Confidence ratings in speech perception research: Evaluation of an efficient technique for discrimination testing*

WINIFRED STRANGE and TERRY HALWES†
University of Minnesota, Minneapolis, Minnesota 55455

Discriminability of voice onset time was determined for four naive Ss in order to evaluate the relative efficiency of a testing procedure which utilized confidence ratings along with the conventional oddity task. Ss rated their judgments on oddity trials as "very very sure," "somewhat sure," or "just guessing." Each S's discrimination scores were given weights according to his ratings, yielding functions that were compared with those computed using unweighted percent correct scores. The confidence-rating technique produced readily interpretable results with about one-third as many judgments as are needed when the conventional procedure is used. Possible problems with the interpretation of the functions and the generalizability of the technique are discussed.

Research on the perception of synthesized stop consonants has revealed discontinuities in discrimination functions which can be predicted quite accurately from performance on an absolute identification test (Liberman, Harris, Hoffman, & Griffith, 1957). In the latter task, the S must assign phonemic labels to the stimuli, while on the discrimination test, he has only to perceive a difference between sounds. Typical results show that discrimination of two sounds that differ along a phonetically relevant physical continuum is only slightly better than chance if the S assigns both sounds the same phonemic label in an identification test. Two stimuli that differ by the same amount, but which the S consistently labels as different phonemes, are discriminated with almost 100% accuracy. While the latter result is to be expected, the poor discrimination within labeled categories is a somewhat unusual phenomenon. (On most psychophysical dimensions, the number of discriminably different stimuli is many times greater than the number that can be identified absolutely.) The outcome of these studies, referred to as "categorical perception," has stimulated much discussion about the mechanism for the perception of highly encoded speech sounds (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

The procedures that have been used for discrimination testing in these studies provide a measure of *relative* discriminability of pairs of speech sounds. Series of stimuli are synthesized, varying in relatively small acoustically equal steps over a range including two or more different phonemes. Then pairs of stimuli are selected from the series such that each pair differs by the same amount; all possible pairs differing by that amount are compared. Ss make judgments about the difference between paired stimuli in one of two types of tests, the ABX test or the oddity test. The comparison pairs are presented randomly over repeated test sessions, and many judgments on each comparison pair are obtained. The percentage of correct judgments for each pair is the measure used to determine relative discriminability along the acoustic continuum.

The first discrimination studies (Liberman et al, 1957) utilized an ABX test. Sequentially presented triads were constructed from each comparison pair along the acoustic continuum, such that the first and second sounds (A and B) were different and the third sound (X) was identical either to A or to B. There are four such arrangements for each comparison pair (ABA, ABB, BAA, BAB); Ss indicated whether X = A or X = B for all permutations. The probability of being correct by chance on any triad is equal to .50, i.e., over repeated tests, a score of 50% correct judgments per pair may be obtained by guessing alone. Relative discriminability scores for pairs can therefore vary from 50% to 100%. (Scores below chance, though often obtained, are uninterpretable. However, they may give an indication of the random variation and possible response biases.) Using the ABX procedure, stable results were obtained only after Ss made from 26 to 42 judgments on each comparison pair. The large number of judgments was needed to differentiate real differences in discriminability from the random flux about chance.

In an effort to reduce the testing time required to obtain stable functions, investigators developed an oddity design for discrimination testing. Triads consisting of three sounds, two identical and one different, are arranged in all six possible permutations (ABB, BAB, BBA, AAB, ABA, BAA). The S's task is to indicate which of the three stimuli is the different or "odd" one. This procedure reduces the probability of a correct judgment by chance to .33, hence, the range over which interpretable discrimination scores can vary is expanded by adding 16.7% to the range possible with the ABX task. While this results in appreciable savings in number of judgments required, testing time is still protracted. Each S has to make at least 18 judgments per comparison pair before stable differences in discriminability along the acoustic continuum are detected (e.g., Abramson & Lisker, 1967).

Reported here is the examination of a procedure that elicits additional information from the S on each trial of the discrimination test, thereby reducing the variance due to chance. A procedure that can be employed in psychophysical testing is to ask the S to rate his confidence in his discriminatory judgment. In other words, the S makes a judgment (e.g., chooses the odd stimulus) and then indicates *how sure he is* that his judgment is correct. These confidence ratings can then be used to assign weights to the discrimination scores. For example, responses and confidence ratings can be combined so that judgments given a low confidence rating (not very sure) contribute less to an overall discrimination measure than do judgments assigned a higher degree of confidence. There are several ways that confidence ratings can be elicited from Ss. One way is for the E to delimit a number of categories alone an ordinal scale from "very confident" to "just guessing," and instruct the S to rate his confidence for each response by assigning it one of the categories on the scale.

Some pilot work by the authors, done with the voice onset time (VOT) continuum (to be described below) constructed and tested by Abramson and Lisker (1967), suggested that experienced Ss were capable of assigning confidence ratings to discrimination judgments on an oddity test with a good deal of consistency. Ss used three degrees of

*This research was supported in part by grants to the University of Minnesota, Center for Research in Human Learning, from the National Science Foundation (GS1761), the National Institute of Child Health and Human Development (HD-01136), the Graduate School of the University of Minnesota, and by Haskins Laboratories, New York. We wish to thank our colleagues associated with Haskins Laboratories for their continuous support in this research. Special thanks to Drs. Arthur Abramson and Leigh Lisker for the use of their materials.

†Now at the University of Connecticut, Storrs, Connecticut 06268.

confidence (very sure, somewhat sure, and just guessing), and scores were adjusted according to these ratings. The adjusted discrimination functions obtained by this method were compared with the functions obtained by Abramson and Lisker (1967) with the conventional oddity testing procedure. The adjusted functions showed very similar results with about one-third as many judgments per comparison pair as were used by Abramson and Lisker in computing discriminability. Although this comparison was made across different Ss and experiments, it suggested strongly that the use of confidence ratings increased the efficiency in discrimination testing of sophisticated Ss.

The present study was designed to investigate the utility of using confidence ratings in combination with the oddity task when testing naive Ss' discrimination of synthetic speech sounds. The Ss tested were unacquainted with the nature of the stimuli as well as with the use of confidence ratings, and were not trained listeners. Given these conditions, discrimination functions computed by the conventional method stabilized only when 18 judgments per comparison pair were used in the comparisons. Discrimination functions computed by adjusting scores according to confidence ratings were compared with the unadjusted functions within individual Ss to determine how many fewer judgments per pair were needed to produce equally clear functions. It was also possible to determine each individual's ability to use confidence ratings consistently and thus indicate how efficient the method was in general.

For the purposes of this study, degrees of confidence were arbitrarily divided into three categories, as in the pilot work. Extremes of "very very sure" and "just guessing" were separated by a middle area defined vaguely as "somewhat sure." By using a small number of categories, it was hoped that variance due to individual interpretation of the definition of the categories would be kept to a minimum.

The Abramson and Lisker VOT continuum was used for this study. Their initial experiments and the pilot data described above show large and consistent discontinuities in discriminability for all Ss. It was expected that the relative merits of the confidence-rating technique could best be evaluated on a stimulus continuum that yielded stable results from the oddity procedure alone, given sufficient data from each S.

Table 1
Adjusted Score Values (Proportions) for ++, +, and φ Confidence Ratings on Run 1

	Two-Step			Three-Step			Four-Step		
Subject	++	+	φ	++	+	φ	++	+	φ
1	.75	.12	−.06	1.0	.20	.15	.87	.17	.04
2	.21	.12	.03	.71	.04	.00	.79	.04	.03
3	.28	.08	.11	.41	.16	−.11	.27	.10	−.02
4	.21	.02	.07	.23	.13	−.01	.46	.18	.17

METHOD

Stimuli and Apparatus

The stimulus tapes were generated at Haskins Laboratories on a computer-controlled parallel-resonance synthesizer. Control parameter values were identical to those determined by Lisker and Abramson (1967) for the initial bilabial stop series. Voice onset time is defined as the interval between the release burst (corresponding to the release of the articulators) and the onset of laryngeal pulsing (voicing). Lisker and Abramson (1964) found that this relationship was sufficient to characterize the voicing and aspiration distinctions which differentiate the English phonemes /b/ and /p/.

The constant portion of all consonant-vowel monosyllables was a 250-msec steady-state three formant vowel /ɑ/ with a fundamental frequency of 114 Hz and falling intonation contour. Thirty-one different VOT variants were added to the beginning of this vowel. Taking the release burst as zero time (0 VOT), voicing onset varied in 10-msec steps from 150 msec before the burst (−150) to 150 msec after the burst (+150). Voicing before the burst in minus VOT variants was synthesized using only low-frequency harmonics of a buzz source. For plus VOT variants the interval between the burst and onset of pulsing was filled with hiss; the first formant was suppressed.

Three sets of stimulus pairs were drawn from the series. The first set consisted of all possible pairs that differed in VOT by 20 msec (e.g., −150/−130; −140/−120 . . . +130/+150). The second set contained pairs differing by 30 msec (e.g., −150/−120, etc.); the third set of pairs differed by 40 msec each. These three sets are referred to as 2-step, 3-step, and 4-step comparisons, respectively. In all, there were 84 comparison pairs. Six oddity triads were constructed from each comparison pair (A, B) by utilizing all the permutations in the manner described above (e.g., AAB, ABA, . . .). Six test series were recorded on magnetic tape. Each series consisted of a randomly selected triad of each of the 84 comparison pairs. The triads were randomly ordered within the series; thus, the 2-step, 3-step, and 4-step comparisons were intermixed within tests. The set of six tests is referred to as a run. The test tapes were reproduced on a two-track Ampex AG500 tape recorder and presented binaurally over Koss Pro-600A earphones at a low, comfortable listening level.

Subjects

Four undergraduate college girls,native speakers of English, completed three runs, yielding a total of 18 judgments per comparison pair (three complete repetitions) over a period of 3 weeks. A session lasted between 1 and 1½ h. During the first session, instructions and the first discrimination test were given; each S was tested individually. For the remaining seven sessions, Ss were tested as a group (with one exception),[1] completing from one to three tests per session.

Procedure

Initial instructions to the Ss included: (1) a brief explanation of the nature of the synthetic stimuli with urging to "hear them as speech," (2) a description of the oddity procedure, (3) a description of confidence ratings and the reasons for wanting this information. Consistent use of the ratings in order to get the "best" functions possible was stressed. Ss were to indicate three levels of confidence in the following way: ++ for "very very sure," + for "somewhat sure," and φ for "just guessing." Score sheets contained rows of three boxes corresponding to each triad. After hearing the three syllables, the S indicated which stimulus she thought was different and how sure she was by placing the appropriate confidence symbol in the first, second, or third box. For instance, if she thought the second syllable was different and was somewhat confident that she was correct, she placed a + in the second box. Ss were required to respond to each triad.

Before each subsequent session, Ss were given reminder instructions and again cautioned to be consistent in their use of confidence ratings. Special emphasis was put on being "very sure" when they used the ++ rating. They were told that they were not necessarily supposed to improve or become more sure of their judgments as they progressed through the tests or the sessions. Upon completion of all three runs, Ss were interviewed about their linguistic backgrounds. None had had any extensive exposure to languages other than American English.

RESULTS

Each S's data were analyzed separately.

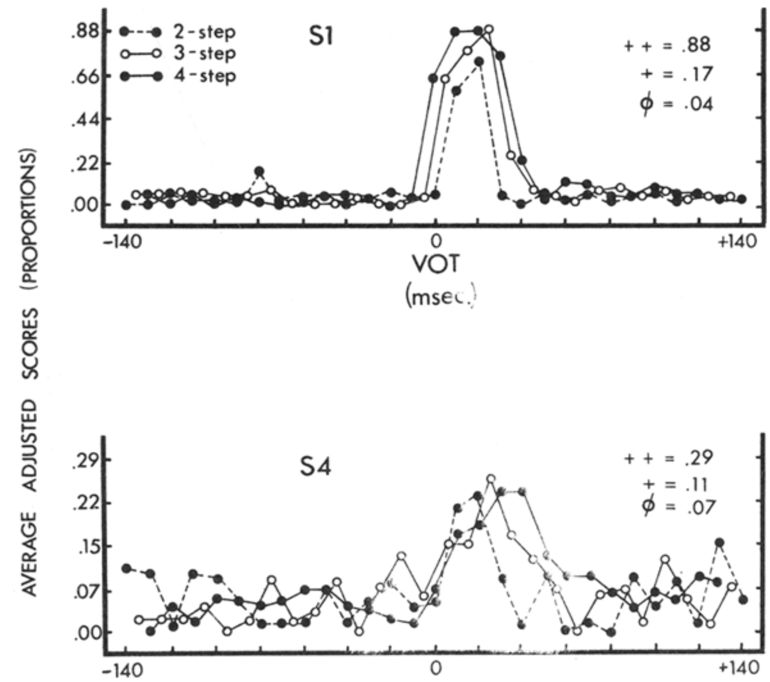

Fig. 1. Average adjusted scores for 2-step, 3-step, and 4-step functions. Individual functions for two Ss.

Tests were divided into runs, and within a run 2-step, 3-step, and 4-step comparisons were separated. Scores were adjusted according to each S's own use of the confidence ratings. For each comparison (e.g., Run 1, 3-step comparison) the scoring procedure was as follows: (1) For each of the three confidence categories, the proportion of correct responses out of the total number of responses that the S assigned that confidence rating was computed (e.g., correct ++/total ++). (2) These proportions were adjusted for chance by the formula: (proportion − .3333)/.6667. The resulting adjusted value can be thought of as the proportion of the "distance" between chance and 100% accuracy of discrimination for responses assigned that confidence rating. For example, if a S was correct on all those responses he assigned ++, then the resulting adjusted value is 1.0. If he was correct on only two-thirds of them, the value is .5. Table 1 shows the adjusted score values for 2-step, 3-step, and 4-step functions on Run 1. These proportions reflect how "accurately" Ss used the confidence ratings, that is, the consistency with which they reported differences in their ability to discriminate the odd stimulus. Ideally, the adjusted values would be ++ = 1.0, i.e., always correct when "very sure," φ = 0.0, i.e., correct only 33.3% of the time when "just guessing," with the value for + falling somewhere in between the two extremes. Note that the adjusted score values for the φ category are sometimes negative. This happens when the proportion correct out of the total assigned a particular confidence rating falls below .33 or chance. (This may reflect the presence of response biases, although no S had negative values over all step functions within a run.)

To compute discrimination functions, each S's appropriate adjusted score values were assigned to all *correct* judgments (e.g., the ++ value was assigned to all correct judgments to which S had assigned a ++ confidence rating, the + value was given to all correct judgments given +, and so on). All incorrect responses became zero, regardless of confidence rating. In those cases where the adjusted value for φ was a negative number, all responses assigned φ, both correct and incorrect, were given a zero value. Therefore, the adjusted score functions never fall below chance.[2] The

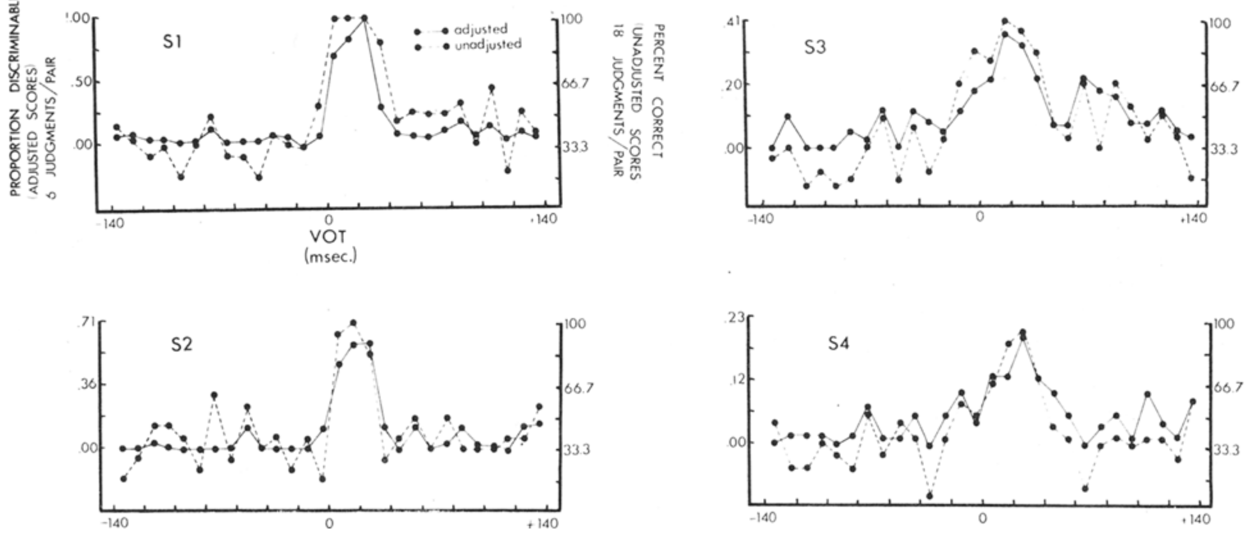

Fig. 2. Comparisons of 3-step adjusted score functions computed over Run 1 (6 judgments) and unadjusted 3-step functions computed over three runs (18 judgments). Individual functions for four Ss. The left ordinate indicates proportion discriminable and the right ordinate indicates the percent correct for each S.

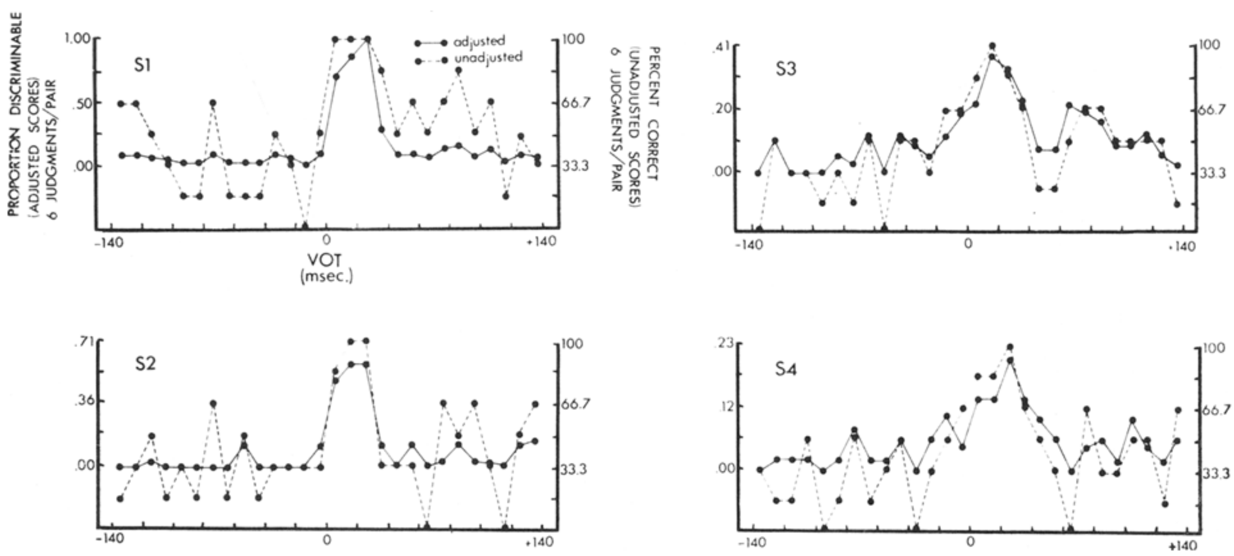

Fig. 3. Comparisons of 3-step adjusted score functions and 3-step unadjusted score functions computed over Run 1. Individual functions for four Ss. The left ordinate indicates proportion discriminable and the right ordinate indicates percent correct for each S.

functions plotted in the figures are the mean adjusted score for each pair across the number of tests of interest.

For the functions shown in Fig. 1, there is an exception to the scoring procedure described above in that the proportions of correct responses out of total responses (Step 1), were computed over all triads in Run 1 without separating 2-step, 3-step, and 4-step comparisons first. (Values for the three confidence categories are given in the figures.) Thus, the three functions for each S were computed using identical values.

For comparison with adjusted score data, scores of percentage of correct responses (disregarding confidence ratings) were computed over Run 1 (6 judgments/pair) and over all three runs combined (18 judgments/pair). These are referred to as unadjusted functions and are comparable to those of Abramson and Lisker (1967).

The 3-step functions were selected for the comparisons to be discussed because they most clearly showed differences in relative discriminability for both adjusted and unadjusted functions. Abramson and Lisker's data (1967) showed "plateaus" for the 4-step comparisons, while the 2-step comparisons failed to reach 100% accuracy, even at the point of best discrimination. Our data follow this same trend. Figure 1 illustrates the adjusted score functions on Run 1 for 2-step, 3-step, and 4-step functions for two Ss: S 1, who used confidence ratings well, and S 4, who used them rather poorly.

Of main interest is the comparison between Run 1 adjusted score functions and the unadjusted functions for combined Runs 1, 2, and 3. Figure 2 shows this for each S individually. Note that the ordinates are scaled so that the highest *possible* adjusted score is equated with 100% (unadjusted) and .00 equals 33.3% unadjusted. In other words, if a S was correct on all six judgments for a particular pair *and* he assigned ++ to every judgment, the point plotted for that pair would be equal to 100% on the unadjusted ordinate.

Inspection of these comparisons shows that adjusted discrimination functions accurately reflect relative discriminability of differences in VOT as measured by the conventional oddity procedure. The "peaks" in discrimination for adjusted and unadjusted functions appear in the same places and are similar in shape. At both extremes of the VOT continuum, discriminability was only slightly better than chance. With the confidence rating scoring adjustment, functions appear somewhat more stable than those computed from percentage scores of *three times* as many judgments per comparison pair.

Figure 3 shows the adjusted and unadjusted functions for Run 1 data only. With only six judgments per pair, the percentage measure appears to be much less useful for differentiating real differences (i.e., "peaks") from random discontinuities in discrimination along the continuum. By the adjustment scoring technique, however, six tests are sufficient to determine where the real discontinuities appear, and to approximate the relative differences in discriminability that would be reflected by the percentage measure, given many more judgments.

DISCUSSION

The results illustrated in Fig. 2 show that the technique of adjusting discrimination scores by confidence ratings yields interpretable functions with a relatively small number of tests. In this particular experiment, as few as six judgments per pair for a S were sufficient to reveal the same clear discontinuities that have been found by other researchers investigating discrimination along the VOT continuum. A comparison of each S with himself showed that the adjusted functions correctly reflected his performance as measured by the conventional testing and scoring procedure when further testing was done.

The table of adjusted score values for Run 1 shows that Ss vary considerably in their ability to use ratings consistently. (Score values for Runs 2 and 3 were very similar to Run 1 scores for each S.) Ss 1 and 2 were very accurate in their use of the highest rating as an indication of those stimuli they could discriminate very easily. Ss 3 and 4 were less capable of reflecting the appropriate information, as demonstrated by the fact that they made incorrect judgments on many of the triads to which they assigned the highest degree of confidence, resulting in small adjusted values for their ++ categories. However, even with these poor Ss, functions computed by the adjustment technique accurately reflected discriminability as measured by the conventional percentage scoring computed over many more tests. Since the scale of adjusted proportions is "coarser" when ++ values are low, one would expect the variance to increase. Despite this, the adjusted functions are still smoother than unadjusted functions. Therefore, the confidence-rating technique could probably be employed with most Ss with reasonable assurance that it would improve the interpretability of the discrimination functions. Pretraining on the use of confidence ratings with stimuli that ranged over larger degrees of difficulty might further improve the utility of this technique. This is now being investigated.

In previous research on discrimination of synthetic stop consonants, discrimination functions obtained by the oddity procedure were compared with functions predicted from identification data. On the assumption that a S's perception of these stimuli is completely categorical (i.e., that he can hear them only as phonemes), the investigators hypothesized that discrimination in the oddity task (which involves only the perception of a *difference* in the stimuli) could be predicted accurately by taking account of the relative frequency with which the S attached one or another label to each of the stimuli. Results showed that predicted and obtained discrimination functions were very similar (see Liberman et al, 1957).

In the present study, it is not clear how discrimination scores obtained with the confidence-rating technique can be compared meaningfully to those predicted from identification data. Predictions cannot be made about a S's confidence in his judgments on the basis of his labeling performance. Adjusted scores might be shown to be unbiased estimates of the asymptotic discriminability function, but this has not been attempted by the authors. An obvious, though undesirable, resolution of the difficulty of comparing adjusted functions with predicted functions, is to avoid using the confidence-rating procedure for investigations requiring these comparisons. The technique could be reserved for use in experiments dealing with further questions about discrimination on continua whose categorical nature is known (e.g., questions about the effects of training on discrimination of VOT, where the basic functions are already known). Another possibility—that the confidence-rating technique constitutes a test of categorical perception independent of any predictions from identification functions—is being considered, but will not be discussed here.

Another difficulty with the confidence-rating technique is that adjusted scores are not as amenable to statistical analysis as are raw percentages of correct responses. However, since most investigators in the area have not utilized statistics in their analysis of percentage data, this criticism is not specific to the technique presented here. Meaningful statistical analyses must be developed with either technique. A danger to be avoided when the confidence-rating technique is used is to base conclusions on too few judgments per stimulus pair. There is a limit beyond which the confidence-rating values themselves become unreliable, especially for infrequently used levels of confidence.

A final consideration is the effect of the additional work the S is required to do when he must assign confidence ratings to judgments. In observing Ss at this task, it appears that, while it may be more fatiguing (Ss must have more breaks and shorter sessions), there is an advantage in that Ss become less "frustrated." It is typical in discrimination tests with these kinds of stimuli that the S is unable to hear differences on a majority of trials, i.e., he "isn't doing very well." With the confidence-rating technique, the S can "disown" trials on which he hears no difference, having been told he should not be able to hear them all. Of course, there is the danger that Ss may become complacent and give only ϕ ratings, but the authors found that urging the Ss to be conscientious with their ratings was sufficient to keep performance stable in the present experiment.

This study has shown that the confidence-rating technique can be used advantageously to determine the presence and location of peaks in cue discrimination functions for some stop consonants, with a great savings in testing time. As yet, the technique has not been tried on continua which are known to be noncategorically perceived (e.g., steady-state vowels) or continua whose categorical nature is unknown.

REFERENCES

ABRAMSON, A. S., & LISKER, L. Discriminability along the voicing continuum: Cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, September 1967. Also in Status Report on Speech Research, SR-11, Haskins Laboratories.

LIBERMAN, A. L., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.

LIBERMAN, A. L., 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.

LISKER, L., & ABRAMSON, A. S. A cross-language study of voicing in initial stops: Acoustical measurements. Word, 1964, 20-3, 384-422.

LISKER, L., & ABRAMSON, A. S. Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, September 1967. Also in Status Report on Speech Research, SR-11, Haskins Laboratories.

NOTES

1. S 2 missed the second group session and was tested separately before the third group session. Therefore, the sequence of tests remained the same for all Ss.

2. Functions were plotted using scores which took account of the minus adjusted values by assigning the minus value to *in*correct judgments given that confidence rating and 0 to correct judgments. There was negligible difference between these functions and those shown in the text.

(Accepted for publication July 6, 1970.)