Perception & Psychophysics
1985, 37 (1), 35-44

Cross-language evidence for three factors in speech perception

JANET F. WERKER and JOHN S. LOGAN
Dalhousie University, Halifax, Nova Scotia, Canada

A continuing controversy concerns whether speech perception can be best explained by single-factor psychoacoustic models, single-factor specialized linguistic models, or dual-factor models including both phonetic and psychoacoustic processes. However, our recent cross-language speech perception research has provided data suggesting that a three-factor model, including auditory, phonetic, and phonemic processing, may be necessary to accommodate existing findings. In the present article, we report the findings from three experiments designed to determine whether three separate processing factors are used in speech perception. In these experiments, English and Hindi subjects were tested in a same-different (AX) discrimination procedure. The duration of the interstimulus interval, the number of trials, and the experimental context were manipulated when testing the English-speaking subjects. The combined results from the three experiments provide support for the existence of three distinct speech-perception factors.

A continuing controversy in the area of speech perception concerns the question of whether speech perception can be best explained by positing a specialized linguistic processor (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), a generalized psychoacoustic processor (Pastore et al., 1977), or a dual-factor processor (Fujisaki & Kawashima, 1969). Researchers attempting to reconcile this controversy have tested human adults, human infants, and infrahumans to determine if they show categorical perception, trading relations, and context effects in the perception of both speech and nonspeech stimuli (see Repp, 1983, for a thorough review). Typically, results suggesting that humans respond to speech stimuli (and not nonspeech stimuli) according to discrete categories, and that they show trading relations and context effects consistent with an articulatory normalization explanation for speech (|ba|-|wa|) stimuli (Eimas & Miller, 1980; Miller & Liberman, 1979), are interpreted as providing support for specialized linguistic processing (see Liberman, 1982). Results suggesting that humans sometimes process nonspeech sounds according to discrete categories (see Pisoni, 1977, with regard to tone-onset-tone) and that infrahumans show categorical perception (Kuhl & Miller, 1975; Kuhl & Padden, 1983) in the perception of human speech sounds are interpreted as providing support for the notion that a generalized psychoacoustic processor can account for all speech-perception findings. Recent work showing that both adult (Pisoni, Carrell, & Gans, 1982) and infant (Jusczyk, Pisoni, Reed, Fernald, & Myers, 1983) listeners show context effects in the perception of nonspeech stimuli similar in form to those context effects between syllable duration and transition duration that Miller and Liberman (1979) demonstrated in the perception of [ba]-[wa] syllables is also interpreted as supporting a single-factor psychoacoustic mechanism. Finally, data indicating that subjects perceive stimuli according to phonetic categories under some testing conditions, and that they can demonstrate finer discriminative capabilities when tested in alternative procedures, suggest a dual-factor model (Fujisaki & Kawashima, 1969, 1970; Pisoni, 1973).

In dual-factor models, it is proposed that the acoustic waveform is stored in both an auditory and a phonetic code, but that the auditory code decays rapidly relative to the acoustic code (Fujisaki & Kawashima, 1969, 1970; Pisoni, 1973). Since the auditory code decays more rapidly than the phonetic code, dual factor models predict that an aoustic level of processing will be evident only when immediate comparisons between stimuli are possible. (See Studdert-Kennedy, 1973, for an explication of the rather vague model proposed by Fujisaki and Kawashima.) As such, an acoustic level of processing should be evident in experimental procedures involving a short interval between stimuli. In support, Pisoni (1973) found an inverse relationship between vowel discrimination and interstimulus interval (ISI) in a same-different (AX) discrimination task. At the shorter ISIs, subjects showed evidence of within-category auditory-level discriminations, whereas at longer ISIs only phonetic categorization was evident. Similar results have been

This work was supported by a National Science and Engineering Research Council grant (A2610) to Janet F. Werker and by NICHD Grant HD12420 to Haskins Laboratories. We thank Al Liberman for making us welcome at Haskins Laboratories and John Baressi, Ray Klein, and Bruno Repp for comments on earlier drafts of the paper and assistance in data analysis. Thanks are also extended to David Pisoni and two anonymous reviewers for constructive suggestions. We appreciate Doug Whalen's help in stimulus preparation and Gordon Troop's technical assistance. John S. Logan is currently at the Department of Psychology, Indiana University, Bloomington, Indiana. Reprint requests should be sent to Janet F. Werker, who is currently on leave from Dalhousie, and is at the Department of Psychology, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6.

reported by Crowder (1982). In addition to these effects due to the length of the ISI, research has shown that in experimental procedures that have high memory demands, such as the ABX task, subjects have access only to a phonetic code, whereas in procedures with low memory demands, access to the acoustic code is facilitated. For example, Pisoni and Lazarus (1974) showed the 41-AX procedure to be more sensitive to acoustic processing than the ABX task, and Carney, Widin, and Viemeister (1977) showed that acoustic processing is facilitated in the AX procedure relative to the ABX task.

Stimulus characteristics also influence whether auditory or phonetic processing will be used. Both Pisoni (1973) and Studdert-Kennedy (1973) speculated that it may be more difficult to demonstrate an auditory level of processing for consonant than for vowel stimuli, because the relevant acoustic cues differentiating consonants are so brief and transient, whereas those cues differentiating vowels are longer in duration and include steady-state parameters. The brief, transient cues for consonants may not be available in auditory short-term memory, especially when they are presented in the context of a longer, steady-state vowel. Such speculation is supported by research showing better within-category discrimination of truncated consonant stimuli (Tartter, 1981) and more categorical perception of shortened vowels (Fujisaki & Kawashima, 1970; Pisoni, 1973).

These results have led many researchers to advocate dual-factor models. Although theorists disagree as to whether the processing levels in a dual-factor model occur sequentially or in parallel, a considerable body of research does suggest that under some testing conditions subjects discriminate speech and speech-like stimuli according to phonetic category boundaries, but that, under other conditions, they are sensitive to the auditory information and can demonstrate more continuous discrimination functions.

In a series of cross-language speech-perception experiments, we found the single-factor and dual-factor models inadequate to account for our findings. In our previous research, we tested adult English speakers, adult Hindi and Thompson speakers, and 6-12-month-old English-learning infants on their ability to discriminate multiple natural exemplars taken from two non-English place-of-articulation distinctions. The Hindi (non-English) pair involved a contrast between the retroflex and dental place-of-articulation, and the Thompson (non-English) pair involved a contrast between glottalized velar and glottalized uvular syllables. Initial results suggested that English-learning infants aged 6-8 months can discriminate these non-English distinctions as well as native Hindi and Thompson speakers, but that English-speaking adults (Werker, Gilbert, Humphrey, & Tees, 1981) and English-learning infants aged 10-12 months (Werker & Tees, 1984a) cannot. Subsequent experiments have shown that although English speakers cannot discriminate these distinctions under most testing conditions (Tees & Werker, 1984), they can differentiate the syllables according to Hindi and Thompson phonetic categories under some testing conditions (Werker & Tees, 1984b).[1] In our previous work, the testing conditions that facilitated sensitivity to nonnative phonetic distinctions were similar to the conditions that have typically been used to demonstrate (within phonetic category) auditory level processing (cf. Pisoni, 1973). When tested in an AX procedure with a long (1,500-msec) ISI, subjects demonstrated phonemic-level processing. It was only when the ISI was shortened in the AX procedure to 500 msec that subjects showed evidence of being sensitive to nonnative phonetic distinctions (Werker & Tees, 1984b). Additionally, when tested in a category change procedure with a long (1,500-msec) ISI, subjects showed phonemic-level processing for full-length syllables. However, when truncated stimuli with much of the steady-state portion removed were employed, English subjects were able to discriminate the stimuli according to non-English phonetic category, even in the 1,500-msec ISI.

In considering these results, we argued that there might be three rather than either one or two factors in speech perception. When subjects perceive stimuli according to native-language phonological categories, they are demonstrating "phonemic" perception. When subjects show a sensitivity to phonetic distinctions that are used in some other (not their native) languages, they are using phonetically relevant (or "phonetic") perception. We argued that generalized "psychoacoustic," or "auditory," level processing is demonstrated only when subjects show a sensitivity to acoustic differences that do not correspond to phonetic boundaries that function phonologically (to contrast meaning) in any of the world's languages.

Although we raised the possibility of a three-factor model in our previous work (Werker & Tees, 1984b), and although we explored the implications of such a model, we did not provide evidence for three separate factors. Rather, we provided evidence differentiating "phonemic" from "phonetic" perception. This raised the possibility that a modified dual-factor model could account for our results. That is, "phonetic" perception might be equivalent to that which had previously been referred to as generalized psychoacoustic, or auditory, processing. The present experiments were designed to test the proposed three-factor hypothesis against a modified dual-factor model by attempting to determine whether phonemic, phonetic, and auditory processing could be differentiated as independent processing factors.

EXPERIMENT 1

The first experiment was designed to test the existence of three separate speech-perception factors by attempting to demonstrate phonemic, phonetic, and auditory processing under varying experimental conditions. In our previous work, the length of the ISI was shown to distinguish phonemic from phonetic processing. English subjects tested on the nonnative speech sounds in an AX procedure with a 1,500-msec ISI could not distinguish be-

tween the non-English phonetic categories. To these listeners, exemplars from both Hindi categories sounded like the English alveolar [ta], and exemplars from both Salish categories sounded like an English [ki]. However, subjects tested in the same procedure with only a 500-msec ISI could discriminate between exemplars from the two Hindi and from the two Salish categories better than would be predicted by chance.

In the present work, we attempted to replicate and extend this finding by testing subjects in 1,500-, 500-, and 250-msec ISI conditions. The present experiment differed from our previous work by utilizing a within-, rather than a between-, subjects design. In addition, subjects were tested only on the Hindi retroflex/dental contrast, since there was likely to be more within-category variability for these tokens than for the Thompson velar/uvular distinctions (see the stimulus descriptions in Werker & Tees, 1984b, for an explanation). Finally, in our present work, a two-choice buttonpress response, rather than a paper and pencil task, was used to record responses.

In addition to modifications in experimental procedures, subjects' responses were scored differently in the present work. In accordance with the terminology introduced by Posner and his colleagues (Posner, 1978; Posner & Mitchell, 1967), the stimuli we used were of three types: (1) physically identical (PI) pairings, (2) name-identical (NI) pairings, and (3) different (DIF) pairings. PI and NI pairings refer to two types of within-category pairings, and DIF refers to between-category pairings. PI pairings have one exemplar paired with itself; thus, there is acoustic identity between the exemplars. NI pairings include two nonidentical exemplars from within the same Hindi (non-English) phonetic category (but still within a single English phonemic category). Pisoni and Tash (1974) were the first speech researchers to test Posner's letter-matching model in speech-perception experiments. Posner gives explicit physical- or name-identity instructions to subjects in his experiments. Pisoni and Tash gave name-identity instructions. Subjects in our experiments were free to adopt their own criterion, and were thus not given explicit instructions. We used a proportion same [P(Same)] measure to gauge the perceived similarity of the three types of pairings. Presumably, if stimuli are *not* discriminated, then the P(Same) responses should be identical across the pairing types.[2]

As an extension of our previous work, one could make the simple prediction that the three hypothesized speech-perception factors would be demonstrated if subjects showed phonemic perception in the 1,500-msec condition, phonetic perception in the 500-msec condition, and auditory perception in the 250-msec condition. Phonemic perception would be indicated if subjects could not discriminate any two exemplars, that is, show equal P(Same) responses to all pairing types. Phonetic perception would be indicated if subjects treated the several exemplars from the retroflex category as different from the several exemplars from the dental category, showed high P(Same) responses to NI and PI pairings and low P(Same) responses to DIF pairings. Auditory perception would be indicated by high P(Same) responses to PI pairings, significantly lower P(Same) responses to NI, and lower yet to DIF pairings (see Figure 1). Single-factor psychoacoustic models would predict that the auditory pattern would be evident in all ISI conditions, whereas single-factor specialized linguistic models would most likely predict phonemic pattern across all ISI conditions. Dual-factor models would predict that subjects would exhibit the phonemic pattern when tested in long ISI conditions, would exhibit the auditory pattern when tested in very short ISI conditions, and would not exhibit the phonetic pattern under any conditions. Results showing three different data patterns corresponding to those shown in Figure 1 would support a three-factor model.

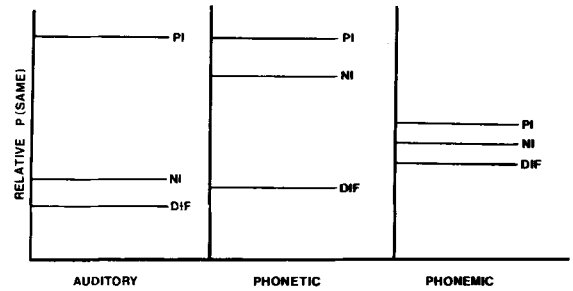

Figure 1. Idealized data patterns that would illustrate use of each of the three hypothesized speech-perception factors as indicated by the relative proportion "same" responses to PI, NI, and DIF pairings.

Method

Subjects. The subjects in Experiment 1 were 30 adults with normal hearing (15 males and 15 females), all of whom were psychology students at Dalhousie University. Participation in this experiment provided course credit. All subjects were unilingual English speakers.

Stimuli. The Hindi (non-English) place of articulation contrast that has been used in our previous research was used in this experiment. This contrasts the voiceless, unaspirated retroflex, and dental consonants /ʈ/ vs. /t/. In Hindi, the distinction between retroflex and dental stops carries phonemic significance (is used to contrast meaning). However, this distinction does not have phonemic significance in English, and both categories of consonants are typically perceived as the English alveolar phone [t].

Eight naturally produced speech syllables, four from the Hindi retroflex and four from the Hindi dental category, each followed by the neutral vowel [a] were used. Each stimulus was approximately 400 msec in duration. These eight syllables were selected from multiple repetitions (approximately 100) recorded by a native Hindi speaker. The eight final exemplars were selected for use because of their similarity in nonphonetic cues such as intonation contour, fundamental frequency, intensity, and duration both between and within categories (see Table 1 for a full description of the acoustic parameters). The four exemplars from each category gave a total of eight acoustically distinct speech stimuli labeled retroflex 1, 2, 3, and 4, and dental 1, 2, 3, and 4.

The stimuli were originally digitized on a Honeywell DDP-224 computer at Haskins Laboratories. They were subsequently redigitized and sequenced on audio tape with a PDP-11/10 computer at Dalhousie University.

Experimental materials. The tapes used in Experiment 1 were assembled using a same/different (AX) format, in which the stimuli

Table 1
Acoustic Analysis of Hindi Syllables

	Retroflex Syllables				Dental Syllables			
	ṭa 1	ṭa 2	ṭa 3	ṭa 4	ta 1	ta 2	ta 3	ta 4
Formant Frequency (Hz)								
1st Formant (F1)								
Center	720	700	660	660	630	660	640	660
2n Formant (F2)								
Starting	1660	1730	1700	1690	1430	1490	1460	1500
Center	1230	1270	1230	1300	1160	1230	1160	1180
3rd Formant (F3)								
Starting	2860	2800	2800	2900	2560	2700	2760	2560
Center	2500	2560	2530	2660	2460	2530	2500	2530
Duration F2 Transition (msec) Burst	40	60	60	42	45	60	65	42
Duration	9.35	8.65	8.65	9.35	11.7	11	10.25	12.15
Frequency Range (Hz)	1530-3530	1460-3260	1590-3200	1560-3600	1400-1730 2500-2660	1460-1933	1260-2230 2450-2900	1330-1660 2530-2860
Intensity (dB)	35	36	37	35	30	31	30	30
Intensity Peak Vowel (dB)	49	49	50	50	50	49	49	49
Pitch Contour (fall then rise)								
Starting (Hz)	135	140	135	145	145	140	140	145
Low Point	110	105	105	106	105	108	104	110
Ending	125	130	130	130	135	135	125	130

are sequenced in pairs (e.g., dental 1-dental 2). The ISI was varied to produce three separate tapes: one containing stimuli separated by a 250-msec ISI, one with a 500-msec ISI, and one with a 1,500-msec ISI. On all tapes, the time interval separating the pairs of stimuli (the intertrial interval, ITI) was 3,000 msec.

All possible pairings of the eight speech sounds were assembled in random order, resulting in the production of a 64-trial block for each tape. The tapes all contained 32 within-category pairings consisting of eight pairs of acoustically identical stimuli (PI; e.g., dental 3-dental 3); 24 pairs of stimuli sharing only a common phonetic category (NI; e.g., dental 3-dental 1); and 32 between-category pairings (DIF; e.g., retroflex 2-dental 4).

Apparatus. Stimuli were recorded and played back on a Teac A-1200 tape recorder through a Harmon-Kardom A-402 amplifier and a Luxman CS-6 speaker. Volume was adjusted to an average of 72-74 dB SPL, as measured by a B&K audiometer calibrator. Free-field rather than headphone presentation was used, for two reasons: (1) We wanted to be able to compare these results to previous (see Werker & Tees, 1984a) and future cross-language research with human infants. In our infant work, a head-turn procedure is used that requires free-field audiometry. (2) Free-field presentation is more similar to the listening conditions that characterize "everyday" speech communication. Since the present research is designed to determine what subjects do under specific testing conditions rather than as an attempt to obtain their optimal performance (as would be the case in most psychophysical experiments, and as would be more readily obtained with headphone presentation), free-field presentation was preferred. An Apple II Plus computer interfaced with a Schmidt trigger and a John Bell 6502 Board was used to control the experiment and to record the discrimination responses of subjects. The subjects were tested in an IAC sound-attenuated booth.

Procedure. An AX task was used to test subjects' discriminative abilities. In this procedure, the subjects heard two sounds and were instructed to press one of two buttons to indicate whether the sounds were the same or different. We chose the AX procedure over other discrimination tasks because it facilitates a listener's ability to distinguish subtle differences between speech sounds (cf. Carney et al., 1977). It also provided continuity with our previous studies, which had used similar stimuli and a similar procedure (Werker & Tees, 1984b). Since the purpose of this research was to determine what processing level subjects would use under different testing conditions, rather than to obtain any particular processing level, feedback was not provided during the testing session.

Each subject was tested under all three ISI conditions in a single testing session, which lasted approximately 30 min. There was a 5-min break between ISI blocks. Order of presentation was counterbalanced across the 30 subjects, resulting in six groups of five subjects each. This provided an orthogonal, blocked design which allowed for within and between comparisons of the effect of the different ISI conditions, as well as testing for possible practice effects. The dependent variable was the proportion of "same" responses for DIF, NI, and PI pairings.

Results and Discussion

The P(Same) responses for PI, NI, and DIF trials were calculated for each subject in each ISI condition. The average P(Same) responses for the three pairing types in each ISI condition are shown in Figure 2. These data were ana-

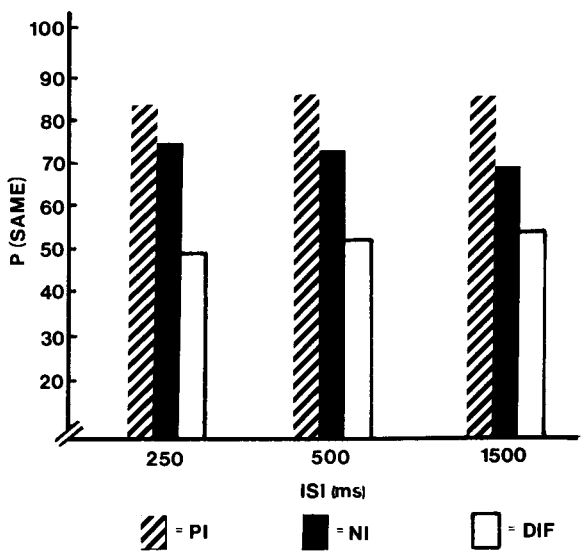

Figure 2. Proportion "same" responses for each type of pairing in the three ISI conditions (Experiment 1).

lyzed in a 3×3 within-group ANOVA. The factors were type of pairing (PI, NI, and DIF) and ISI (250, 500, and 1,500 msec). Since order of presentation of ISI condition and sex were both counterbalanced, these variables were pooled in the first analysis. A significant main effect for type of pairing [$F(2,58) = 48.19$, $p < .001$] was obtained in this analysis. There was no main effect for ISI, and no significant interaction.

Multiple comparisons using the Tukey method revealed that there was no significant difference between the means of the PI and NI pairings in either the 250-msec or the 500-msec ISI conditions ($p > .05$). The means of the P(Same) to the NI and PI pairings, as compared with the DIF pairings, did, however, differ significantly at these ISIs ($p < .05$). In the 1,500-msec ISI condition, the means for all three types of pairings were significantly different from each other ($p < .05$).

The data pattern across the ISI conditions does not match our predictions or those of the other models discussed earlier. It was particularly surprising that there was no difference between the NI and PI pairings in the 250-msec ISI, and that the difference appeared only in the 1,500-msec ISI. This data pattern suggests that subjects were using a phonetic processing strategy in the two shortest ISI conditions, and were possibly using both a phonetic and an auditory strategy in the 1,500-msec condition.

Further analyses were carried out on the data to assess the possibility that in a within-groups design the potential effects of ISI were mitigated by the context in which the ISI conditions were presented. This possibility was examined in two ways: (1) with an evaluation of the effect of position, and (2) with an analysis of the effect of order. Position refers to the three possible places each ISI block could occupy in the presentation sequence—first, second, or third. The position analysis evaluated whether practice resulted in enhanced discrimination abilities. This analysis consisted of a mixed-group ANOVA in which the between-subjects factor was position of ISI presentation (three positions—first, second, or third) and the within-subjects factors were type of pairing (PI, NI, and DIF) and ISI (250, 500, and 1,500 msec). As in the previous analysis, a significant main effect for type of pairing was obtained [$F(2,54) = 58.58$, $p < .001$]. There was also a main effect for position [$F(2,27) = 7.78$, $p < .005$], as well as a significant interaction of position and type of pairing [$F(4,54) = 3.63$, $p < .05$]. Post hoc comparisons indicated that subjects perceived PI pairings as "same" for a consistently high proportion of the trials over all three blocks, whereas their "same" responses to NI and particularly to DIF pairings declined as a function of each successive block. This suggests that practice may enhance discrimination performance for all types of pairings, but that the effect may be greater for DIF than for NI pairings. This again suggests the use of a phonetic processing strategy (refer to Figure 1). The effect of ISI was negligible.

The final analysis of the above data examined the six different orders of stimulus presentation in an effort to determine if this variable affected subjects' performance. The six orders of presentation (each block with respect to ISI: 250-500-1500, 250-1500-500, 500-250-1500, 500-1500-250, 1500-250-500, and 1500-500-250) constituted the between-subjects factor; ISI and type of pairing were the within-subjects factors. As before, a significant main effect for type of pairing was obtained [$F(2,48) = 47.13$, $p < .001$]. In addition, there were two significant interactions: a two-way and a three-way interaction among order of presentation, type of pairing, and ISI [$F(20,96) = 2.97$, $p < .001$].

These interactions suggest that order of presentation interacted with both ISI and pairing type in affecting subjects' scores. In examining the data, it appeared that subjects could not easily switch processing strategies when shifting from one ISI condition to another. In most cases, performance in an ISI block was better if that block had been preceded by a 250-msec condition. These complex interactions indicate that the within-subjects design was not appropriate for assessing the proposed three factors.

Typically a within-subject design with blocked ISIs is preferred in experimental work because it allows for more power in the data analysis, and is thought to facilitate optimal performance. The design of Experiment 1 was chosen for those reasons. However, when feedback is not provided and the experimenter is interested solely in determining how ISI affects the level of processing, a blocked, within-subjects design may not be optimal. Without feedback to influence adoption of a particular processing strategy within a block, subjects appear to have relied on information obtained and strategies developed in prior testing blocks.

EXPERIMENT 2

The complex interactions that resulted from the within-subjects design used in Experiment 1 may have masked important data patterns. Experiment 2 was designed to eliminate this problem by testing subjects on their ability to discriminate the several Hindi retroflex and dental exemplars in a between-subjects design. Three groups of subjects were tested: one group in the 250-msec ISI condition, one in the 500-msec ISI condition, and one in the 1,500-msec ISI condition. Three other changes were made in Experiment 2: (1) Since position interacted with pairing type in Experiment 1, many more testing trials were used in Experiment 2 [480 vs. 192 (containing only 64 for each ISI)]; (2) the proportion of PI pairings was increased to be equal to that of NI pairings in an effort to obtain clearer evidence distinguishing auditory from phonetic perception; and (3) a measure of reaction time (RT) was included in addition to type of response.

These changes rendered the experiment more comparable to work completed by other researchers. For example, one of the first demonstrations of sensitivity to within-category distinctions was provided by Pisoni and Tash (1974). Using a synthetic [ba]-[pa] speech continuum, Pisoni and Tash showed that RT provides a measure of

subjects' certainty in discriminating within- and between-category speech stimuli. Pisoni and Tash found that subjects took significantly longer to respond "same" to stimuli that shared only a common phonetic category (similar to our NI pairings) than they did to pairs of stimuli that were acoustically identical (as in our PI pairings). Similarly, subjects responded "different" more slowly to phonetically distinct pairs of stimuli that were only two steps apart on a synthetically produced continuum than they did to phonetically distinct stimuli that were four to six steps apart. Similar results were obtained by Howell and Darwin (1977) with regard to the [ba]-[da] place-of-articulation continuum. The number of trials was increased in this experiment because it is known that increasing the number of trials can significantly improve performance (cf. Samuel, 1977). By increasing trials without providing feedback, it can be determined whether subjects adopt and perfect a particular processing strategy in a certain ISI condition, or whether they shift processing levels as a function of increased practice.

Method

Subjects and Stimulus materials. Thirty adults, 15 females and 15 males, served as subjects. Ten subjects were tested in each of the three ISI conditions. All subjects were unilingual English speakers with no history of hearing problems. Twenty subjects received credit in an introductory psychology course at Dalhousie University for participation in this experiment, and the remainder were paid $4 for their participation.

The four retroflex and four dental syllables used in Experiment 1 were used in this experiment to construct AX discrimination tapes. For each of three ISI conditions, the pairings were randomized into five blocks of 96 trials. In all cases, there was a 3,000-msec ITI. Each 96-trial block contained 48 within-category trials and 48 between-category (DIF) trials. The within-category trials contained 24 PI stimulus pairs (e.g., dental 4-dental 4) and 24 NI stimulus pairs (e.g., retroflex 1-retroflex 3).

Procedure and Apparatus. The apparatus used in Experiment 2 was identical to that used in Experiment 1. In addition, the Apple II Plus computer was programmed to measure subjects' reaction times for making a same/different judgment. Reaction times were measured in milliseconds from the onset of the second stimulus in each pair.

The procedure in this experiment was similar to the procedure used in Experiment 1. The 10 subjects in each ISI group were tested on all five trial blocks using the buttonpress response in an AX task. After presentation of each 96-trial block there was a 5-min break to compensate for fatigue and adaptation effects. The task required approximately 1 h.

As in Experiment 1, subjects' responses were scored using the P(Same) measure for DIF, NI, and PI pairings. In addition, reaction times to both "same" and "different" responses were recorded (as in Pisoni & Tash, 1974). Reaction times for "same" responses were included in our experiment to clarify the results in case the P(Same) responses were ambiguous. If P(Same) responses provided clear results, RT data would provide converging evidence in support of these results. Reaction times for "different" responses were recorded to determine whether the pattern of results for "different" responses was similar to that obtained for "same" responses.

Results and Discussion

In the first analysis, the P(Same) responses for the three types of pairings and the three ISIs were averaged across the five blocks of stimuli for each subject. These results are illustrated in Figure 3. The data were analyzed via a 3×3 mixed ANOVA in which the between-subjects factor was ISI (250, 500, and 1500 msec), and the within-subjects factor was type of pairing (PI, NI, and DIF). This analysis yielded two significant main effects: ISI [$F(2,147) = 3.79$, $p < .05$] and type of pairing [$F(2,294) = 399.81$, $p < .001$]. There was also a significant interaction between ISI and type of pairing [$F(4,294) = 14.64$, $p < .001$]. Multiple comparisons using the Tukey method revealed that in all three ISI conditions there was a significant difference between the P(Same) response for each type of pairing, but that the effect was less pronounced at the longest ISI ($p < .01$ at 250 and 500 msec; $p < .05$ at 1,500 msec).

The next analysis measured the effect of practice on subjects' performance in a 3×3×5 mixed ANOVA. The between-subjects factor was ISI (three levels), and the two within-subject factors were type of pairing (PI, NI, and DIF) and position (five positions—one for each block of stimuli). There were two significant main effects: type of pairing [$F(2,54) = 171.05$, $p < .001$] and the other between type of pairing and position [$F(8,216) = 11.25$, $p < .001$]. The mean P(Same) for each type of pairing at each position is shown in Figure 4. Although the second-order interaction between ISI, type of pairing, and position did not reach significance ($p = .13$), the data pattern suggests that the P(Same) responses to the three pairing types did tend to separate at a dissimilar rate at the

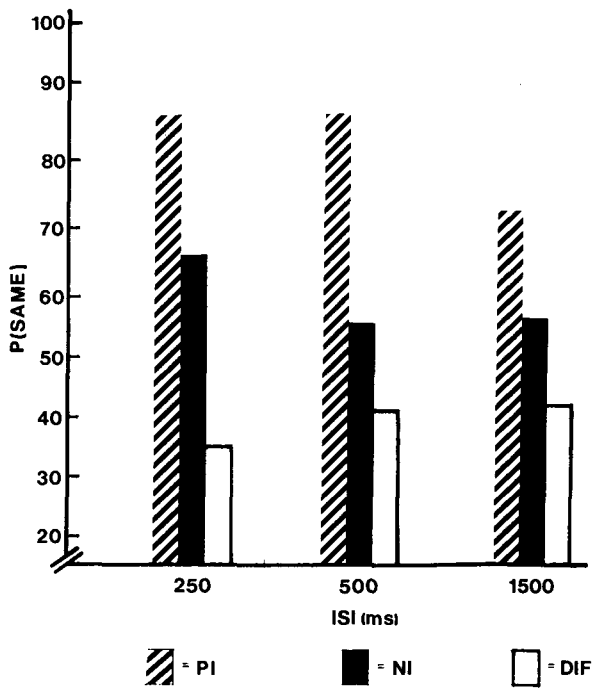

Figure 3. Proportion "same" responses for each type of pairing in the three ISI conditions (Experiment 2).

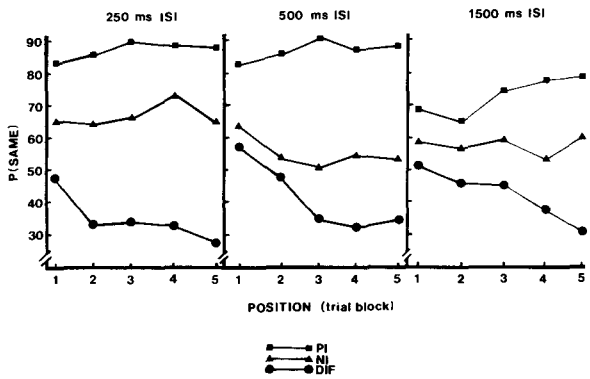

Figure 4. Proportion "same" responses as a function of position for each type of pairing in each ISI condition.

different ISIs.[3] Post hoc comparisons using the Tukey method were used to compare performance between the three pairing types in each position at each ISI. Results showed that in the 250-msec ISI condition, the P(Same) responses to the PI, NI, and DIF pairings were significantly different from each other ($p < .05$) even in the first position. In the 500-msec condition, P(Same) responses to PI pairings were significantly greater than those to NI and DIF pairings, but NI and DIF pairings did not diverge significantly until the third block of trials.[4,5] Finally, in accordance with our initial predictions, the P(Same) responses to the three pairing types in the 1,500-msec ISI condition were not significantly different until the third block of trials, at which time all three (PI, NI, and DIF) were significantly different. This suggests that the three ISI conditions affect performance differentially, and is consistent with previous work showing that ISI affects the use of phonetic vs. auditory levels of processing (Crowder, 1982; Pisoni, 1973). It is the first demonstration, however, of the effect of ISI of three processing levels. The data pattern obtained in the first two trial blocks closely matches the predictions in support of a three-factor model in both the 250-msec condition and the 1,500-msec condition. Subjects appear to be using an "auditory" factor in the 250-msec condition and a "phonemic" factor (at least initially) in the 1,500-msec condition (refer to Figure 1 for clarification). In the first two blocks in the 500-msec condition, it appears that subjects show both auditory *and* phonemic processing, since the P(Same) responses are very high for the PI pairings but not for the other pairing types. In the final three trial blocks in the 250-msec condition, the data pattern approximates that predicted for phonetic processing.

An analysis comparing the average reaction time (RT) for each type of pairing in each ISI condition was also done. These RTs were then subjected to a $3 \times 3 \times 2$ mixed ANOVA in which ISI (250, 500, and 1,500 msec) was the between-subjects factor and type of pairing (PI, NI, and DIF) and type of response ("same" vs. "different" RT) were the within-subjects factors. There was a significant main effect for type of pairing [$F(2,54) = 7.20$, $p < .002$] and two significant interactions: one between type of response and type of pairing [$F(2,54) = 55.46$, $p < .001$] and the other among ISI, type of response, and type of pairing [$F(4,54) = 5.97$, $p < .001$]. The mean RT for each type of pairing in each ISI condition is shown in Figure 5. "Same" RTs are shown in (a) and "different" RTs in (b). This graph clearly illustrates the nature of the interactions. Subjects responded "same" fastest to PI pairings, followed by NI and DIF pairings. Conversely, subjects responded "different" fastest to DIF pairings, followed by NI and PI pairings. The extent of this effect varied with the ISI. As with the P(Same) results, this spread between the RT values for the three types of pairings decreased as the ISI increased. Thus, the RT measure provided converging evidence to support the data pattern obtained with the proportion "same" responses.

Similar RT results were reported by Pisoni and Tash (1974) when measuring the RT response to synthesized stimuli in an AX discrimination task. Pisoni and Tash were testing a model of speech perception based on Posner's work with letter-matching experiments (see Posner, 1978) in which comparisons between physically identical stimuli required less processing time than comparisons between stimuli sharing more abstract similarities (that is, NI pairings). The RT data we obtained in Experiment 2 replicate their findings. The P(Same) data reported by Pisoni and Tash differ markedly from our results, presumably because Pisoni and Tash gave their subjects explicit NI instructions.

EXPERIMENT 3

This experiment was conducted to determine how native Hindi speakers would respond to these Hindi syllables when tested in an AX task at the longest ISI. This would provide data on the natural categories used by native speakers, and would provide us with important comparative data to use in attempting to understand the data

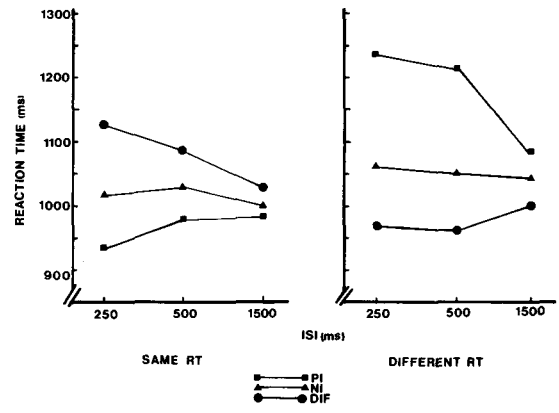

Figure 5. Average reaction time for "same" and "different" responses for each type of pairing in the three ISI conditions.

pattern supplied by English speakers. A single, long ISI was chosen for use in this experiment for two reasons: (1) It was already known from Experiment 2 that English-speaking subjects have access to an auditory processing level, and it can be assumed that Hindi-speaking subjects would also have access to this level; and (2) there was only a small sample of available native Hindi speakers in the Halifax area. Given the interactions introduced by the within-groups design in Experiment 1, it was felt that all Hindi subjects should be tested in the ISI condition most likely to tap linguistic categories rather than auditory sensitivities. Finally, since we were interested in how native speakers naturally respond to these syllables, the subjects were given only one block of trials.

This experiment was designed to compare the categories used by speakers of two different language groups, and was in no way meant to test possible alternative processing levels among the Hindi listeners. These results can be compared with those obtained in the first trial block in Experiment 2.

Method

Subjects were four native Hindi speakers presently living in Halifax, Nova Scotia, three males and one female between the ages of 25 and 45. All subjects reported that their first language was Hindi, although three of the four spoke additional Indian languages as well. Only subjects who had lived in India at least through their mid-20s were selected for participation in this study. In addition, two of the subjects had lived outside of India for less than 2 years; the other two subjects had been back to India for extended visits every few years and still spoke Hindi much of the time at home and in their involvement with the Halifax Hindu religious community. All four subjects also spoke English.

The subjects were tested in the AX procedure with the same stimuli as used in the two preceding experiments. As noted above, all subjects were tested in the 1,500-msec ISI condition, and were given only one block of testing trials.

Results and Discussion

The P(Same) responses for each type of pairing were calculated and analyzed in a one-way ANOVA. This yielded a significant effect for type of pairing [$F(2,12) = 380.79$, $p < .001$]. This result was due solely to the classification of PI and NI pairings as "same" an equally high proportion of the time ($\overline{X}PI=99.16$, $\overline{X}NI=94.16$); the DIF pairings were consistently perceived as "different" ($\overline{X}DIF=9.8$). Those data map perfectly onto the predicted "phonetic" pattern illustrated in Figure 1. This clearly indicated that Hindi subjects were using a single processing strategy in perceiving these syllables—a strategy corresponding to what would be called "phonetic" processing for the English subjects tested in Experiments 1 and 2 but which represents phonemic processing for the Hindi listeners. The similarity in responses to PI and NI pairings indicates no evidence of an auditory level of processing among the Hindi listeners when tested in this ISI condition and given only one block of trials. Rather, it suggests that the most available processing level for native Hindi speakers when tested in a high-uncertainty procedure (high memory demands, little practice) is that level, and *only* that level, corresponding to the linguistically relevant phonological categories used in the speaker's native language. This is isomorphic to the results obtained with English listeners in the first 1,500-msec trial block in Experiment 2, wherein they showed a data pattern corresponding to English (rather than Hindi) phonological categories.

CONCLUSIONS

Let us consider these findings in the light of current models of speech perception. Single-factor models suggesting that all speech perception findings can be explained by specific phonetic processing mechanisms (see Liberman, 1982) have typically used the term "phonetic" to refer to that which we call "phonemic" perception (see discussion in Werker & Tees, 1984b). If these models were specifically addressing phonemic perception, they would predict that subjects would respond "same" equally often to all three pairing types in all ISI conditions, since all eight speech syllables were classified according to a single phonemic category in English and are thus equivalent. If these models were addressing what we call "phonetic" perception, they would predict that subjects would respond "same" to NI and PI pairings equally often, and significantly more than they would to DIF pairings. The subjects tested in the 1,500-msec condition in Experiment 2 showed a data pattern consistent with the prediction for "phonemic" perception in the first two trial blocks (see Figure 1 for clarification). Thus, it appears that, without practice, subjects rely on phonemic categories when responding to speech syllables in paradigms that have high memory requirements. The strong prediction of "phonetic" processing was not supported in the data collected from English listeners in Experiment 2, but their data were in the predicted direction in the last three trial blocks in the 250-msec ISI condition, in which the difference between proportion "same" responses to NI and PI pairings was less than that between NI and DIF pairings. Support for the phonetic level was provided in Experiment 1 in both the 250- and 500-msec ISI conditions. Clear support for this universal "phonetic" level was supplied by the Hindi subjects in Experiment 3.

A single-factor psychoacoustic theory (see Schouten, 1980) would predict that the relative proportion of "same" responses would vary as a function of acoustic dissimilarity in all positions in all ISI conditions (although there might be some effects due to decay in the longer ISIs, and some improvements due to practice). This prediction is supported by the data in Experiment 2 in the first two blocks of the 250-msec ISI condition, but not until the third trial block in the other two conditions. It is not supported by two of the three ISI conditions in Experiment 1 and not at all by Experiment 3. It thus appears that a single-factor psychoacoustic explanation is inadequate to explain the current data pattern.

A dual-factor model would predict that subjects would respond "same" equally often to all pairing types under

testing conditions that encourage reliance on the phonetic code (would show what we call "phonemic" perception), and would respond differentially according to acoustic dissimilarity under conditions that facilitate use of the auditory code (see Fujisaki & Kawashima, 1969, 1970; Pisoni, 1973). According to a dual-factor model, when testing was done without considerable practice in longer ISI conditions, auditory information in short term-memory would have decayed, forcing subjects to classify stimuli according to the more robust language-specific (phonemic) codes (Crowder, 1982; Pisoni, 1973). The predictions from a dual-factor model are partially supported by the findings in Experiments 1 and 2. However, the data from all three experiments together, indicating the presence of a phonetic level as well a phonemic and auditory levels, suggest that dual-factor models may also be inadequate to explain the data pattern.

The results of these three experiments thus provide support for the hypothesis that subjects can use three distinct processing strategies when responding to speech syllables. Evidence for these processing strategies is dependent upon task conditions. Subjects can (1) classify the syllables according to familiar phonemic categories, (2) show a perceptual sensitivity to nonnative, phonetically relevant category boundaries, and (3) discriminate syllables on the basis of any acoustic variability between individual exemplars. These findings have important implications for current models of speech perception. By raising the possibility that there may be three, rather than one or two factors involved in speech perception, these results mitigate against the argument that all speech perception data can be explained by a generalized psychoacoustic mechanism. It is clear, at least in the adult (also see Jusczyk, 1984), that under task conditions that are most similar to those used in everyday oral communication (long intervals between repetitions of the same exemplar; high memory demands) subjects rely on a language-specific phonemic processing strategy. That is, they classify syllables according to the phonological categories used to contrast meaning in their native language. Under these task conditions, English and Hindi adults show little sensitivity to any acoustic variability. However, under other task conditions (short ISI and practice) there is clear evidence for an auditory processing level.

These experiments provide evidence for three processing strategies, and show that phonemic perception is clearly distinct from auditory perception. These experiments also provide clear support for an intermediate, phonetically relevant level of perception. The experiments do not, however, explain the derivation of either phonetic or phonemic processing. It is not clear whether phonetically relevant perception is a function of a specific linguistic processor or the result of second-order auditory factors resulting in perceptual classification on the basis of physical similarity. Also, it is not clear whether phonemic processing is based on a modification of innately determined universal phonetic sensitivities or is a reflection of learned (auditory based) linguistically relevant categories. Further research using different testing procedures, different populations (infants and young children), and different stimuli is needed to answer these questions.

REFERENCES

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961-970.

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **8**, 153-162.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science*, **209**, 1140-1141.

Fujisaki, H., & Kawashima, T. (1969). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute* (Tokyo), **28**, 67-73.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute* (Tokyo), **29**, 207-214.

Hays, W. L. (1973). *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston.

Howell, P., & Darwin, C. J. (1977). Some properties of auditory memory for rapid formant transitions. *Memory & Cognition*, **5**, 700-708.

Jusczyk, P. W. (1984). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion*. Hillsdale, NJ: Erlbaum.

Jusczyk, P. W., Pisoni, D. B., Reed, M. A., Fernald, A., & Myers, M. (1983). Infants' discrimination of the duration of rapid spectrum changes in nonspeech signals. *Science*, **222**, 175-177.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosic consonants. *Science*, **190**, 69-72.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, **73**, 1003-1010.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist*, **37**, 148-167.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Pastore, R. E., Ahroon, N. A., Buffuto, K. J., Friedman, C., Puleo, J. S., & Fink, E. A. (1977). Common-factor model of categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 676-696.

Pisoni, D. B. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 297-314.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and non-speech signals. *Perception & Psychophysics*, **34**, 314-322.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, **15**, 285-290.

Posner, M. (1978). *Chronometric explorations of mind*. Hillsdale, NJ: Erlbaum.

POSNER, M., & MITCHELL, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, **74**, 392-409.

REPP, B. H. (1983). Categorical perception: Issues, methods and findings. In N. L. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10). New York: Academic Press.

SAMUEL, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics*, **22**, 321-330.

SCHOUTEN, M. E. H. (1980). The case against a speech mode of perception. *Acta Psychologia*, **44**, 71-98.

STUDDERT-KENNEDY, M. (1973). The perception of speech. In T. A. Sebok (Ed.), *Current trends in linguistics* (Vol. 12). The Hague: Mouton.

TARTTER, V. C. (1981). A comparison of the identification and discrimination of synthetic vowel and stop consonant stimuli with various acoustic properties. *Journal of Phonetics*, **9**, 477-486.

TEES, R. C., & WERKER, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate nonnative speech sounds. *Canadian Journal of Psychology*, **38**, 579-590.

WERKER, J. F., GILBERT, J. H. V., HUMPHREY, K., & TEES, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, **52**, 349-355.

WERKER, J. F., & TEES, R. C. (1984a). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, **7**, 49-63.

WERKER, J. F., & TEES, R. C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866-1878.

NOTES

1. This is similar to Pisoni, Aslin, Perey, and Hennessy's (1982) recent findings, which showed that English listeners could easily learn to differentiate three categories of voice onset time. Our work suggests that this is more difficult, but still possible, in the case of nonnative place of articulation distinctions.

2. In our previous work, an A' analysis was used because we were attempting to see whether subjects could discriminate the syllables according to Hindi phonetic categories. In the present work, we wanted to see if different task conditions would encourage different processing strategies. Hence, there was no absolute right or wrong way to respond, making it inappropriate to use a signal detection analysis. For the same reasons, responses were not compared with chance, since the chance level would vary depending upon processing strategy.

3. It is recognized that a p value of .13 is not significant. However, statisticians have pointed out that the major problem in concentrating solely on avoiding alpha errors is that it precipitates many potential beta errors and leads to the premature (and often incorrect) rejection of new and interesting hypotheses. Thus, it is recommended that the researcher proceed to post hoc tests when he/she perceives regularities in the overall data pattern regardless of significance (cf. Hays, 1973, p. 582).

4. These results are similar to those obtained by Werker and Tees (1984a, 1984b) and in Experiment 1. However, in those experiments, there was evidence for nonnative (phonetic) discrimination in the 500-msec condition for DIF pairings within 126 and 64 trials, respectively. In the present experiment, this effect is not evident until close to 300 trials. The discrepancy may be explained by the greater proportion of PI pairings' providing a context effect in this experiment.

5. This data pattern is corroborated by a block-by-block analysis of RT responses. The RT to NI and DIF pairings is virtually identical in Block 1 and much slower than that to PI pairings; RTs to NI and DIF begin to spread apart in Block 2, foreshadowing the P(Same) pattern evident in Block 3.

(Manuscript received July 2, 1984;
revision accepted for publication December 27, 1984.)