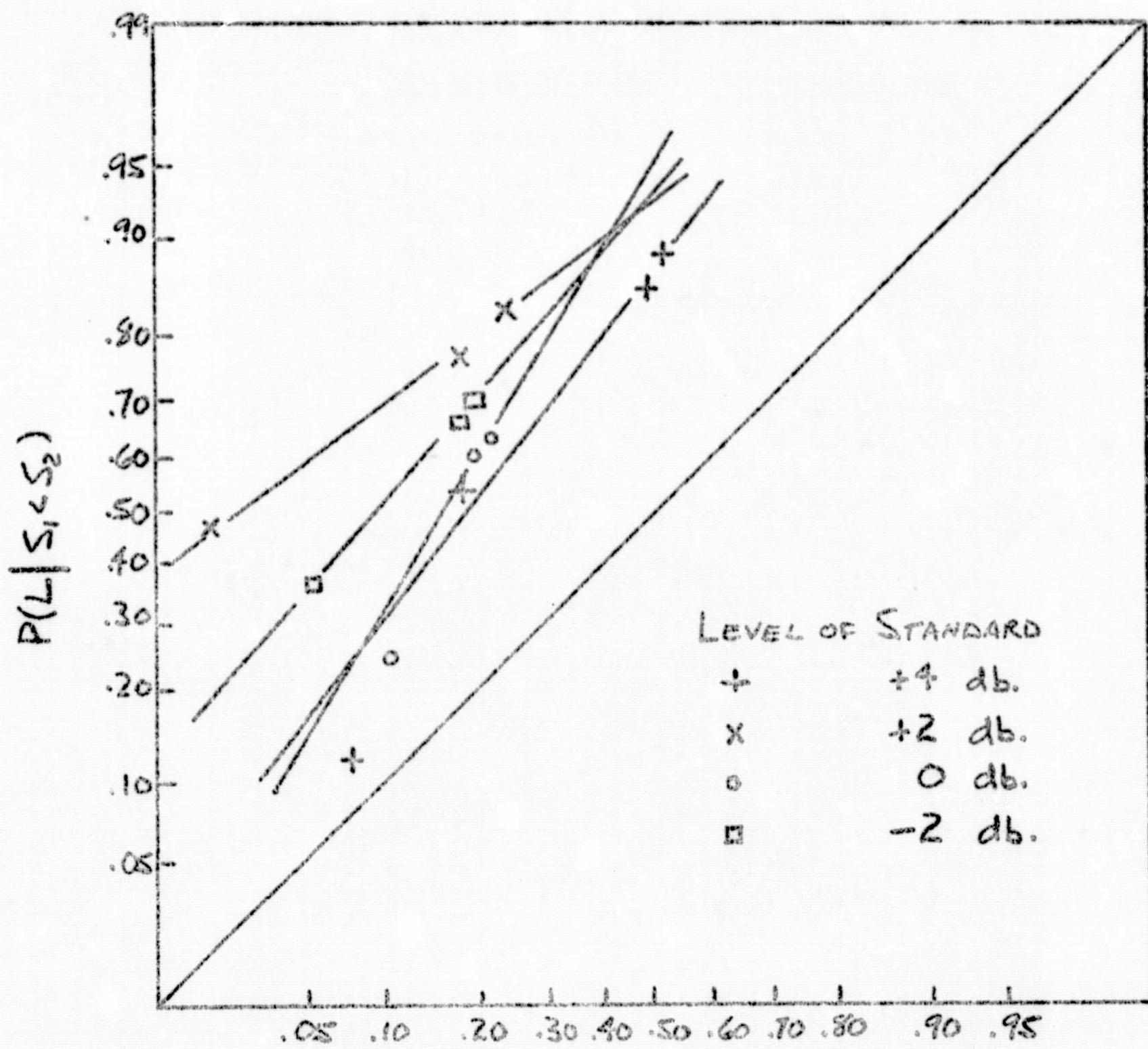


FIGURE 8



$P(L|S_1 > S_2)$

FIGURE 9

variety of factors, not the least of which are factors that influence the subject's responses directly, and therefore, can be rightly thought of as non-sensory.

Our results do not, however, lead to the conclusion that the constant error is simply a question of response biasing. There is still an open issue as to whether the sensory system and the residual traces induced by the sensory processing may give rise to these biases. In order to study this question, we have shown that it is necessary to alter the classical format of the method of constant stimuli to eliminate the demonstrated sequential effects that certainly exist in all of the classical forms of the experiment. By forcing the judgments into a paradigm that demands relative comparisons, we are of necessity required to present a variety of stimuli in any given session. The consequences of this necessity lead to extremely elaborate experiments in order to assess the role of experimental variable admixtures within the experiments themselves. For example, if varying intervals between the standard and comparison signal are to be used, should they all occur in an intermixed fashion within a single session or should the intervals be presented on separate sessions and the subject trained until asymptotic performance with that interval has been obtained? The same question can be raised for the various signals for which comparisons are to be made. Clearly, within the context of the experiment as described here, it is necessary to present signal pairs of mixed intensities throughout a single situation, because a critical parameter of the signal is its presentation probability relative to other signals in the stimulus set.

Finally, the flexibility of stimulus programming needed for experiments of this kind make it impossible to perform simple pilot experiments; we must perforce invest a great deal of time and energy in the elaborate programming equipment that is needed to effect accurate control over the various stimulus parameters that we are forced to use. If we are lucky, it may be possible to demonstrate that these effects are quite independent of the stimuli in every sense of the term including independent of the representation of the stimulus. This appears to be an unlikely event, but if this were demonstrated to be the case, the entire constant error phenomenon could be relegated to a decent burial as part of simple response bias effects induced by variations in

payoffs, expectations, signal presentation probabilities, and all of the other paraphenalia that produce changes in subject's response independent of the sensory events. If, on the other hand, the constant error is shown to depend upon features of the signal or its representation, we shall be forced to explore the way in which it enters into theories of detection and recognition before these theories can be considered complete.

Appendix

Rationale for confidence ratings as a response measure

The assumptions underlying this experiment follow from an extension of the well-known theory of signal detection. The relevant signal in the present case is formed by the difference between the pair of signals presented on each trial. We shall consider a simplified case of the present experiment, where there are only two signals which are not identical in terms of acoustic energy but alike in all other respects such as frequency, duration, phase, etc.

The notation $s_1 < s_2$ is used to indicate that the second signal of the pair is the member having greater acoustic energy. Since the two signals differ on only this dimension, it is natural to use this ordering to establish the identification function for the subject, that is, to define a correct response. Thus for the pair $s_1 < s_2$, "louder" is correct, while for $s_1 > s_2$, "softer" is the appropriate response.

The stimulation is considered to be reduced by the subject to an observation which is a random variable X on a unidimensional continuum. The different signal pairs produce different distributions of the random variable. In order to make the mathematics tractable, the theory of signal detectability has typically assumed that both distributions are continuous, normal and of equal variance.

The observer supposedly has acquired experience regarding his observations and the feedback of correct responses, so that upon making an observation x , he can estimate which distribution was more likely to have produced it. He is thought to do this systematically, but not without motivational bias, by establishing a "cut point" c on the X axis. The system of responding under these conditions is:

if $x \geq c_1$, respond with "louder"
if $x < c_1$, respond with "softer"

The following demonstrates that this arrangement provides a legitimate probability measure on the space consisting of all possible outcomes of the experiment.

The relevant distributions to consider are the two conditional probability density functions,

$$g(x|s_1 < s_2) \quad \text{and} \quad f(x|s_1 > s_2),$$

illustrated in Fig. 10. Clearly, these each satisfy the criterion

$$\left. \begin{array}{l} f(x|s_1 > s_2) \geq 0 \\ g(x|s_1 < s_2) \geq 0 \end{array} \right\} \quad \text{for all } x.$$

$$\int_{-\infty}^{+\infty} f(x|s_1 > s_2) dx = 1$$

and

$$\int_{-\infty}^{+\infty} g(x|s_1 < s_2) dx = 1$$

so that the probability of an event A defined as $a \leq x \leq a'$ for $s_1 < s_2$, is given by

$$P(A) = \int_A g(x|s_1 < s_2) dx = \int_a^{a'} g(x|s_1 < s_2) dx.$$

A particular event of interest is $x \geq z_1$, for which the subject is assumed to respond "louder":

$$P(L|s_1 < s_2) = \int_{z_1}^{+\infty} g(x|s_1 < s_2) dx.$$

It follows from the cut point definition that

$$P(S|s_1 < s_2) = \int_{-\infty}^{z_1} g(x|s_1 < s_2) dx$$

and thus

$$P(L|s_1 < s_2) + P(S|s_1 < s_2) = 1.$$

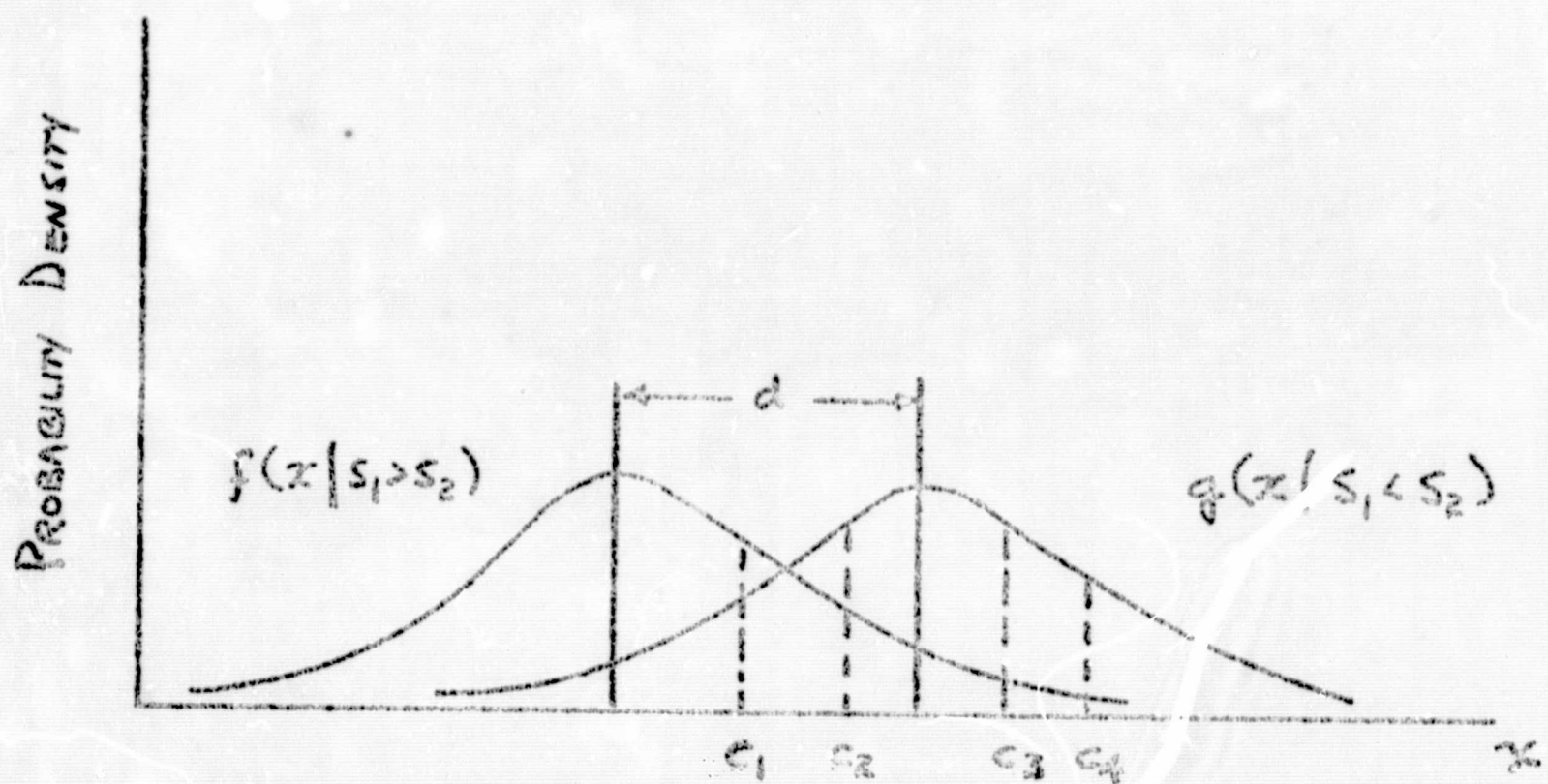


FIGURE 10

Another relevant event is the generation of "false alarms", those occasions when the observer incorrectly identified an observation produced by $s_1 > s_2$. This is given by

$$P(L|s_1 > s_2) = \int_{-\infty}^{+\infty} f(x|s_1 > s_2) dx.$$

The values $P(L|s_1 < s_2)$ and $P(L|s_1 > s_2)$ are used to determine the separation between the means of the two distributions. This distance, d , is a function of the energy difference between the two signals, expressed in standard deviation units. The extension of this characterization of the discrimination process will now proceed to the confidence rating paradigm.

It has been shown (Egan et al., 1959) that human observers can maintain several criteria simultaneously. This provides the basis for requesting the subject to report not only that an observation fell above or below a criterion, but something of the nature of how far it exceeded the cut point. Only the case for the $s_1 < s_2$ distribution will be examined, the other being symmetrical.

Under the rating procedure, the observer indicates the highest cut point which the observation exceeds. That is, in the case of 4 criteria:

- if $x \geq c_4$, respond with L_4
- if $c_3 \leq x < c_4$, respond with L_3
- if $c_2 \leq x < c_3$, respond with L_2
- if $c_1 \leq x < c_2$, respond with L_1

We assume that a response made to a high criterion such as L_4 would have been made to L_3 , L_2 , and L_1 if we were to ask the subject to do so. This is the basis for the cumulative relation used to calculate the probability that the response attains a given criterion:

$$P(L_4|s_1 < s_2) = \int_{c_4}^{+\infty} f(x|s_1 < s_2) dx$$

$$\begin{aligned}
 P(L_3 | s_1 < s_2) &= \int_{c_3}^{c_4} g(x | s_1 < s_2) dx + \int_{c_4}^{+\infty} g(x | s_1 < s_2) dx \\
 &= \int_{c_3}^{c_4} g(x | s_1 < s_2) dx + P(L_4 | s_1 < s_2)
 \end{aligned}$$

and so forth, so that the total probability of saying louder to any criterion is:

$$P(L | s_1 < s_2) = \int_{c_1}^{\infty} g(x | s_1 < s_2) dx .$$

This is the same relation obtained for the binary case, so it is seen that the use of confidence ratings can also produce the data in binary form by ignoring all but one cut point.

References

- Atkinson, R.C., Carterette, E.C. & Kinchla, R.A. Sequential phenomena in psychophysical judgments. Inst. Radio Engineers Trans. on Information Theory, IT-8, S155-162, 1962.
- Atkinson, R.C., Carterette, E.C. & Kinchla, R.A. The effect of information feedback upon psychophysical judgments. Psychon. Sci., 1964, 1, 83-84.
- Bush, R.R. Estimation and evaluation. In Luce, Bush & Galanter (Eds.) Handbook of Mathematical Psychology. Vol. 1. New York: Wiley, 1963, Pp. 429-469.
- Bush, R.R., Galanter, E. & Luce, R.D. Characterization and classification of choice experiments. In Luce, Bush & Galanter (Eds.) Handbook of Mathematical Psychology. Vol. 1. New York: Wiley, 1963, Pp. 191-243.
- Clarke, F.R. Confidence ratings, second choice responses and confusion matrices in intelligibility tests. J. acoust. Soc. Am., 1960, 32, 35-46.
- Egan, J.P., Schulman, A.I. & Greenberg, G. Operating characteristics determined by binary decisions and by ratings. J. acoust. Soc. Am., 1959, 31, 768-773.
- Fernberger, S.W. Interdependence of judgments for the method of constant stimuli. J. exp. Psychol., 1920, 3, 126-150.
- Friedman, M.P. & Carterette, E.C. Detection of Markovian sequences of signals. J. Acoust. Soc. Am., 1964, 36, 2334-2339.
- Galanter, E. Contemporary Psychophysics. In New Directions in Psychology. New York: Holt, Rinehart, & Winston, 1962.
- Helson, H. The use of comparative rating scales for the evaluation of psychophysical data. Amer. J. Psychol., 1954, 67, 321-326.
- Irwin, F.W. & Preston, M.G. Avoidance of repetition of judgments across sense modalities. J. exp. Psychol., 1937, 21, 511-520.
- Köhler, W. Zur theorie des Sukzessivvergleichs und der Zeitfehler. Psychol. Forsch., 1923, 4, 115-175.
- Luce, R.D. & Galanter, E. Discrimination. In Luce, Bush & Galanter (Eds.) Handbook of Mathematical Psychology. Vol. 1. New York: Wiley, 1963, Pp. 191-243.
- Michels, W.C. & Nelson, H.A. A quantitative theory of time-order error effects. Amer. J. Psychol., 1954, 67, 327-334.

- Needham, J.G. The time error in comparison judgments. Psychol. Bull., 1934, 31, 229-243.
- Needham, J.G. The effect of the time interval upon the time error at different intensive levels. J. exp. Psychol., 1935a, 18, 530-543.
- Needham, J.G. Contrast effects in judgments of auditory intensities. J. exp. Psychol., 1935b, 18, 214-226.
- Pollack, I.J. Sound level discrimination and variation of reference testing conditions. J. acoust. Soc. Am., 1955, 27, 474-480.
- Pollack, I.J. & Decker, L.R. Confidence ratings, message reception and the receiver operating characteristic. J. acoust. Soc. Am., 1958, 30, 286-292.
- Pratt, C.C. Time errors in the method of single stimuli. J. exp. Psychol., 1933, 16, 798-814.
- Swets, J.A., Tanner, W.P. & Birdsall, T.G. Decision processes in perception. Psychol. Rev., 1961, 68, 301-340.
- Verplank, W.S., Collier, G.H., & Cotton, J.W. Nonindependence of successive responses in measurements of the visual threshold. J. exp. Psychol., 1952, 44, 273-282.
- Woodrow, H. Weight discrimination with a varying standard. Amer. J. Psychol., 1933, 45, 391-416.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) University of Washington Seattle, Washington 98105 Eugene Galanter, Principal Investigator		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP	
3. REPORT TITLE STUDIES OF THE CONSTANT ERROR			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) A technical report			
5. AUTHOR(S) (Last name, first name, initial) Ronken, Don A. and Galanter, Eugene			
6. REPORT DATE 1 November 1965		7a. TOTAL NO. OF PAGES 38	7b. NO. OF REFS 23
8a. CONTRACT OR GRANT NO. NONR 477(34)		9a. ORIGINATOR'S REPORT NUMBER(S) PLR - 14N	
b. PROJECT NO.			
c.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
d.			
10. AVAILABILITY/LIMITATION NOTICES This report is available upon request to the Principal Investigator, Eugene Galanter, University of Washington, Seattle, Washington 98105.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY Office of Naval Research	
13. ABSTRACT A series of experiments are reported in which attempts are made to study the role of stimulus variables that may affect the magnitude of the constant error. The results indicate that in addition to bias effects that cannot rightly be ascribed to the stimulus, there are certain forms of potential sequential biases that may or may not be stimulus induced. The classical experiment to investigate the phenomenon is redesigned, and some preliminary results are presented to show that the technique is feasible.			

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
time-order error time error constant error effects of stimuli and responses on "time error" median bias obtaining relative judgments without absolute biases obtaining comparative judgments without contextual effects						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those

imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.