Thomas and Legge's matching hypothesis for detection and recognition tasks: Two tests

A. E. DUSOIR

City of London Polytechnic, Whitechapel High Street, London E1 7PF, England

Thomas and Legge have proposed a *matching hypothesis* for two-alternative detection and recognition tasks. Given symmetric payoff, S is supposed to match his unconditional response probabilities to the presentation probabilities. More generally, even when payoff is asymmetric, S is supposed to keep his unconditional response probabilities constant across different discriminability levels. Time series analysis is applied here to the response sequences from two auditory amplitude recognition tasks to test both hypotheses. Experiment I obtains a single sequence from each of 16 Ss. Experiment II obtains two sequences at different discriminability levels from each of 24 Ss. Departures from the matching hypothesis are extremely gross in both cases. In addition, Experiment II shows clear changes in the response probabilities across discriminability conditions, though these are not systematic in direction across Ss.

Thomas and Legge (1970), generalizing earlier suggestions (Parks, 1966; Creelman & Donaldson, 1968), have proposed a *matching hypothesis* for detection and recognition situations. Suppose that one of two stimuli, s_1, s_2 , is presented on each trial with probability $(1 - \gamma)\gamma$ and followed by one of two responses, r_1, r_2 ; and suppose that payoff, o_{ij} , is made contingent on the joint occurrence of s_i and r_j . Their proposal is that S makes $p(r_2)$ equal to γ , provided the payoff matrix is symmetric ($o_{11} = o_{22}, o_{12} = o_{21}$). More generally, $p(r)_2$ is supposed to depend only on γ and payoff and to be constant over discriminability. Thus, letting $p(r_2)/\gamma = k$,

$$p(r_2 | s_2) = k - [(1 - \gamma)/\gamma]p(r_2 | s_1) \quad (1)$$

must always hold as γ and payoff are held constant and discriminability is varied; and when payoff is symmetric, Eq. 1 must hold with $k = 1$.

Evidence for the matching hypothesis is reviewed in detail elsewhere (Dusoir, in press). Creelman and Donaldson (1968) and Thomas and Legge (1970) each quote results from very experienced Ss run with trial-by-trial feedback. Neither paper, however, gives any formal test of fit, and in each case there seem to be systematic, if small, deviations from the matching hypothesis (see Creelman & Donaldson, 1968, Fig. 2, and Thomas & Legge, 1970, Fig. 1). Parks (1966) applies the proposal to the recognition memory performance of naive Ss working without feedback. Unfortunately, he uses a test of fit suggested by Grant (1962) that compares the correlation between observed and predicted group means with the intragroup variance. This allows large intragroup

variance to be mitigated by a high correlation between observed and predicted group means. Parks's results, therefore, show that *group mean response rates* tend to obey the matching prediction, whereas Thomas and Legge's proposal is meant to apply to individual Ss. An adequate formal test of the matching hypothesis is surely overdue.

A similar matching hypothesis, of course, has long been discussed in the related field of probability learning [Lee (1971) gives a summary]. In this context, Norman (1971) has used *time-series analysis* to demonstrate very significant differences in response rate across Ss even though group data show excellent probability matching. A similar method of analysis is applied here to two signal-recognition tasks. An important advantage of the analysis is that it makes only minimal assumptions about the stochastic process generating the data; thus it does full justice to Thomas and Legge's wish to avoid specific assumptions about how and whether the response probabilities vary from trial to trial. A brief account of the method is given below; for a full discussion, see Norman (1971, 1972) and Anderson (1971).

It should be noted that Thomas and Legge do not limit their model to the case where Ss are practiced and trial-by-trial feedback is given. On the contrary, they consider the model especially likely to apply to naive Ss working without feedback. This, therefore, is the case considered in this paper. Separate tests of the model's applicability to practiced Ss, working with and without feedback in both detection and recognition tasks, will be reported later.

METHOD OF ANALYSIS

Suppose, for a single S, that on the i th trial,

Thanks are due to W. Donnelly and R. Brown for technical assistance.

$0 \leq i \leq n$, a random variable, X_i , takes the value 0 or 1, where 0 corresponds to r_1 , 1 to r_2 . Disregarding the initial (probably transient) behavior, it is assumed that the shifted process $X_m, X_{m+1}, \dots, X_n$ is *second-order stationary*; that is, it is assumed that the expectation $E(X_j) = \mu$ and the autocovariance of lag k $\text{cov}(X_j, X_{j+k}) = c(k)$ both exist and do not depend on j . [This assumption is weaker than it might appear; the suspicious reader should consult the numerous examples in Anderson (1971, Section 7.2), many of which are agreeably counterintuitive.] Let

$$N = n - m + 1$$

$$\bar{X} = \sum_{j=m}^n X_j / N$$

$$\hat{c}(k) = (1/N) \sum_{j=m}^{n-k} (X_j - \bar{X})(X_{j+k} - \bar{X})$$

$$\hat{\sigma}^2 = \hat{c}(0) + 2 \sum_{k=1}^K \hat{c}(k). \quad (2)$$

Thus N gives the number of trials, $\bar{X}$, the sample mean, and $\hat{c}(k)$ the sample autocovariance of lag k . $\hat{\sigma}^2$ plays the same role as the sample variance in classical statistics. K in Eq. 2 is chosen so that $c(k) = 0$ for $k > K$. Norman (1971) shows that the generalized t statistic

$$t = \frac{(\bar{X} - \mu)\sqrt{N}}{\hat{\sigma}} \quad (3)$$

is asymptotically standard normal as $N \rightarrow \infty$; sufficient conditions for this are given in Ibragimov (1962) and Anderson (1971). If t_j denotes the value of t in Eq. 3 for the j th of I Ss, then

$$T = \sum_{j=1}^I t_j / \sqrt{I}$$

is also standard normal for a sufficiently large N (Feller, 1971, p. 46) and

$$\chi^2 = \sum_{j=1}^I t_j^2$$

is asymptotically chi square with I degrees of freedom (Mood & Graybill, 1963, p. 227), the T statistic tests for the $(\bar{X} - \mu)$ being large and *systematic in sign across Ss*, and the χ^2 statistic tests for the $(\bar{X} - \mu)$ being too large, *regardless of sign*.

Often there is no sound theoretical basis on which to choose K in Eq. 2, nor any reason to assume identity of K across Ss; this is true here, as in Norman (1971). Fortunately, firm conclusions can be reached, in both cases, because the argument proves to be quite

insensitive to choice of K . A usefully modest statistic, from this point of view, is $\chi_{\min}^2$ ($0 \leq K \leq M$), defined analogously to

$$\chi^2 = \sum_{j=1}^I t_j^2,$$

except that K in Eq. 2 is chosen for each S separately, in the range $0 \leq K \leq M$, so as to maximize $\hat{\sigma}^2$ and minimize $|t|$ for that S . M , here, is chosen informally as a liberal upper bound to K ; M cannot simply be chosen as large as possible, since K must be small relative to N for the estimator, Eq. 2, to be consistent (Norman, 1971, p. 447). Thus $\chi_{\min}^2$ ($0 \leq K \leq M$) provides a (very conservative) test for too large deviations $(\bar{X} - \mu)$ without committing the argument to any particular choice of K in Eq. 2, while allowing individual protocols to differ in their covariance structure and in σ^2 .

Extension to the case where each S generates two sequences, with no carry-over, is straightforward. The analogue of Eq. 3 is

$$t = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)}}, \quad (4)$$

where N_i , X_i , μ_i , $\hat{\sigma}_i^2$ denote the number of trials, mean, expectation, and σ^2 estimate for the i th sequence. t in Eq. 4, like t in Eq. 3, is asymptotically standard normal as N_1 and $N_2 \rightarrow \infty$ (Norman, 1971); thus the definition, asymptotic distribution, and use of T and χ^2 are exactly as above, except that one is now making inferences about $(\mu_1 - \mu_2)$ rather than μ . The $\chi_{\min}^2$ statistic, in this case, is defined so as to allow K to be chosen separately for each sequence, not just for each S .

EXPERIMENT I

Method

Sixteen Ss were run in a signal recognition task, $\gamma = 0.7$, with symmetric payoff $\alpha_{11} = \alpha_{22} = +1p$ (one English penny), $\alpha_{12} = \alpha_{21} = -1p$, and no feedback. A single response sequence is taken from each S and used to test the matching hypothesis, $\mu = \gamma = 0.7$.

Subjects. Ss were 16 staff or students at the City of London Polytechnic, naive to the present task and with no recent exposure to other detection or recognition tasks.

Task. S heard continuous broadband noise through earphones. Every 4 sec, a clearly audible, 1,000-cps sinusoid was superimposed for 0.25 sec. The amplitude of the sinusoid was greater on s_2 trials than on s_1 trials. S responded "loud" (r_2) or "soft" (r_1) after each tone by means of two keys; assignment of keys to hands was balanced randomly. Ss took part in three 100-trial blocks, with a 2-min interblock break. Each block was preceded by four examples of s_1 and s_2 in fixed alternating order. Time series analysis is applied below to the last 85 trials of the third block.

Stimuli. Trial sequences were determined by random numbers generated on a computer by applying the MacLaren-Marsaglia algorithm to two linear congruential sequences (Knuth, 1969). Separate unconstrained binomial sequences were used for each S . Overall s_2 frequency was 0.701.

The 1,000-cps sinusoids were generated by a phase-shift oscillator, gated by a solid state gate using F.E.T.s, and mixed with broadband noise from a reverse-biased diode. Two timers (Behavioural Research and Development, Ltd.) were used to control the signal gate. Noise and sinusoid were passed via a Ferrograph amplifier to Koss ESP6 earphones (binaural presentation). The s_1 and s_2 signal levels were 0.475 and 0.575 V rms measured at the earphones, corresponding to 74 and 75.7 dB SPL. The background noise, measured over the bandwidth 40-10,000 cps, averaged 0.650 V on an rms meter calibrated for sinusoids. Its approximate sound pressure level was 76.5 dB. Fluid-filled ear cushions and a sound-attenuating booth gave more than 40 dB attenuation of external noise.

Instructions. S was given a full account of the trial cycle, previews, and payoff matrix. The instructions about γ were: "The probability that the louder tone will be presented is 0.7. The probability that the softer tone will be presented is 0.3. The sequence of trials has been decided by using random number tables; this means, in practice, that on about 70% of the trials the loud tone will, in fact, be presented, and it means also that there is no regular pattern in the sequence of trials which you can use to improve your performance." (At the end of the experiment, Ss were asked what γ had been; all replied correctly.) Ss were invited to earn as much money as possible.

Results

Table 1 gives $p(r_2)$, $p(r_2 | s_2)$, and d' for each S, calculated over the last 85 trials of Block 3.¹ d' is calculated as $Z_2 - Z_1$, where Z_i is the normal deviate corresponding to $p(r_2 | s_i)$. Note that the variability of the d' estimates is bound to be large, given the paucity of s_1 trials. Group mean $p(r_2)$ was 0.648, somewhat less than the matching hypothesis prediction. Thirteen Ss showed better than chance performance in the sense that $d' > 0$ ($p = .011$, one-tailed sign test); this shows that Ss were not being run on a covert probability learning task. On the other hand, no S showed perfect discriminability.

Even low-order $\hat{c}(k)$ are very small relative to $\hat{c}(0)$: mean $\hat{c}(0)$ for the 16 Ss was 0.219, and the standard deviation was 0.028; this is to be compared with mean $\hat{c}(k)$ for $1 \leq k \leq 20$, shown with their standard deviations in Fig. 1. The $\hat{c}(k)$ are initially negative and

Fig. 1. Mean $\hat{c}(k)$, $1 \leq k \leq 20$, for the 16 Ss of Experiment I. (The vertical bars show ± 1 standard deviation.)

rapidly approach zero; thus the $\hat{\sigma}^2$ (see Eq. 2) tend to decrease as the $\hat{c}(k)$, $k > 0$, are included, leading to an increase in the significance of the deviations $(X_i - \mu)$.

The deviations $(X_i - \mu)$ are much too large. Taking K in Eq. 2 to be anywhere in the range $0 \leq K \leq 20$ gives $\chi^2 > 66.7$ ($p < .001$). As Table 1 suggests, the deviations are systematic in direction across Ss: for $0 \leq K \leq 20$, $T < -3.66$ ($p < .00025$, two-tailed), except for $K = 18$, where $T = -2.36$ ($p < .02$, two-tailed). The assumption that K is constant across Ss is not at all critical: $\chi^2_{\min}$ ($0 \leq K \leq 20$) = 64.64 ($p < .001$). Since the .05 critical level for chi square ($df = 16$) is 26.30, obtained $\hat{\sigma}^2$ would need to be more than doubled, on average, even to reduce the conservative $\chi^2_{\min}$ statistic to a nonsignificant level. [It will be recalled that

$$\chi^2 = \sum_{j=1}^I t_j^2 = N \sum (X_i - \mu)^2 / \hat{\sigma}^2 \}.$$

Thus the hypothesis that $\mu = 0.7$ for all Ss seems to be seriously in error.

EXPERIMENT II

Method

Twenty-four Ss were run in a signal recognition task, $\gamma = 0.8$, with zero payoff and no feedback. Two response sequences were taken from each S, under different discriminability conditions; these were used to test the hypothesis that r_2 rate is constant within Ss across conditions, or $\mu_1 - \mu_2 = 0$ (Eq. 4). The matching hypothesis, $\mu_1 = 0.8$, $\mu_2 = 0.8$, was also tested (Eq. 3).

Subjects. Ss were 24 staff or students at the Polytechnic, none of whom had taken part in Experiment I.

Task. The task was identical to that of Experiment I except that the intertrial interval was 1.75 sec instead of 4 sec, and each S was

Table 1
Response Probabilities and d' Estimates in Experiment I

$p(r_2)$	$p(r_2 s_2)$	d'
.61	.63	0.14
.66	.66	-0.02
.51	.52	0.10
.81	.79	-0.33
.69	.81	1.14
.73	.69	-0.43
.68	.72	0.40
.46	.48	0.24
.61	.67	0.63
.71	.73	0.21
.55	.62	0.75
.78	.80	0.27
.62	.66	0.34
.59	.64	0.54
.61	.64	0.30
.74	.81	0.38

run for 250 trials under each of two discriminability conditions, viz. high (H) and low (L) discriminability. For the L condition, signal levels were the same as in Experiment I; for the H condition, the s_2 level remained unaltered, but the s_1 level was lowered to 0.335 V rms (71 dB). Ss were run through a 100-trial block under each condition, then through a 150-trial block under each condition, in one of the orders HLHL or LHLH. A balanced factorial design was used to determine condition order and assign hands to response keys. Time series analysis is applied below to the last 85 trials of each condition.

Stimuli. Stimulus sequences were determined and presented as in Experiment I, except for the difference in γ . Overall s_2 frequency was 0.799.

Instructions. Instructions were identical to those in Experiment I, mutatis mutandis, with the following additions: S was told that there were two discriminability conditions, was told their order, and knew that only s_1 intensity varied between conditions; he was explicitly told that γ remained at 0.8 for both conditions. (Again, all Ss could tell E γ at the end of the experiment.)

Results

All results below are calculated on the last 85 trials of each condition.¹ Two Ss failed to respond on one or more of these trials and were replaced.

Table 2 shows $p(r_2)$, $p(r_2 | s_2)$, and d' for L and H conditions. Equation 1 predicts that the two $p(r_2)$ values for each S should be identical, subject to sampling error; given the additional assumption $k = 1$, both values should equal 0.8. Group mean $p(r_2)$ was, in fact, 0.722 for the L condition and 0.724 for the H condition. Twenty-one Ss showed $d' > 0$ in the L condition, 24 in the H condition; no S ever showed perfect discriminability. The difference $p(r_2 | s_2) - p(r_2 | s_1)$ increased for all Ss in passing from the L to the H condition; this shows that discriminability did indeed vary between conditions, insofar as Thomas and Legge's Eq. 1 makes $p(r_2 | s_2) - p(r_2 | s_1)$ a monotonic function of discriminability for fixed k and γ . Average sample autocovariances were similar in both conditions to those of Experiment I and are not given here.

Despite the similarity of group mean $p(r_2)$ across L and H conditions, the intercondition differences ($X_1 - X_2$) in mean r_2 frequency were much too large to be compatible with the model $\mu_1 = \mu_2$. Thus $\chi^2 > 91.6$ for K anywhere in the range $0 \leq K \leq 20$; allowing K to vary within and between Ss does nothing to save the model, since $\chi^2_{\min} (0 \leq K \leq 20) = 62.22$ ($df = 24$, $p < .001$). On the other hand, the differences in response rate, though large, are not systematic in direction across Ss: $0 < T < 2.3$ ($p > .02$, two-tailed) for $0 \leq K \leq 20$, except for $K = 19$, where $T = 4.19$.

As in Experiment I, the $p(r_2)$ tend to undershoot γ in both conditions. In fact, using Eq. 3 to calculate t for each S within each condition, $T < -4.47$ ($p < .00001$, two-tailed) for $0 \leq K \leq 20$ in both conditions separately, while $\chi^2_{\min} (0 \leq K \leq 20) = 284.02$, for the L condition, and $\chi^2_{\min} (0 \leq K \leq 20) = 113.59$, for the H condition; the .001 critical level

Table 2
Response Probabilities and d' Estimates in Experiment II

L			H		
$p(r_2)$	$p(r_2 s_2)$	d'	$p(r_2)$	$p(r_2 s_2)$	d'
.55	.58	0.26	.41	.44	0.88
.63	.73	1.09	.76	.90	**
.91	.90	-0.13	.54	.63	1.09
.76	.84	1.16	.66	.79	1.63
.95	.97	0.63	.74	.83	1.38
.85	.86	0.34	.80	.93	2.12
.69	.73	0.53	.74	.88	1.68
.81	.79	-0.72	.76	.78	0.55
.86	.88	0.41	.81	.93	1.94
.71	.76	0.50	.69	.82	**
.86	.88	0.47	.86	.97	2.23
.74	.80	0.67	.88	1.00	*
.76	.78	0.54	.72	.85	1.83
.92	.96	1.29	.82	.94	2.06
.76	.73	-0.40	.75	.88	1.55
.11	.12	0.34	.61	.72	1.51
.80	.83	0.39	.80	.94	1.94
.71	.76	0.54	.73	.81	1.20
.49	.51	0.15	.46	.58	**
.76	.84	1.16	.80	.89	1.74
.83	.86	0.58	.87	.91	1.49
.76	.84	0.98	.79	.96	2.89
.62	.64	0.22	.65	.66	0.35
.42	.45	0.37	.73	.85	1.71

*undefined, $p(r_2 | s_2) = 1.00$ **undefined, $p(r_2 | s_1) = 0.00$

for chi square, $df = 24$, is 51.18. Thus the deviations from the matching prediction are very gross.

DISCUSSION

The results of both experiments are strikingly at variance with Thomas and Legge's hypothesis. Ss did not match $p(r_2)$ to γ under symmetric payoff, nor did they keep $p(r_2)$ constant across discriminability conditions. This last result is important in ruling out two modifications to Thomas and Legge's hypothesis. It shows, first, that it is insufficient to assume that Eq. 1 holds, but with $k \neq 1$ for symmetric or zero payoff. It shows, second, that it is insufficient to assume that S matches *subjective* rather than *objective* probabilities, unless the function relating subjective to objective probability is allowed to vary with discriminability; but this last concession would remove all content from Thomas and Legge's proposal. Thus the proposal seems quite unpromising for the situations studied here, though it may, of course, fare better when Ss are more experienced or are given feedback. As was pointed out earlier, previous work sheds little light on this possibility, since adequate *formal* tests of the hypothesis have not been carried out. It should be noted, however, that Thomas and Legge suggest that their proposal is *most* likely to apply to naive Ss under minimal information conditions, and this at least seems false.

A natural question is whether Thomas and Legge

need the crucial assumption made here, that the response sequences are second-order stationary. Certainly other assumptions are possible: for example, the random variables $X_m, X_{m+1}, \dots$ might have a constant expectation but a covariance structure which changes over time. Unhappily, any such model would probably need to be made quite specific in order to give predictions about the sampling distribution of $(X - \mu)$ (see Anderson, 1971). Thus Thomas and Legge would need to give up their wish to avoid specific models of trial-by-trial events in order to make their model formally testable. At the very least, as the present results show, the model would need to be made sufficiently specific to exclude stationary stochastic processes.

REFERENCES

- ANDERSON, T. W. *The statistical analysis of time series*. New York: Wiley, 1971.
- CREELMAN, C. D., & DONALDSON, W. ROC curves for discrimination of linear extent. *Journal of Experimental Psychology*, 1968, 77, 514-516.
- DUSOIR, A. E. Treatments of bias in detection and recognition models: A review. *Perception & Psychophysics*, in press.
- FELLER, W. *An introduction to probability theory and its applications*. Vol. 2. New York: Wiley, 1971.
- GRANT, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 1962, 69, 54-61.
- BRAGIMOV, I. Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, 1962, 7, 349-382.
- JONCKHEERE, A. R., & BOWER, G. H. Non-parametric trend tests for learning data. *British Journal of Mathematical & Statistical Psychology*, 1967, 20, 163-186.
- KENDALL, M. G. *Rank correlation methods*. London: Griffin, 1970.
- KENDALL, M. G., & STUART, A. *The advanced theory of statistics*. Vol. 2. London: Griffin, 1973.
- KNUTH, D. E. *The art of computer programming*. Vol. 2. Reading, Mass: Addison-Wesley, 1969.
- LEE, W. *Decision theory and human behaviour*. New York: Wiley, 1971.
- MOOD, A. M., & GRAYBILL, F. A. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1963.
- NORMAN, M. F. Statistical inference with dependent observations: Extensions of classical procedures. *Journal of Mathematical Psychology*, 1971, 8, 444-451.
- NORMAN, M. F. *Markov processes and learning models*. New York: Academic Press, 1972.
- PARKS, T. E. Signal detectability theory of recognition memory performance. *Psychological Review*, 1966, 73, 44-58.
- THOMAS, E. A. C., & LEGGE, D. Probability matching as a basis for detection and recognition decisions. *Psychological Review*, 1970, 77, 65-72.

NOTE

1. This note concerns the stationarity of these last 85 trials. Given N consecutive trials, N large, it would be possible to take two blocks of trials, say $1-j, k-N$, to which Eq. 4 could be applied ($k-j$ would need to be sufficiently large to ensure no carry-over). This would provide a strong parameter-free test of stationarity, but is impossible here, since trials are few and the small-sample properties of the generalized t statistics are unknown.

As a considerably weaker alternative, number the last 85 trials from 1 to 85, and define

$$Y_1 = \sum_{j=1}^{15} X_j, Y_2 = \sum_{j=36}^{50} X_j, Y_3 = \sum_{j=71}^{85} X_j.$$

Given the assumption $c(k) = 0, k > 20$, the Y_i are independent. Given strict stationarity, they are identically distributed. Thus the usual trend-test theory applies (e.g., Kendall & Stuart, 1973, Chapter 31).

Tests for an upward trend in the Y_i were carried out on the three crucial 85-trial blocks from Experiments I and II, using Kendall's S statistic (Kendall, 1970, 5.1-5.5). All results were nonsignificant ($p > .1$). There is, then, no evidence here of a continuing approach to the matching hypothesis prediction.

It should be noted that the much stronger trend test of Jonckheere and Bower (1967), which might seem to apply directly to the random variables $X_1, X_2, \dots$, is irrelevant here, since its null hypothesis attributes equal probability to all distinguishable permutations of the obtained sequence. This attribution does not follow from either strict or second-order stationarity. An extreme counterexample is the strictly stationary sequence $X_1, X_2, \dots$, defined by

$$\begin{aligned} p(X_1 = 0) &= p(X_1 = 1) = 0.5 \\ X_j &= 1 \text{ if } X_{j-1} = 0 & j > 1 \\ X_j &= 0 \text{ if } X_{j-1} = 1 & j > 1. \end{aligned}$$

(Received for publication November 26, 1973;
revision received June 18, 1974.)