A psychophysical interpretation of a "categorical perception" experiment by Hary and Massaro

NEIL A. MACMILLAN

*Brooklyn College, City University of New York
Brooklyn, New York*

In this comment, I offer a psychophysical account of an experiment on "categorical perception" recently reported by Hary and Massaro (1982). The account is psychophysical in that it uses theoretical models of the procedures used by Hary and Massaro, which have been developed in other stimulus contexts, but which are believed to be general. The present explanation invokes no mechanisms or effects which have not already been found in studies of intensity discrimination, and is thus a step towards parsimony. In particular, this description of the data does not refer to categorical perception, in any of its guises.

Hary and Massaro Summarized

Hary and Massaro performed identification and discrimination experiments, of the sort usually done to test whether a continuum is "categorically" perceived, using the "pluck-bow" continuum of sawtooth-wave bursts differing in rise time. The identification experiment allowed only two responses, "pluck" and "bow." Discrimination was of the variable-standard same-different design: each discrimination trial contained two signals whose rise time was either the same or different. On "different" trials, the rise time of the first signal could be either less than or greater than that of the second. The discrimination task was *roving-level*, in that the particular pair of signals presented varied from trial to trial across the stimulus range.

Two rise-time continua were used, each with 15-msec steps. The "standard" continuum ranged from 0 to 120 msec, and the "bipolar" continuum ranged from -60 to +60 msec. (A "negative" rise time is the time to fall from an initial level to the point at which a more gradual decline begins.) Figure 1, adapted from Hary and Massaro (1982), shows the main result of the experiment.¹ The predicted discrimination points in the figure were computed from the thresh-

old model of Liberman, Harris, Hoffman, and Griffith (1957), on the assumption that discrimination is completely determined by labeling.

The main conclusions reached by Hary and Massaro were these:

(1) Agreement between observed and predicted discrimination, a criterion of categorical perception, is better in the bipolar condition.

(2) The identification function is steeper for the bipolar condition; this is also a characteristic of categorical perception.

(3) Discrimination is highly context-dependent; this is because the underlying categories that partially determine discrimination are different in the two conditions for some stimulus pairs.

(4) By criteria 1 and 2, the bipolar pluck-bow continuum is categorically perceived. Yet the results of the standard condition show that the continuum can also be continuously perceived. Thus "categorical results found in the identification-discrimination task do not necessarily imply that perception is categorical" (p. 418).

A Psychophysical Approach to the Data

Conclusion 4 above suggests that a paradoxical result is being reported, but the paradox clearly depends on the definitions of categorical and continuous perception. Suppose that the reader had never heard of categorical perception (as the author sometimes wishes were true for him); would the data seem surprising in any way?

To answer this question, it is necessary to recast the data into a conventional form. In experiments on categorical perception, percent correct discrimination and percent of identification responses falling into one class (e.g., "plucked") are generally reported; as has been pointed out elsewhere (Macmillan, Kaplan, & Creelman, 1977), this is tantamount to assuming a special case of a low-threshold model of discrimination (Luce, 1963). In psychoacoustic experiments using the same tasks, signal detection theory (SDT; Green & Swets, 1974) and Thurstonian models are used to compute values of d' . This is not because SDT has been proved correct and threshold theory wrong, but because (1) it is much more difficult to abstract an unbiased sensitivity measure from threshold theory than it is to compute d' , and (2) threshold theory offers only a very primitive model of identification. Virtually all existing psychophysical work on discrimination and identification is presented in Thurstonian and SDT terms.

Experiments like that of Hary and Massaro are not ideal for SDT/Thurstonian analysis. Relatively large numbers of subjects serve for relatively few trials. Perfect performance in discrimination (100% cor-

This article was written while I was on leave at the Research Laboratory of Electronics at M.I.T. and was supported by a National Research Service Award from N.I.H. and Grant 13489 from the PSC-CUNY Research Award Program. Correspondence address is Department of Psychology, Brooklyn College, Brooklyn, New York 11210. I thank Dom Massaro for his comments, and especially for providing the data from Hary and Massaro (1982). I am grateful to Terry Allard for provocative discussions.

Figure 1. Identification, discrimination, and discrimination predicted by the threshold model of Liberman et al. (1957) for two rise-time continua. Adapted with permission from Hary and Massaro (1982), Figures 4 and 5.

rect) and identification (100% of the presentations of a particular stimulus assigned the same response) sometimes occurs; both results imply infinite values of d' . One would also prefer to find d' for each subject and average the result, rather than compute a d' from averaged data. These difficulties are sometimes used as a justification for preferring threshold analysis. But threshold analysis of these experiments is also flawed: percent correct, the measure of discrimination performance, is far from an unbiased measure of sensitivity when bias is substantial.² Observed identification proportions of 0 and 1 represent floor and ceiling effects which conceal the true sensitivity of the subject. Certainly, from either a threshold or an SDT point of view, experiments in categorical perception are poorly designed. However, the d' computed from averaged data does not differ greatly from the average d' unless subjects have very different biases (not true in the present case), or unless some of the observed proportions of 0 and 1 reflect true d' values that are extremely large. Cautious application of SDT to Hary and Massaro's data yields insights unavailable through (equally cautious) application of threshold analysis. The floor and ceiling effects in identification, however, are severe enough that few conclusions can be stated with any confidence about that task; the present analysis is largely confined to discrimination.

Estimating d'

Hary and Massaro report values of d' , but these values are not unbiased and do not correspond to a

distance along a psychological dimension. According to SDT, each signal in an experiment generates an internal distribution of effect; the distributions generated by different signals are, in the simplest version of the model, normal, differing only in mean. To apply SDT to a particular experiment, one must hypothesize the decision variable used by the subject. This leads to a description of the decision space of the subject in terms of two or more distributions and one or more criterion lines.

In the simplest and best known applications of SDT, such as one-interval (yes-no) and two-alternative forced-choice experiments, the optimal decision variable yields only two normal distributions, differing in mean. In that case, d' can be estimated by converting the hit rate, H , and false-alarm rate, FA , to z scores and computing $d' = z(H) - z(FA)$.

In the same-different experiment of the variable-standard type, no decision variable gives rise to two normal distributions. The most thoroughly investigated rule (Sorkin, 1962; Wickelgren, 1969) is to compute the difference between the observations in the two intervals, $X_1 - X_2$. This leads to three normal distributions: same pairs have a distribution with a mean of 0, different pairs in which $X_1 < X_2$ have a distribution with a mean less than 0, and different pairs in which $X_1 > X_2$ have a distribution with a mean greater than 0. There is a variable which generates two distributions for this experiment, namely $|X_1 - X_2|$, but these distributions are not normal and do not have equal variance (see Macmillan et al., 1977). It is possible to estimate d' from same-

Table 1
Values of Discrimination (d') From Hary and Massaro
Experiment 2

Rise Times (msec)	Standard d'	Bipolar d'
-60, -30		*
-45, -15		0.84
-30, 0		2.08
-15, 15		3.34
0, 30	2.79	4.07
15, 45	2.15	1.87
30, 60	1.85	0.71
45, 75	1.53	
60, 90	0.71	
75, 105	0.50	
90, 120	0.69	

* $P(\text{"same"} | \text{same})$ was greater than .99; d' cannot be reliably computed.

different data, however, and tables for doing so have been presented by Kaplan, Macmillan, and Creelman (1978).³

Values of d' for Experiment 2 of Hary and Massaro, assuming the differencing model, are given in Table 1. Although d' as a function of rise time has much the same shape as that presented in the original paper (where d' was estimated by simply subtracting z -scores), there are some important differences. Let us compare the largest and smallest d' in each condition; since d' is a distance measure, the appropriate measure of this discrepancy is a ratio. This ratio, based on Hary and Massaro's reported d' 's is 16.1 in the standard condition and 10.9 in the bipolar condition. According to the differencing model, however, these ratios are 5.6 and 6.1. Hary and Massaro's misapplication of SDT leads to an exaggerated impression of the peak-to-trough discrimination ratio, that is, a relative underestimate of "within-category" discriminability. Their attempt to relate observed and predicted d' is similarly flawed.

Subjects in this study maintained an essentially constant hit rate [$P(\text{"same"} | \text{same})$] across the stimulus range, for each condition of each experiment. This is what would be expected from a subject using the differencing model.

Effects of Range and Anchors on Discrimination

The best known generalization about how discrimination varies along a stimulus continuum is Weber's law; van Heuven and van den Broecke (1979), using an adaptive procedure, found this relation to be approximately true for rise time. If stimuli are logarithmically spaced, Weber's law implies that d' will be the same for all adjacent pairs. When stimuli are arithmetically spaced, as in Hary and Massaro's experiment, some assumption about the form of the

psychometric function (d' as a function of stimulus difference) must be made to predict the data. If this function is presumed to be a normal ogive when stimuli are on a log axis, then $d' = c \cdot \log(I_1/I_2)$. If $I_1 - I_2$ is constant, this reduces to $d' = c \cdot \log(1 + 1/k)$. As k increases ($k = 1, 2, \dots$), d' decreases; the function is concave upwards, with an asymptote at 0.

Weber's law is not generally found in experiments with tones differing in intensity. In a series of papers using such stimuli (Berliner & Durlach, 1973; Berliner, Durlach, & Braida, 1977; Durlach & Braida, 1969), two important effects have been discovered which do not seem a priori to have to do with auditory intensity particularly. First, discriminability is an inverse function of stimulus range; subjects can discriminate a given pair better if the possible stimuli to be presented are limited to a narrow region of the continuum. Second, discriminability improves near the edges of the stimulus range. Berliner, Braida, and Durlach attribute this to the operation of "anchors" at the edges: since the end stimuli are easily remembered by the subjects, the variability limiting the perception of such stimuli is reduced.

In experiments on auditory intensity, the only available anchors are the endpoints. In Hary and Massaro's experiment, the endpoints presumably serve as anchors, but the bipolar condition contains one other memorable stimulus: the 0-msec rise-time signal, which is the only signal to have neither a rising nor a falling envelope. Let us assume that this anchor serves to divide the bipolar stimulus continuum into two smaller continua; what should the data look like?

Figure 2a shows predictions for the standard and bipolar conditions assuming only Weber's law and the range effect; since the central anchor in the bipolar condition reduces the effective stimulus range, bipolar performance will be higher. In Figure 2b, anchor effects are added to points near the anchors. Anchors may or may not be strong enough to produce an upturn in discriminability at the edges of the range. Hary and Massaro's recomputed data, from Table 1, are presented in Figure 2c.⁴ Clearly, the range and anchor effects provide a good qualitative description of the data.

Relation Between Identification and Discrimination

As remarked earlier, a detailed analysis of the identification data cannot be performed because of the large number of proportions that are 0 or 1. It is worth making an attempt, however, to test Hary and Massaro's conclusion 1, that identification predicts discrimination better in the bipolar than in the standard condition. According to a Thurstonian model (e.g., Braida & Durlach, 1972), d' can be estimated from identification, for the two-response task used

Figure 2. Predicted and observed discrimination along "standard" and "bipolar" rise-time continua. In (a), d' s for the standard condition are assumed to follow Weber's law; for the bipolar condition, Weber's law is assumed to apply both above and below the 0-msec anchor. Since the anchor effectively reduces the range, performance levels are higher. In (b), anchor effects (increased sensitivity at the edges, and at 0 msec) are added to the effects shown in (a). The reduced range and the anchors are arbitrarily assumed to raise d' by 25%. Data from Hary and Massaro (1982) are plotted in d' terms in panel c.

here, by converting to z scores the estimated probabilities of identifying signals as "pluck" and subtracting; that is,

$$d'(s_i, s_j) = z[\text{prob}(\text{"pluck"} | s_i)] - z[\text{prob}(\text{"pluck"} | s_j)]$$

for stimuli s_i and s_j , the rise time of s_i being less than that of s_j . These values of identification d' can be compared with corresponding values obtained from the discrimination data.

Figure 3 plots discrimination d' vs. identification d' for both standard and bipolar conditions. Because of ceiling and floor effects in identification, there are only two points for the bipolar condition. Discrimination is well predicted from identification: the best-fitting line through the origin accounts for 81% of the variance. Its slope, an estimate of the ratio of discrimination to identification d' , is 1.33. In intensity-resolution experiments, discrimination and identification d' s are also found to be proportional, and ratio between them also exceeds 1. Thus, Hary and Massaro's discrimination data are not only well predicted by their identification data in a statistical sense, but are related, as would be expected for a simpler psychoacoustic continuum.

A Brief Comment About Identification

The range and edge effects described earlier for roving-level discrimination have also been found to occur for identification. In particular, sensitivity is greatest near an anchor. Thus, if the 0-msec signal is serving as an anchor in the bipolar condition, the

identification function should be steep in that condition, as it is.

Conclusions

Hary and Massaro comment that "models of continuous perception should be able to reproduce both

Figure 3. Relation between discrimination and identification d' for Hary and Massaro's Experiment 2. The line through the origin which best fits the data (from both conditions) is also shown. See text for details.

the categorical and continuous functions found in this experiment" (p. 415). A reanalysis of Hary and Massaro's data in conventional psychophysical terms vindicates this faith. The bipolar condition has a steep identification function because of the existence of an anchor in midrange for that condition. This anchor effect, and an effect of stimulus range, both of which have been documented in experiments on auditory intensity, account for the dependence of discriminability on context. Where the relation between discrimination and identification can be assessed, these tasks are related as they are in intensity perception. No reference to hypothetical categories is needed to describe the data.

Postulating an anchor at the middle of the bipolar stimulus range captures some of the spirit of postulating two categories in that condition, or of postulating, as Hary and Massaro do, a "perceptual discontinuity" near the 0-msec stimulus. The observant reader may suspect that the present analysis differs only semantically from Hary and Massaro's. But it is important that new data be referred, wherever possible, to existing models in the hope of finding common explanations of effects in different stimulus domains. To many, stimulus continua studied in categorical perception research are interesting because they appear to follow different patterns from other continua. Some of these differences, as this comment illustrates, arise from methodology rather than substance.

REFERENCES

- BERLINER, J. E., & DURLACH, N. I. Intensity perception. IV. Resolution in roving-level discrimination. *Journal of the Acoustical Society of America*, 1973, **53**, 1270-1287.
- BERLINER, J. E., DURLACH, N. I., & BRAIDA, L. D. Intensity perception. VII. Further data on roving-level discrimination and the resolution and bias edge effects. *Journal of the Acoustical Society of America*, 1977, **61**, 1577-1585.
- BRAIDA, L. D., & DURLACH, N. I. Intensity perception. II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, 1972, **51**, 483-502.
- DURLACH, N. I., & BRAIDA, L. D. Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, 1969, **46**, 372-383.
- GOUREVITCH, V., & GALANTER, E. A significance test for one-parameter isosensitivity functions. *Psychometrika*, 1967, **32**, 25-33.
- GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics*. Huntington, N.Y.: Krieger, 1974.
- HARY, J. M., & MASSARO, D. W. Categorical results do not imply categorical perception. *Perception & Psychophysics*, 1982, **32**, 409-418.
- KAPLAN, H. L., MACMILLAN, N. A., & CREELMAN, C. D. Tables of d' for variable-standard discrimination paradigms. *Behavior Research Methods & Instrumentation*, 1978, **10**, 796-813.
- LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, **54**, 358-368.
- LUCE, R. D. Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.
- MACMILLAN, N. A., KAPLAN, H. L., & CREELMAN, C. D. The psychophysics of categorical perception. *Psychological Review*, 1977, **84**, 452-471.
- NOREEN, D. L. Optimal decision rules for some common psychophysical paradigms. In S. Grossberg, (Ed.), *Mathematical psychology and psychophysiology*. Providence, R.I.: American Mathematical Society, 1981.
- SORKIN, R. D. Extension of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America*, 1962, **34**, 1745-1751.
- VAN HEUVEN, V. J. J. P., & VAN DEN BROECKE, M. P. R. Auditory discrimination of rise and decay times in tone and noise bursts. *Journal of the Acoustical Society of America*, 1979, **66**, 1308-1315.
- WICKELGREN, W. A. Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology*, 1969, **6**, 13-61.

NOTES

1. Hary and Massaro actually conducted three experiments. Experiment 1 consisted of only a standard condition. Experiments 2 and 3 used both a standard and a bipolar condition, but in Experiment 3 discrimination preceded identification and larger step sizes (45 msec instead of 30) were used in discrimination. Figure 1 shows the data from Experiment 2; Experiment 3 yielded qualitatively similar results. Since overall performance was very poor in Experiment 3 (in other pluck-bow studies, differences of 5 to 20 msec are sufficient to produce discriminability at this level), only Experiment 2 is reanalyzed here.
2. Bias was, in fact, substantial in this experiment; subjects favored the "same" response. The average value of prob ("same" | different) in Experiment 2 was .62; prob ("different" | same) averaged .11. (My computations are from data provided by Massaro.)
3. Noreen (1981) has presented another model for same-different, which predicts slightly higher performance levels than the differencing model. It is optimal for the situation in which the subject knows the pair of stimuli to be presented exactly; for the roving-level experiment of Hary and Massaro, the differencing model is more appropriate.
4. Only proportions between .01 and .99 are used in this and the following analyses (see Macmillan et al., 1977, p. 465, and Braida & Durlach, 1972, Appendix A). The variance of estimated z scores increases dramatically as proportions approach 0 or 1 (Gourevitch & Galanter, 1967).

(Manuscript received May 26, 1983;
revision accepted for publication August 23, 1983.)