COGNITIVE PSYCHOLOGY 28, 17–64 (1995)

On the Nature of Streaks in Signal Detection

DAVID L. GILDEN

University of Texas at Austin

AND

STEPHANIE GRAY WILSON

Seton Hall University

Human performance in the domain of signal detection is analyzed with respect to the formation of streaks. Streakiness was found to be a general property of auditory and visual discrimination in the sense that correct and incorrect responses have a positive sequential dependency. Success tends to follow success and failure tends to follow failure. Level of streakiness was discovered to be a function of the attentional demand required by the discrimination. Discriminations that make the least demand on attentional resources produce the highest level of streakiness. Monte-Carlo simulations of the observed data sequences suggest that streaky performance is a residue of wave-like variations in perceptual and attentional resources. © 1995 Academic Press, Inc.

There are potentially two levels of information that are present in a response to a stimulus. At a nominal level, there is the identity of the response; what the response was. In some domains of judgment the response may be evaluated as an assertion about a matter of fact, and if so, it may or may not be correct. Different traditions within psychology have focused on one or the other level of information. In magnitude estimation and category judgment, for example, the response *per se* is the only datum as the immanent experience of a stimulus can be neither right nor wrong. In signal detection methodologies, however, there is a concept of error, and both levels of information are required to develop the theory of receiver sensitivity. In correspondence with the two levels of information that are present in a response, there are two kinds of biases that inform the interpretation of data: response bias in the sense of identity and re-

This work was supported by a grant from Vanderbilt University to David Gilden and by NIH Grant P30-EY08126. We are especially grateful to Wes Grantham for his generous assistance in developing the audition study and in allowing us to use his laboratory. We thank David Bloom for programming assistance. We also extend special thanks to Kerstin MacDonald for assistance in running the experiments and preparing the data for analysis. Reprint requests should be addressed to David Gilden, Department of Psychology, Mezes 330, University of Texas at Austin, Austin, TX 78712.

0010-0285/95 $9.00

sponse bias in the sense of correctness. In signal detection both types of response bias are potentially present.

Response bias generally refers to the identity of a response and typically appears in two forms, as unequal representation and as nonstationarity. For example, unequal representation in a yes/no methodology may appear as a tendency for the subject to say "yes" rather than "no." This form of response bias is well understood and is handled by the theory of receiver operator characteristics (ROC) which disentangles response representation from response correctness. Nonstationarity in response identity, however, is not handled by any general theory and can create problems of interpretation in situations where there is no external criterion of correctness. It usually manifests itself as a positive sequential dependency known as response assimilation. In this regard, several studies have demonstrated that a categorized "yes" or "no" response to the presence of dim test light depended on the history of previous responses (Verplanck, Collier, & Cotton, 1952; Verplanck, Cotton, & Collier, 1953; Verplanck & Blough, 1958) such that if a subject has just responded "yes," they are biased to respond "yes" again. Consequently runs of "yeses" and "noes" were longer than expected under the null hypothesis of response independence. Similar assimilation effects have also been found in magnitude estimation (Luce, Nosofsky, Green, & Smith, 1982; Staddon, King, & Lockhead, 1980); i.e., if a subject has just said "very loud," they are biased to say something similar to "very loud" on the subsequent trial. Early work on the nonindependence of successive threshold measurements (Wertheimer, 1953) was later shown to be accounted for by runs of like responses (Howarth & Bulmer, 1956), consistent with the general result that responses are assimilatively biased. It should be noted that negative sequential dependency in response has also been reported. Fernberger (1920) found a contrast effect in perceived weight. Near threshold, subjects are biased to judge a comparison weight as heavier than a standard weight if the weight on the previous trial was judged to be lighter and vice versa.

In the context of signal detection, we can also inquire into the existence of response bias in the sense of outcome, whether there is nonstationarity in correct and incorrect responses across trials. This type of bias has not been well studied and is generally presumed not to occur. Trial independence in correctness of response is in fact assumed by ROC theory (Falmagne, 1985) which is designed to give a bias-free estimate of discriminability. This type of nonstationarity, were it to occur, would indicate that receiver sensitivity is not constant in time and would necessitate a revision in the way threshold measurement is conceived. In other contexts, especially in sports, there is a related type of nonstationarity that is referred to as streakiness. Streakiness is characterized by extreme nonstationarity in hit rate and is widely perceived to be a real aspect of skilled performance. In this article we investigate the nature of streakiness in signal detection.

Atkinson (1963), in an isolated effort, developed a semi-analytic theory of signal detection based upon the notion that momentary sensitivity to threshold stimuli is dynamic and conditionalized upon both the stimuli received on earlier trials and the responses made to them. While the theory is difficult to evaluate, primarily because the formality of its axiomatic structure is compromised by a surfeit of free parameters, Atkinson does report the results from a tone detection study in which streaks in outcome were distinguished from streaks in response identity. As will be discussed in greater detail below, Atkinson found that there was positive sequential dependency in both response identity and response correctness, although the dependency in correctness was rather small in amplitude. In this article we continue Atkinson's work by generalizing the discrimination tasks, by clarifying the circumstances that promote streakiness, and by developing a different theoretical focus that is based on attention.

Our experiments differ from typical signal detection experiments in that we are not interested in threshold measurement *per se* but rather in the fluctuations that occur at a single point on the psychometric curve relating hit rate to stimulus contrast. The data from these experiments are sequences of binary numbers that reflect the production of correct and incorrect responses (hits and misses) as a discrimination task is iterated at fixed stimulus parameters. These sequences provide an opportunity to examine receiver nonstationarity in a context that is free from the additional complexity associated with stimulus uncertainty and variation in task difficulty. The primary statistical problem is to discern whether outcome sequences are consistent with the output of a Bernoulli process. That is, determining whether the instantaneous hit rate is stationary such that the outcome of a given trial is independent of outcomes on previous trials.

Gilovich, Vallone, and Tversky (1985) describe a number of statistical tests that can be used in assessing departures from a Bernoulli process. We shall employ several of these in our analyses in order to test the null hypothesis that outcome sequences derive from a Bernoulli process. However, it must be recognized that tests of this null do not provide explanations of why sequences might exhibit trial dependency, and there are several ways in which successive outcomes could become correlated. A first-order Markov process, for example, will induce local trial dependence by virtue of making the probability of a hit on a given trial contingent upon performance on the previous trial. In this case the marginal probability would be stationary, but individual trial probabilities would fluctuate. The hit rate might also vary secularly as would be the case if fatigue or learning effects influenced performance. Secular trends in hit rate might also be a signature of the existence of underlying mechanisms controlling attention and perception that fluctuate over time. In these latter cases, the outcome of a given trial might depend upon its position within the trial sequence rather than explicitly on the outcomes of earlier trials. Discovering the source of nonstationarity in observed sequences is a much subtler endeavor than demonstrating the existence of nonstationarity. In this article we shall attempt to develop explanatory constructs for the etiology of streaky performance, and we present a theoretical analysis based upon Monte-Carlo simulation that provides some insight into the mechanisms of streak production.

PRELIMINARY STUDIES

The basic designs in signal detection (two-alternative forced choice, two-interval forced choice, yes/no) all explicitly incorporate the notion of correctness of response. It is this feature that allows any signal detection task to be a candidate for assessing streakiness in outcome. Outside of this observation, there was initially no theoretical motivation or obvious reason to study any particular discrimination task with respect to the formation of streaks. The tasks that form the preliminary set of studies were chosen not on the basis of a prior theory, but only because they required different kinds of judgments. We repeat the Atkinson (1963) experiment for detection of a tone embedded within noise. Also included are judgments of shape defined by motion (structure-from-motion also known as kinetic depth), judgments of relative line length, and detection of a brief flash.

Subjects

Four subjects participated in each of the vision studies. Nine subjects participated in the audition study. Subjects were recruited through advertisement and were paid $5 per session.

Stimuli

All visual stimuli were displayed on a 13" Apple Macintosh color monitor. Viewing conditions in all experiments were mesopic. In all of these experiments the correct choices were randomized and counterbalanced.

Flash detection. Two uniformly illuminated squares subtending 4 degrees were placed horizontally 1 degree apart. The luminance of the squares was a gray near the midpoint of the gray scale. The squares were shown on a uniform black field. On each trial, one square would brighten for 16 ms. The subject's task was to indicate which of the two squares brightened. The level of brightening was adjusted individually for each subject to achieve hit rates near 60, 75, and 90%. Stimuli were presented in blocks of 500 trials without feedback.

Ovateness detection. Shapes were depicted as random dot kinetic depth forms. The forms were composed by placing 1000 dots over the bounding surface. The forms subtended 4 degrees and oscillated back and forth through 40 degrees in smooth motion for 2 s. On each trial either a sphere or a slightly prolate ellipsoid was shown. The subject's task was to indicate whether a sphere or prolate ellipsoid was displayed. The deviation from sphericity was adjusted individually for each subject to achieve hit rates near 60, 75, and 90%. Stimuli were presented in blocks of 200 trials with feedback.

Distance ratio detection. The stimuli consisted of two sets of three vertical tick marks. The set on the left was defined as a reference and the distance between the outer tick marks was always the same. The middle tick mark varied in its horizontal position and divided the space into two regions of size r_1 and r_2. The set of tick marks on the right was defined as a test set. In this set, there were seven different distances between the outer tick marks, none of which equaled the distance in the reference set. The middle tick mark in the test set divided the space into regions of size t_1 and t_2. The subjects task was to indicate whether $r_1/r_2 > t_1/t_2$ or $r_1/r_2 < t_1/t_2$. The ratio r_1/r_2 varied randomly on each trial. The hit rate was controlled by the placement of the middle tick mark in the test set. The absolute values of t_1/t_2 were adjusted for each subject individually to achieve hit rates near 60, 75, and 90%. Stimuli were presented in blocks of 210 trials without feedback.

Tone detection. Sinusoidal signals were generated digitally and were produced at a 20 kHz sampling rate. The masker was a continuous 8 kHz low pass filtered noise that was present throughout all of the trials. The noise level was 75 dB SPL. Signals had a rise and decay time of 10 ms and a total duration of 300 ms. The frequency of the amplified tone that was to be detected was 1000 Hz. The tone detection task was designed as a two-interval forced detection task. In one interval, the masker alone was presented, while in the other the masker was accompanied by the target 1000 Hz tone. The two intervals were separated by 500 ms. Lights were used to define the observation intervals and to provide correct-answer feedback. A 300 ms warning light and a 300 ms delay preceded each trial. The subject's task was to indicate which of the two intervals contained the tone. The amplitude of the tone was adjusted for each subject individually to achieve hit rates near 60, 79, and 87%. Stimuli were presented in blocks of 300 trials.

Procedure

In the vision studies (flash, distance ratio, and ovateness detection), each subject participated in three blocks of trials at each level of difficulty for a total of nine blocks. In the audition study, each subject participated in one block of trials at each difficulty level. At the beginning of each trial block, the subject was calibrated in order to determine what stimulus parameters were required to achieve the desired hit rate. In the vision studies calibration was accomplished by the method of constant stimuli. In the audition study calibration was determined by an interleaved staircase. The 60, 79, and 87% percentage correct points were estimated using the appropriate up–down rules. In the flash detection experiment, subjects were dark adapted for 10 min before calibration took place. Trials were self-paced in all experiments except for ovateness detection. The presentation program for displaying kinetic depth required 7 s to compute the individual animation sequences.

Analysis

The data from this experiment consist of sequences of zeros (misses) and ones (hits). There are a number of statistical measures that may be defined on such binary sequences that measure deviation from the output of a Bernoulli process. Gilovich et al. (1985) used conditional probabilities, run counts, and serial correlations to characterize the basketball sequences in their studies. These measures are all related to some extent. For a given hit rate, sequences with fewer runs than expected under the null hypothesis of a Bernoulli process must have more internal repetition than expected and consequently a positive serial correlation. In addition, for such sequences $p(h|h) > p(h)$—the probability of a hit following a hit is greater than the probability of a hit.

Unlike the serial correlation between successive trials, denoted here as r_{12}, and the contingent probability difference, $\Delta p = p(h|h) - p(h)$, the number of expected runs is influenced by the hit rate. In order to develop a useful statistic from run counts it is necessary to refer each sequence to the ensemble formed from all of its permutations. The sampling distribution of run counts is approximately normal with mean $2Np(1 - p) + 1$, where N is the number of trials and p is the probability of a hit. Deviations from normality are sufficiently large for $p \neq .5$ that we use the exact hypergeometric distribution (Hays, 1988) to compute the probability of observing R or fewer runs for given numbers of hits and misses. R here refers to the number of runs that were actually observed in the sequence under consideration. It is convenient to convert this probability to a z score by inverting the cumulative Gaussian distribution—a quantity that will be referred to throughout as the *runs z score*. The runs z score is a measure of outcome clustering that is not biased by either sequence length or hit rate.

Although r_{12}, Δp, and the runs z score are related measures of sequence structure, they are not identical. It is the case that the Fisher Z associated with r_{12} and the runs z score are virtually identical, they cannot be distinguished to three significant digits. Δp, however, is not a function of the runs z score. Simple regressions of these two variables captures only about 50% of the variance for the sequences discussed in this article. The reason for this is that the first order transition probability $p(h|h)$ is not as sensitive to global structure as the serial correlation or run count. It is necessary to look at the first, second, and third order transition probabilities to begin to adequately characterize a sequence. Often it is the case that sequences that are easily discriminated in terms of runs z score are distinguished by subtle differences in the relative amplitudes of the higher order transition probabilities. Transition probabilities are more useful as measures of local sequence structure, and for this reason we shall use the runs z score to characterize the departure from a Bernoulli process.

The statistical tests that will be presented in this article are of two kinds, depending on the type of null hypothesis that is being considered. In every experiment reported here we shall ask the question whether the sequences of hits and misses can be distinguished from a Bernoulli process. The null hypothesis that the sequences cannot be so distinguished entails both that (1) each sequence is independent (as are parts of sequences) from all others and (2) the ensemble of runs z scores form a normal distribution with unit variance and zero mean. In testing this null we simply form the distribution of runs z scores and ascertain the significance of the deviation of the mean from zero. Each sequence in such an analysis forms a separate degree of freedom by virtue of their mutual independence required by the null. In other cases we shall ask whether the results from one or more experiments can be distinguished, i.e. whether different tasks have different levels of streakiness. Here the null hypothesis has nothing to do with whether the sequences are Bernoulli or not, and we shall resort to more traditional repeated measure analyses where the different subjects constitute the degrees of freedom.

Results

The distributions of runs z scores from the four preliminary signal detection experiments are shown in Fig. 1. The results from the flash detection task were particularly striking. The distribution of runs z scores across subjects had a mean of -1.18 and a median of -1.33. Twenty-eight of 36 sequences had fewer runs than expected ($p < .0006$). Individual sequences were often significantly anomalous and could be distin-

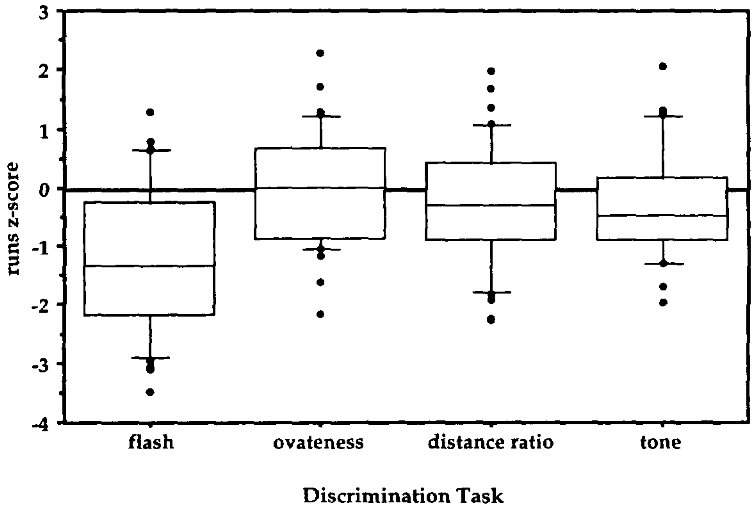


FIG. 1. Box plots of runs z score for the initial set of signal detection tasks.

guished from a Bernoulli process at constant hit rate. Forty-two percent of the sequences had $z < -1.65$ while 28% had $z < -2$. The expected percentages from a Bernoulli process are 5 and 2.3%, respectively. The magnitude of the streakiness was so large in this study that we performed a replication with two additional subjects who participated in nine blocks of 500 trials each. The mean runs z score for these additional subjects was $-.96$, 33% of the sequences had $z < -1.65$, and 22% had $z < -2$. It is evident that performance in this task is streaky.

In contrast, there was no evidence for streaky performance in ovateness discrimination. Both the ensemble of sequences as well as individual sequences conformed to the expectation from a Bernoulli process. The distribution of runs z scores had a mean of $-.021$ and a median of $-.016$. Eighteen of 36 sequences had fewer runs than expected ($p < .57$). The distribution of runs z scores could not be distinguished from a normal distribution with mean of zero.

The sequences from the two remaining tasks were intermediate in their level of streakiness. The distance ratio discrimination task generated a distribution of sequences that was overall biased toward negative z scores and produced a relatively large number of sequences that were individually anomalous. The distribution of runs z scores had a mean of $-.24$ (this is a marginal result as $t(35) = -1.38$, $p < .089$) and a median of $-.29$. Fifteen percent of the sequences had $z < -1.65$, while 6% (two sequences) had $z < -2$. The evidence for streaky performance in the context of tone detection was somewhat weaker. The distribution of runs z

scores had a mean of $-.32$ ($t(26) = -1.72$, $p < .048$) and a median of $-.46$. Nineteen of 27 sequences had fewer runs than expected ($p < .026$). However, there was not a preponderance of anomalous sequences as was observed in the other tasks. Only 7% (two) of the sequences had $z < -1.65$ and there were none with $z < -2$. This is roughly the tail distribution you would expect from 27 Bernoulli sequences. Thus, although the ensemble of sequences could be distinguished from the output of a Bernoulli process at constant hit rate, the individual sequences could not. The runs z score distributions of sequences from the ratio and tone studies could not be distinguished from each other.

In the design of the initial signal detection tasks, we calibrated subjects to perform at three distinguishable levels of hit rate. This was motivated by an earlier finding in the domain of motor skills that streakiness is related to level of difficulty (Gilden, Gray, & MacDonald, 1990). Golf putting and dart throwing tend to generate Bernoulli-like outcome sequences at low and high hit rates and streaky sequences at an intermediate hit rate. There was no evidence for quadratic trends in the runs z score in signal detection sequences, although those sequences with the most negative z scores occurred in the hit rate interval (.6, .85).

Discussion

It is instructive to analyze Atkinson's (1963) results for two-interval forced choice (2IFC) tone detection in the dual contexts of response assimilation and positive sequential dependency in response correctness. Atkinson provides data on the probability of a response contingent upon both the current stimulus as well as upon the response and stimulus on the previous trial. Atkinson uses the following notation for conditional probabilities which we shall also adopt; letting A refer to responses and S to stimuli presented, $Pr(A_i|S_jA_kS_l)$ is the probability that the subject responded that the tone was in the ith interval given that the tone was in the jth interval and on the preceding trial the subject responded that the tone was in the kth interval when it was in fact in the lth interval. The subscripts i, j, k, and l range over the set (1, 2). In Atkinson's theory and in the data he presented, there was no response bias for preferring one interval over another. Therefore the data are invariant under the global substitution of 1 for 2 and 2 for 1. The average hit rate in this experiment was $Pr(A_1,S_1) = .73$. This value is not of interest in itself and indicates only the relative amplitude of the tone.

In reviewing Atkinson's data it may appear at first that there was substantial positive sequential dependency because $Pr(A_1|S_1A_1S_1) = .80$. In this case $Pr(A_1|S_1A_1S_1)$ is the probability of a hit given that the signal was in the same interval on the previous trial and that trial was also a hit. This is a large increase and if it generalized to trial pairs where the tone was in a different interval on the previous trial would correspond to an average z score much less than -1. However, $Pr(A_1|S_1A_2S_2) = .67$, indicating that hits tend to follow hits only if the same response can be given. The two conditionals are nearly symmetrically placed around the basal hit rate of .73. On average, the probability of a hit following a hit was .735, evidence that there was little (although significant) sequential dependency in correctness of response. From Atkinson's results we would conclude that to the extent that streaks exist, they primarily result from the bias of response assimilation, a result that has been continuously reiterated since it was noticed that people tend to repeat themselves.

Atkinson's (1963) results appear to be weaker than ours. In our tone detection experiment the probability of a hit following a hit was 2% larger than the probability of a hit, while in Atkinson's experiment the increment was only ½%. It is difficult to say how Anderson's results would compare if streakiness were measured by the runs z score. As we have discussed above, when the runs z score was introduced, global sequence structure is not well characterized by the first order conditional probabilities. For example, it is necessary to look at probabilities that are conditionalized upon the previous two and three trials to distinguish flash sequences which were highly streaky from tone sequences which generally were not. Atkinson (1963) does not report the higher moments.

THE ROLE OF ATTENTION IN STREAK PRODUCTION

In attempting to understand the results from the psychophysical studies we must confront the fact that the tasks were very different from one another and varied across a myriad of dimensions. The data indicate that the tasks divide into three separate groups; extremely streaky (flash detection/average runs z score ~ -1), moderately streaky (distance ratio and tone detection/average runs z score $\sim -.3$), and not streaky (ovateness detection/average runs z score ~ 0). In order to account for the existence of these groups we have focused on two distinctions; the attentional allocation required to process the stimuli within a task, and the time delay between presentation. We will consider the effect of attention first.

Our initial studies suggest that the level of attentional demand required for stimulus identification is an important variable in the production of streaks. The streakiest outcome sequences were associated with a stimulus that under some circumstances may be discriminated preattentively; a brief flash. The perceptual processing of a superthreshold flash is preattentive in the sense that detection of a flash does not depend on the number of distractors; things that flash pop out. It is not necessary to conduct detailed experiments to validate this notion. A blinking light in the night sky will pop out even against the backdrop of rich star fields. On the other hand, neither distance ratio discriminations nor kinetic depth shape judgments led to a high level of streakiness. It is also the case that both of these tasks require focused attention—even for superthreshold differences and errorless performance. In the case of distance ratios, a target that has the larger interval on the right will not pop out from a field of distractors that have the larger interval on the left. Left–right reversals do not pop out primarily because the power spectrum is not changed by such a transformation (Julesz, 1975, 1981). The shapes of random-dot kinetic depth stimuli do not pop out for different reasons. Such stimuli create an impression of depth that is labile (it reverses in parallel projection) and builds up over time as the animation unfolds. It is necessary to focus attention on random-dot animations in order to perceive their structure.

Having made this distinction among our stimuli for superthreshold differences, it must be recognized that all of our experiments were carried out at threshold. Had we used superthreshold differences, discrimination performance would have been virtually errorless. It is not possible to study outcome sequence structure when there are no errors. In our studies it was necessary to induce error and this can only be done when the discriminations are conducted near threshold. We are thus led to consider the nature of visual search for barely discriminable differences. In particular, we are interested in differences that would be preattentively identified at superthreshold. While the existence of a distinction at superthreshold is sufficient for the purpose of understanding our empirical results, the theoretical analysis given below will attempt to interpret sequence structure in terms of attentional demand. Consequently, the issue of processing at threshold is not one that can be finessed.

The nature of attentional limitations at threshold has not been systematically studied. As Shiffrin (1988) remarks, this is an intricate question that touches on a number of difficult issues in signal detection, ideal observer theory, and decision modeling. Bergen and Julesz (1983) have argued that search processes which are parallel at superthreshold become increasingly serial as threshold is approached. The basic idea here is that attention is intrinsically limited in capacity and that it may be allocated broadly in space if it is not needed to resolve small stimulus differences. At threshold, the spatial field covered by attention is conceived to "zoom" in to provide greater local resolving power. This conception of attention as having variable coverage and resolution is certainly consistent with naive experience and is supported by a number of independent studies (Jonides, 1980, 1983; Eriksen & Yeh, 1985). However, empirically distinguishing a parallel process from one that is serial is problematic because there are often limited-capacity parallel models of serial searches and vice versa (Townsend & Ashby, 1983; Townsend, 1990). In fact, the Bergen and Julesz (1983) studies do not adequately address the distinction between serial and limited-capacity parallel processing.

Assessment of the nature of visual attention at threshold difference levels requires a methodology that incorporates an analysis of error. Palmer, Ames, and Lindsey (1993) have recently developed a method based on error rate that provides evidence that stimulus differences that pop out at superthreshold are also processed in parallel at threshold. In their methodology, difference magnitudes at fixed error rate are measured as a function of set size. Set size effects are conceived to arise from decision phenomena (opportunity for false alarm increases with number of distractors) in addition to sensory limitations. Palmer et al. consistently found that for simple visual search, search that would be manifestly parallel at superthreshold, the increases in difference magnitudes with set size that are required to maintain a constant rate of error could be completely accounted for in terms of decision processes. There was no evidence for attentional limitations in the extraction of information about the various stimuli. The model that best fit the data was one in which noisy percepts were presented to a decision maker such that the variance of the noise distributions was invariant with set size. This latter invariance is the hallmark of parallel processing that has unlimited attentional capacity. In what follows we shall make the conservative conjecture that stimulus differences that pop out at superthreshold values also make minimal claims on attentional resources at threshold when stimulus sets do not exceed two in number; i.e., in 2AFC designs.

Processing style at superthreshold permits a distinction that allows the flash discrimination task to be separated from the other visual tasks in the initial group of studies. The flash discrimination task was unique in two senses; it generated the sequences with the largest sequential dependencies, and it was the only task in the initial group that was preattentive at superthreshold. These observations lead to the following conjecture:

> *A*. Superthreshold stimulus differences that are preattentively identified will create a higher level of streakiness in threshold discriminations than differences that require focused attention.

This condition on the magnitude of streak production may be further refined. All parallel unlimited-capacity processes are naturally on the same logical footing with regard to their usage of attentional resources—they essentially do not use any. Thus, there is no reason to distinguish between any stimuli that pop out at superthreshold in the production of sequences of hits and misses in threshold discrimination. The level of streakiness that was observed for flash discrimination may have been maximal and should be representative of all stimulus differences that support pop out. This is a second conjecture:

> *B*. All stimulus differences that are preattentively identified at superthreshold will produce a maximal level of streakiness (mean runs z score ~ -1.0) in threshold discrimination.

These two conjectures make nontrivial predictions about streak formation as a function of attentional resource usage. They have been framed so as to be falsifiable and were tested in the following five studies.

Generalization beyond the flash detection experiment required additional tasks incorporating stimulus differences that pop-out at superthreshold. Two suitable candidates are orientation and brightness discrimination. A tilted line is known to pop out in a field of vertical lines, as will one bright object among a number of faint ones. These feature differences support flat reaction time functions in singleton search (Treisman & Gelade, 1980; Treisman, 1982), and they can be incorporated into segmenting textures (Julesz, 1975, 1981; Bergen & Julesz, 1983). For a third task we chose an extremely elementary discrimination; whether a contour is present on the left or right of two parallel lines. The stimulus looks like a square with a missing side and we shall refer to it in this way. The evidence that all these differences are processed preattentively—in parallel and with unlimited capacity—is well established. If the second conjecture is true, then brightness, orientation, and missing side discrimination at threshold will yield runs z scores near -1, the value found for flash discrimination.

In order to provide further evidence for the first conjecture, that superthreshold stimulus differences that require focused attention are only moderately streaky at threshold, we required a task that mandated refined positional judgment—such as discriminating ratios of length. In another context, Gilden, Schmuckler, and Clayton (1993) had been assessing people's abilities to discriminate between fractal contours. The contours employed in these studies are known as fractional Brownian noises and are examples of random fractals. Fractional Brownian noises are defined by having power-law power spectra; power ~ (spatial frequency)$^{-\beta}$. The exponent of the power law, β, determines the fractional dimension of the curve and its roughness; the larger the value of β, the smoother the fractal contour. Although these fractals are not familiar stimuli in the psychological literature, they served our purpose here quite well.[1] Such discriminations manifestly demand focused attention even when contour differences are sufficiently large that performance is errorless.

The distinctions that we are drawing here between different types of

[1] Several experiments in which 2AFC comparisons of fractional Brownian noises were made (Gilden *et al.*, 1993) have shown that people possess an intuitive and immediate understanding of the degree of roughness in a noisy contour. These noises are of interest in their own right as they provide a useful geometric description of natural forms (Mandelbrot, 1983). Contours such as tree lines are characterized by $\beta \sim 2$ (Keller, Crownover, & Chen, 1987). Mathematicians have also had reasonable success in rendering natural scenes using fractional Brownian noises (Voss, 1985, 1988).

stimuli are not subtle. For illustrative purposes we show in Fig. 2 how examples of superthreshold differences vary in their ability to create perceptually segmenting regions. Segmentation and boundary formation are naturally allied with preattention; a boundary forms as a result of a global percept of stimulus difference. In all four panels the middle column is distinguished from the remaining columns by a large difference in the individual elements. In Panels A and B the middle column is clearly segmented showing the perceptual signature of parallelism and unlimited capacity for differences in brightness and orientation. In Panels C and D we illustrate that length ratios and contour roughness (represented here as fluctuations in brightness) do not lead to segmentation. Note that segmentation does not occur in Panels C and D even though elements in the middle column would not be confused with elements in the other columns (Gilden & Schmuckler (1989) have shown this for the $\beta = 1$ and $\beta = 2$ fractals represented exactly as in the figure).

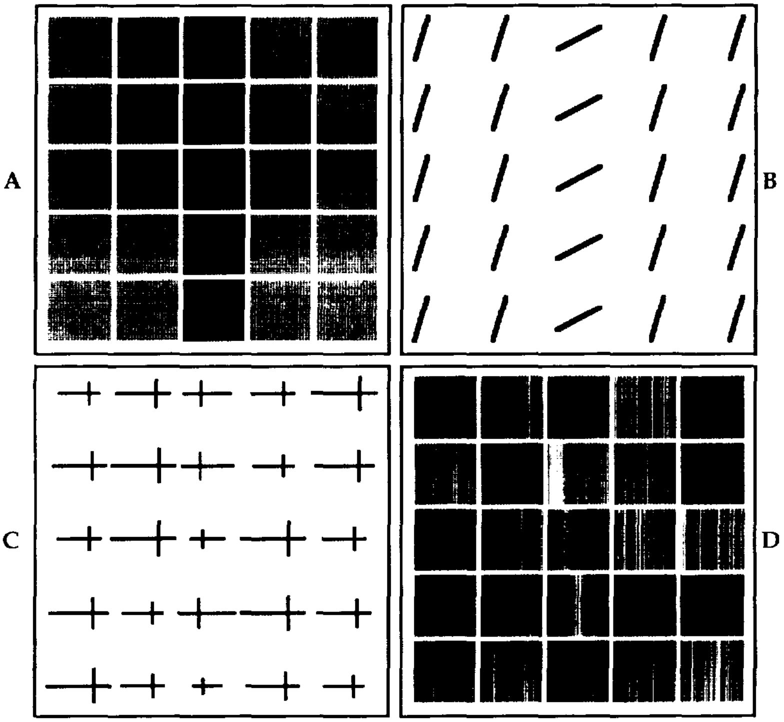


FIG. 2. Illustration of the formation of boundaries and groups on the basis of a preattentive difference. Orientation and brightness differences are processed preattentively and groups segment. Distance ratios and fractal power law differences do not generate boundaries even when groups would not be confused.

Streak Formation and Preattention

Subjects

Four subjects participated in the brightness and fractal discrimination experiments. Five subjects participated in the orientation discrimination experiment. These subjects were recruited through advertisement and were paid $5 per session. Fourteen subjects drawn from a course in experimental design participated in the missing side experiment. These subjects received course credit in lieu of payment. One subject failed to follow directions, evidenced by a hit rate that did not exceed chance guessing, and so was dropped from the study.

Stimuli

All visual stimuli were displayed on a 13″ Apple Macintosh color monitor. Viewing conditions in all experiments were mesopic. In these two alternative forced choice (2AFC) experiments, the correct choices were randomized and counterbalanced.

Brightness discrimination. Two uniformly illuminated squares subtending 4 degrees were placed horizontally 1 degree apart. The luminance of the squares was a gray near the midpoint of the gray scale, with one being slightly brighter. The squares were shown on a uniform black field. The subject's task was to indicate which square was brighter. Each subject was calibrated to achieve hit rates in an intermediate range of hit rate, on the order of .75. The stimulus was displayed for 500 ms. Stimuli were presented in blocks of 500 trials.

Orientation discrimination. Two lines subtending 2 degrees were placed horizontally 4 degrees apart. One line was tilted 4 degrees in a clockwise direction, the other was vertical. The contrast of the lines against the background was calibrated for individual subjects to a value that ensured hit rates on the order of .75. The subject's task was to indicate which of the two lines was tilted from vertical. The stimulus was displayed for 16 ms. Stimuli were presented in blocks of 500 trials.

Missing side discrimination. A single square was presented that subtended about 3° on a side. Either the left or right side was missing. The contrast between the square and the background was set so that a 68 ms stimulus duration would yield a hit rate of .75. The subject's task was to indicate whether the left or right side was missing. Stimuli were presented in blocks of 300 trials.

Fractal discrimination. Two versions of this experiment were conducted. In the first version, two stimuli appeared simultaneously, side by side in a 2AFC design. Here the subject's task was to indicate which of the two noises was rougher. In the second version, stimuli were presented sequentially. Here the subject's task was to categorize each stimulus as being in the rougher or smoother class. In either case, half of the contours had a power law exponent $\beta = 2$. We chose this value on the basis of earlier experiments (Gilden et al., 1993) which demonstrated that people's discrimination sensitivity to fractal structure is maximal near this point. The exponent of the other noise was calibrated separately for each subject to achieve a hit rate near .75. Fractals were represented as line drawings as in Gilden et al. (1993). This was a choice based primarily on convenience as Gilden and Schmuckler (1989) have shown that fractals are roughly equally discriminable independent of the medium of presentation; i.e. whether they are represented as fluctuations in height or as fluctuations in brightness. Stimuli were presented in blocks of 500 trials.

Procedure

Each subject in the brightness, orientation, and fractal discrimination experiments participated in 10 blocks of trials. Subjects in the missing side experiment completed 5 blocks of trials. At the beginning of each trial block, subjects were calibrated in order to determine what stimulus parameters were required to achieve the desired hit rate. In the brightness detection experiment, subjects were dark adapted for 10 minutes before calibration took place. Trials were self-paced in all experiments.

Results

The results from these three experiments are shown in Fig. 3. We have also replotted the results from the flash and distance ratio discrimination studies in order to make it clear that the attention variable cleanly separates our studies into two discrete groups and that runs *z* scores from different studies are distributed alike within the same group. A repeated measures analysis of variance (with attentional demand as a between variable) verified that preattentive tasks were streakier than tasks requiring focused attention ($F(1,38) = 16.0$, $p < .001$). However, a main effect for the attention manipulation is not sufficient for our purposes; the data must indicate that outcome sequences from the flash, brightness, orientation, and missing side experiments have equivalent levels of streakiness. The satisfaction of this additional requirement is evident in the figure; all four runs *z* score distributions overlap. A repeated measures analysis (with task as a between variable) showed that none of the studies in the preattentive group could be distinguished in terms of their level of

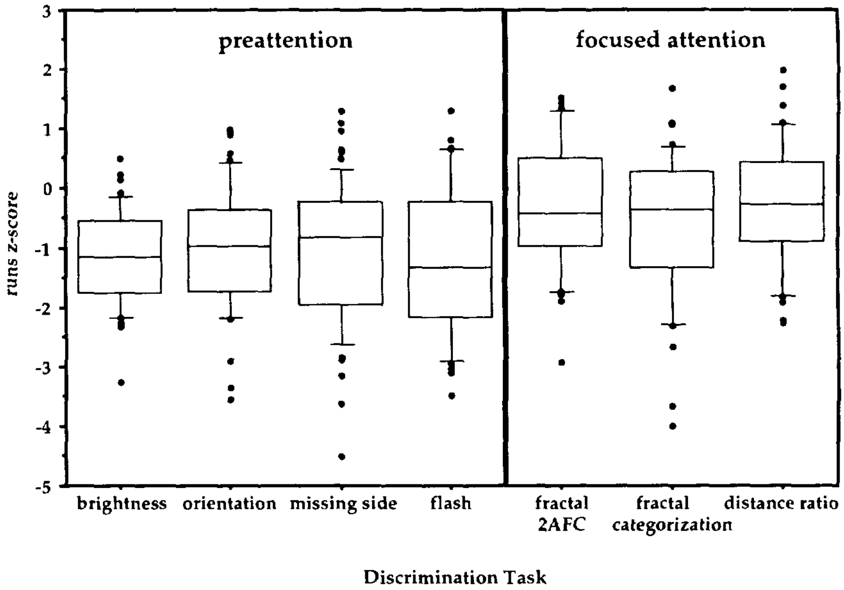


FIG. 3. Box plots of runs *z* score for tasks requiring preattention and focused attention respectively. The attentional difference cleanly separates the *z* score distributions into two groups. Within a group, the distributions are indistinguishable. Data from the flash and distance ratio study are replotted here for comparison.

streakiness ($F(3,24) = .05, p < .98$). The means and negative tails of the runs z score distributions are also quite similar within the preattentive group. In the brightness experiment, the average runs z score was -1.15 which is quite close to the value of -1.18 obtained in the flash experiment. The brightness experiment also generated a large number of anomalous sequences; 15% had $z < -2$ and 28% had $z < -1.65$. In the orientation study the average runs z score was -1.04, 20% had $z < -2$, and 30% had $z < -1.65$. Finally, in the missing side experiment, the average runs z score was -1.08, 23% of sequences had z scores less than -2, and 34% had z scores less than -1.5. In all four preattentive discrimination tasks the mean runs z score is quite close to -1, and all showed a large number of individually anomalous sequences.

It is also clear from Fig. 3 that the ratio and the fractal discrimination studies produced comparable levels of streakiness. Both fractal studies showed significant departures from a Bernoulli process. The mean runs z score for 2AFC fractal discrimination was $-.30$ which is significantly less than zero ($t(39) = -1.73, p < .046$). Thirteen percent of the sequences in this experiment had $z < -1.65$, which compares well with the 15% found in the ratio study. The mean runs z score for fractal categorization was slightly more negative than that for 2AFC discrimination, $-.67$ versus $-.30$, but this mean was influenced by two large negative outliers ($z < -3.6$). With these two outliers removed, the mean was $-.50$. The medians for the distance ratio, fractal categorization, and 2AFC fractal discrimination experiments were $-.30$, $-.37$, and $-.43$, respectively. A repeated-measures analysis (with task as a between variable) showed that none of the studies in the focused attention group could be distinguished in terms of their level of streakiness ($F(2,9) = .9, p < .44$). The differences between these medians is small compared with those from the four tasks that incorporated pop-out stimuli. We thus have two coherent groupings that are distinguished by a single attention variable.

These results support the conjectures relating attention and streakiness. Not only are the tasks involving preattentive discriminations streakier but also they all share a limiting mean runs z score of -1. The empirical situation appears to be quite straightforward: Sequences deriving from tasks requiring focused attention are not very streaky although they can be distinguished from a Bernoulli process. Sequences deriving from tasks permitting preattentive discrimination are all of one kind and are maximally streaky.

There appears to be an underlying coherence between preattention and streakiness that permits a definite number to be attached to the runs z score. The lower limit of resource allocation that is represented by preattention is apparently reflected as a ceiling in streaky performance. In this situation we are able to make a much stronger claim than is usually found in psychological research—we are able to make a point prediction, a prediction that goes beyond ordinal comparison. This is the strongest form of prediction that can be made; falsification arises if the average runs *z* score in a preattentive task is found to be different than -1.

Streak Formation with Extended Practice: Automaticity

The experiments so far presented have developed the conjecture that attentional demand makes a difference in the amount of streakiness that characterizes the hits and misses in signal detection. Tasks that require attention have been found to be less streaky than tasks that do not. This observation suggests that extended practice in a task might lead to streakier outcome sequences because practice tends to make execution more automatic and less demanding of attentional resources. The validity of this hypothesis is relatively easy to test and there is a highly developed literature at hand that may be used for the assessment.

Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) have developed a methodology that offers a potential context for studying streak formation as a function of practice. In principle, almost any task that can be fruitfully practiced would serve our purposes here, but these authors have established a particular form of visual search as a test-bed for the development of automaticity. In the consistent mapping condition where target and distractor identities remain constant over trials, they argue that an initially effortful and serial process is replaced by an automatic and parallel process (see also Shiffrin, 1988) during extended practice. We shall use their paradigm to assess whether runs *z* scores become more negative as the consistent mapping search task becomes increasingly practiced.

A distinction that we wish to make clear at the outset of this section is that there are two ways that practice can influence the statistic that we are using to analyze streakiness; run production. The first way is the sense in which practice effects are manifest across blocks of trials. Here later blocks are associated with increased skill in the task and the issue of automaticity and attention arises. This is the sense that is pursued in this section. A second way that practice can influence the runs *z* score is by a palpable increase of hit rate within a given block of trials. Since the unit of analysis in our studies is the trial block, a secular trend in hit rate will not be resolved, and such trends will cause run counts to appear low for the average hit rate in the block. This sense of practice effect offers an account for why individual runs *z* scores are negative; that is, as an artifact of learning. We will address practice effects within blocks in the theoretical section where we consider generally what forms of hit rate nonstationarity cause run deficits.

Subjects

Five undergraduates at Vanderbilt University participated. Subjects were recruited through advertisement and were paid $5 per session.

Stimuli

Each trial consisted of the presentation of 20 card images. Card images contained four stimuli at the corners of a square region subtending 5.4 degrees. At two corners there were letters subtending .86 degrees. In the remaining corners a square random dot field subtending .86 degrees was shown. Individual cards were displayed for 17 ms with a 50 ms interstimulus interval. Corners containing letters were randomly selected for each card image.

Design

Twelve letters were divided into a group of 4 that were designated targets and a group of 8 that were distractors. For 3 subjects MNOP were targets and RSTUVWXY were distractors. For 2 subjects HIJK were targets and QRSTUVWX were distractors. Letters from the distractor set were selected at random for the 20 card images that comprised a trial. On half of the trials a randomly selected letter from the target group was displayed on a single (randomly selected) card image. Only cards in the positions 4–17 were permitted to carry targets. Blocks consisted of 250 trials. Subjects generally completed 2 blocks each day and participated on as many consecutive days as required to finish the number of blocks required of them. Three subjects completed 15 blocks and 2 subjects completed 8 blocks.

Results

The results from this experiment are shown in Fig. 4. The two columns depict respectively graphs of hit rate and runs z score as a function of block number. All subjects showed virtually monotonic improvement in discrimination accuracy over the first 8 trial blocks (2000 trials). Initial learning of the task was assessed by comparing performance in blocks 1–4 with performance in blocks 5–8. A repeated measures analysis revealed that the hit rate in blocks 5–8 was significantly larger ($F(1,4) = 24.3, p < .008$). The three subjects that continued for another 7 blocks did not show any substantial further improvement. This is evident from the lack of secular trend in the hit rate oscillations that appear following block 8. The results for the runs z scores are quite different. It is evident that practice has no systematic effect on the development of sequential dependence. There was no significant difference in runs z score between the first four blocks and the second four for the 5 subjects examined ($F(1,4) = 4.6$, p $< .10$). In fact the sign of the effect was opposite to what might be expected; runs z scores were on average more positive ($\Delta z = .24$) in blocks 5–8. A simple regression of runs z score and hit rate confirmed the impression that z scores were uninfluenced by practice; the percentage of variance accounted for was only .03%. However, there was evidence for a level of streakiness consistent with tasks that require attention. The average runs z score was $-.27$ which is significantly less than zero ($t(60) = -1.96, p < .027$). The similarity with the fractal and distance ratio

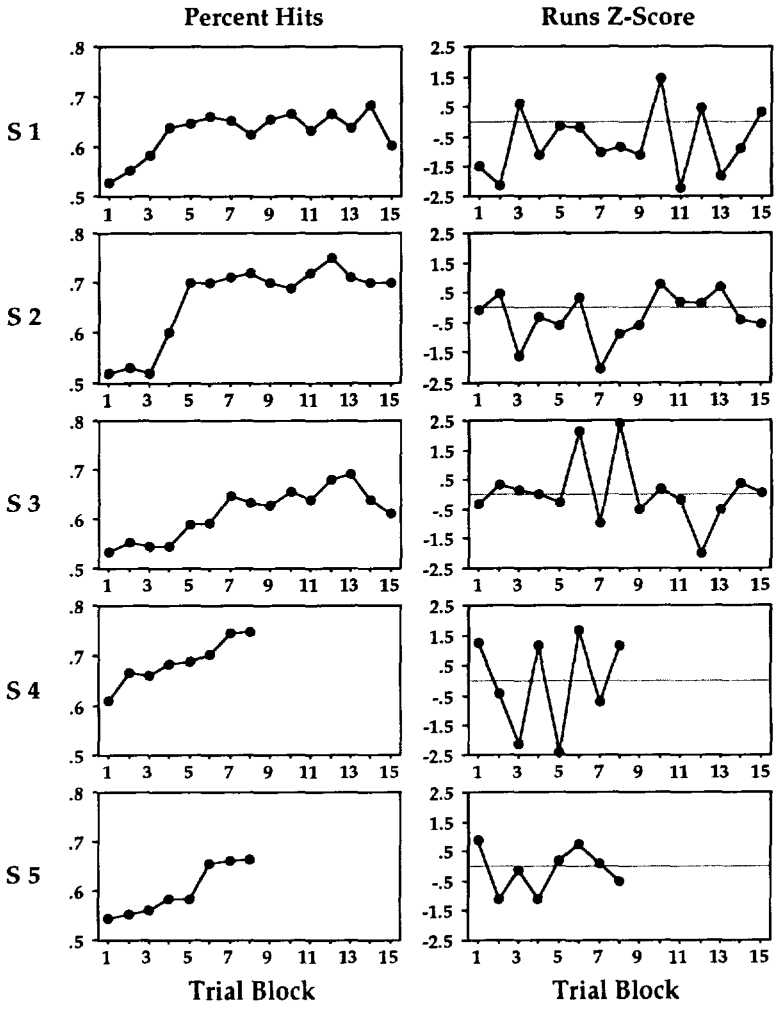


FIG. 4. Hit rates and runs z scores are depicted as a function of trial block for five subjects in the consistent mapping letter search experiment. Learning and increased automaticity, to the extent that it exists, is uncorrelated with streakiness as measured by the runs z score.

studies extended to the filling of the negative tail of the z score distribution; 13% of the sequences had $z < -1.65$.

Discussion

In this experiment we have manifestly failed to find any evidence that extended practice in the Shiffrin/Schneider consistent mapping detection task leads to streakier performance. Rather, the distribution of runs z

scores indicated that attention was required independent of hit rate and experience with the task. These results admit of several interpretations. First, we may have failed to sufficiently practice our subjects so that they did not reach an automatic processing regime. In our experiment subjects did not exceed hit rates of 70–75%, while those in Experiment 1 of Shiffrin and Schneider (1977) achieved hit rates of 90%. This difference may be critical, but our subjects completed as many trials as those in Experiment 1 of Shiffrin and Schneider (1977) and did show nearly monotonic improvement over the first 2000 trials. Shiffrin and Schneider explicitly claimed that following 2100 trials their subjects were utilizing automatic detection. Furthermore, the 3 subjects that received extended training in our experiment clearly saturated in discrimination performance after about 2000 trials. It therefore appears that our subjects had ample opportunity to realize any benefits that practice might bring.

A second interpretation is that the consistent mapping task never becomes completely automatic and that it requires attention at all levels of experience. There is ample evidence for this point of view in that curvilinear reaction time–set size functions in visual search persist even after asymptotic training in consistent mapping (Fisher, 1982, 1984). Shiffrin (1988) argues for an interpretation of load effects wherein consistent mapping involves a hybrid process with parallel and automatic detection occurring within a moving and controlled focus of attention. Evidence that attention consists of two concurrent processes—one effortless and automatic and one effortful and controlled—has also been reported by Weichselgartner and Sperling (1987).

Other researchers have questioned more generally the equation of automaticity and preattention. Logan (1992) argues that the two terms should be kept distinct because they refer to different psychological structures. Following Ullman (1984), Logan identifies preattention with a form of processing that is locally parallel. Automaticity, Logan suggests, is single-step direct-access memory retrieval (Logan, 1988)—an entirely different kind of process than preattention. In this sense, automaticity can be learned to the extent that such memories may be created through repeated exposure. Preattention, however, cannot be learned because it is allied with basic neural architecture. Treisman, Vieira, and Hayes (1992) present a concordant view in their analysis of what takes place as performance in a visual search task improves with practice. They suggest that extended practice does not result in the formation of a new preattentively detectable features, but rather in the accumulation of specific memories for individual stimuli.

The moderate level of streakiness that was found in our letter-search study is only consistent with our earlier results if practice does not produce preattention. Otherwise, the failure to find a relationship between practice and runs z score is a counterexample to the major empirical claims made in this article. However, the lack of correlation between trial block and runs z score, as well as the z score distribution itself, is predicted by our conjectures on the production of streaks if the consistent mapping task always requires focused attention. And it is apparent that the current thinking about attention and automaticity is that Shiffrin and Schneider's original claims were mistaken and that practice does not produce preattention.

TIME DELAY AND STREAK SUPPRESSION

Ovateness detection was the single task that generated runs sequences that were indistinguishable from Bernoulli trials. If some level of streakiness is normative, as appears to be the case, then the absence of streakiness in ovateness discrimination becomes problematic. This task was also distinguished by the large amount of time that elapsed between trials. The presentation program required about 7 s to prepare each animation sequence before it could be shown. In all other tasks the trials were self-paced. In the initial construction of this experiment, this delay was considered to be a nuisance, but not necessarily a relevant variable in streak formation. However, it seemed plausible after reviewing the results, that a long waiting period between trials could generate independence of trials. The qualitative difference in generation time could not *a priori* be ruled out as the cause of the differences in run structure that were found.

Delayed Flash Trials and Minimization of Delay in Ovateness Trials

In order to determine the effect of timing empirically, we artificially imposed a 7 s delay between trials in flash detection, and found a way of generating kinetic depth shapes that required only a 1 s delay. The flash detection task is an appropriate foil for estimating the effect of time delay as it generated sequences with the greatest run deficits.

Subjects

Two subjects were recruited by advertisement. Subjects were paid $5 per session.

Stimuli

The same stimuli were used as in the first ovateness and flash detection studies. Presentation was identical except that a 7 s delay was imposed between flash trials and the delay between ovateness trials was minimized to 1 s.

Design and Procedure

Each subject completed 9 blocks of 200 trials in each discrimination task; 3 blocks at each of 3 levels of hit rate (.6, .75, .9). In all other respects the procedure was as in the earlier experiments.

Results

Box plots of runs *z* scores are shown in Fig. 5 together with the original ovateness and flash data. It is evident from this figure that ovateness discrimination was not influenced by reducing the delay in presentation time from 7 s to 1 s; the mean runs *z* score for fast presentation could not be distinguished from zero ($t(17) = .719, p < .24$). However, the 7 s delay had manifest consequences for streakiness in flash detection where the mean runs *z* score increased from -1.18 (no delay) to $-.54$ (delayed trials). A repeated measures analysis (with presentation speed as a between variable) showed that this increase was significant ($F(1,6) = 8.0, p < .03$). The delayed trial sequences in flash detection were still distinguishable from a Bernoulli process ($t(17) = -3.17, p < .0028$).

Discussion

The observation that ovateness detection was not streaky, even when the time between trials was significantly decreased, admits of two interpretations. The first is that there is a latent variable that is present in ovateness detection that is causing the generation of independent trials, regardless of timing delay. This variable may not be influencing flash detection so that the timing effect is visible. In other words, there may be something intrinsic about the ovateness stimuli *per se* that is unrelated to

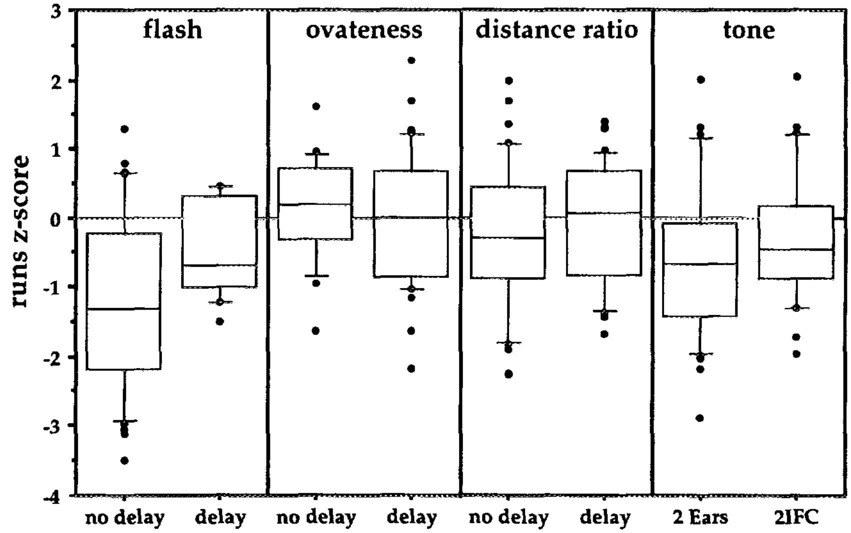


FIG. 5. Box plots of runs *z* score depicting the effect of delayed presentation in the flash, distance ratio and ovateness discrimination studies. Also shown are box plots for the 2IFC and two-ear tone detection studies. Data from the initial set of psychophysical studies are replotted here for comparison.

time that makes flash detection hot and ovateness detection cold. Attentional demand is not the latent variable as we found streakiness in several tasks that required effortful attention.

A second possibility is that reducing the time delay in computing the ovateness animation sequence from 7 s to 1 s was not sufficient. There is, in fact, no way to remove time as a variable from ovateness detection because the stimulus exists only as an extended event in time. Kinetic depth shapes are perceptual entities only when they are moving, and it takes at least one period of oscillation to perceive the shape. Now time clearly is an important component in streak formation as the flash detection part of this study makes clear, and ovateness detection is not expected to be streakier than other tasks which require effortful attention. It may be that the few seconds that elapse during the computation of the animation sequence and the necessary viewing of the stimulus, coupled with the fact that the task is not likely to be very streaky anyway, is sufficient to bring the mean runs z score to zero.

Delayed Distance Ratio Trials

We have tested the importance of time delay in an effortful attention task drawn from the first group of signal detection studies—distance ratio discrimination. The level of streakiness in a task requiring effortful attention is moderate when not delayed and it may be entirely suppressed when delay between trials is introduced. If so, then an account would be provided for the absence of streaks in kinetic depth shape discrimination. In this experiment we chose a delay commensurate with the amount of time consumed in creating and viewing a kinetic depth shape—3 s.

Subjects

Four subjects were recruited by advertisement. Subjects were paid $5 per session.

Stimuli

The distance ratio task was repeated with the imposition of a 3 s time delay between trials. In all other respects the task was unchanged.

Design and Procedure

In all respects the design and procedure were identical to the earlier distance ratio study.

Results

The effect of time delay was pronounced as is shown in Fig. 5. The mean runs z score was brought to $-.04$, indistinguishable from the output of a Bernoulli process. Thus it is possible to render a task that requires effortful attention to resemble a Bernoulli process by imposing a time interval between trials. This result suggests that the absence of streaks in the ovateness task was caused by the intrinsic time delay coupled with the attentional demands required for discrimination.

Discussion

These results indicate that delays of even a few seconds can induce independence in the outcomes between successive trials. This observation may be relevant not only to the ovateness study but also to the tone detection experiment. The nature of a two-interval forced choice paradigm requires that a decision be delayed until both stimuli have been presented. The inevitable elapse of time associated with the two interval design may account for the relatively weak level of streakiness that was found in tone discrimination. A second and equally plausible accounting for the level of streakiness in tone detection concerns the role of memory. In this one experiment, it was necessary for the subjects to rate their impressions of the existence of a signal in stimuli that were not ongoing. The level of streakiness found in this study may be indicative of the attention required for memorial comparisons. We note that this task generated the fewest sequences with anomalously few runs (only 7% had $z < -1.65$), and it is possible that both timing and attention variables are relevant here.

Two-Ear versus Two-Interval Choice in Tone Detection

We created a version of the tone detection experiment that minimized usage of memory and minimized the time consumed by the presentation of stimuli. The design in our first study was a two-interval forced choice (2IFC). The duration of a trial was 1100 ms (300 ms for each interval of sound and a 500 ms interstimulus interval) which we have shown is substantial in the context of suppressing streaks in signal detection. Furthermore, the two-interval design raises the issue of memory in a way that is not present when the stimuli to be discriminated are presented simultaneously. A two-ear discrimination where a pure tone + noise is presented to one ear and noise to the other allowed a design where the trial consumed only 300 ms (the time to present the sounds) and where memory usage was equated to that of the visual discriminations.

Subjects

Four subjects were recruited by advertisement. Subjects were paid $5 per session.

Stimuli

The equipment used and the generation of stimuli were identical to the previous audition study.

Design and Procedure

The two types of stimuli (pure tone + noise, just noise) were presented to different ears. The subjects' task was to identify which ear contained the tone. The ear to which tone was presented was randomized over trials. Subjects were individually calibrated to determine the amplitude of the tone for a 75% hit rate prior to each trial block. Subjects completed 10 blocks of 300 trials.

Results and Discussion

As illustrated in Fig. 5, sequences generated by two-ear presentation were generally streakier than those generated by two-interval presentation. The average runs z scores for the two methods was $-.57$ and $-.32$, respectively. This difference, however, failed to reach significance ($t(11) = -.64$, $p < .27$). However, whereas only 7% of two-interval sequences had $z < -1.65$, 18% of two-ear sequences had $z < -1.65$. Furthermore, no two-interval sequences had $z < -2$ while 8% of two-ear sequences had $z < -2$. The negative tail of the runs z score distribution from the two-ear experiment is similar to that encountered in the distance ratio and fractal discrimination experiments. We conclude that audition is not inherently a weak modality for producing anomalous sequences. Rather, a two-interval design tends to suppress streak formation. It is important to note that even simultaneous presentation of auditory stimuli failed to produce the level of streakiness observed in sequences derived from preattentively discriminated visual stimuli.

RUN LENGTHS OF HITS AND MISSES

In the common parlance regarding streaky performance there is a special role assigned to success. Streaks are, cognitively at least, related to intermittent but unusually long runs of successful trials. It is not clear that the characterization of streaky performance that has been presented here tallies with this notion. A deficit in the number of runs does not imply that hits are preferentially segregated into longer runs than misses. Determining whether or not hits and misses are on an equal footing with respect to run length is an empirical matter that can be decided by computing the respective run length distributions for hits and misses within the sequences observed in the various studies.

The issue of whether runs of hits have a different length distribution than runs of misses has to be approached with some care. The two length distributions deriving from a given sequence cannot be directly compared except in the degenerate case when the sequence has a hit rate exactly equal to .5. If the hit rate is greater than .5, as is generally the case in 2AFC designs, the hit distribution will naturally have a greater mean and variance than the miss distribution. A meaningful comparison of the two distributions must take the hit rate into account.

The appropriate question to ask in this context is whether runs of hits have a more deviant length distribution than runs of misses. In order to address the issue of deviance, it is necessary to characterize the respective length distributions in terms of specific statistical quantities, and to refer the observed values of these statistics to the relevant sampling distributions. Here we characterize the distributions of run lengths in terms of their lower order moments; mean, variance, skew, and kurtosis. The z scores for these four quantities are calculated for both hits and misses by constructing their sampling distributions from sequences that share a common hit rate. We would assert that hits cluster into longer runs than misses within in a given sequence only if the mean of the observed run length distribution of hits was at a larger z score than the mean of the observed run length distribution of misses.

The following procedure formed the basis of an algorithm that calculated z scores for the first four moments of the hit and miss run length distributions. For a given sequence of N observed trials:

1. Compute the hit rate and the frequency distribution of run length for hits and misses separately. Compute the moments for both distributions.

2. Form the ensemble of all sequences of N trials with the given hit rate. In practice, we created ensembles that contained 1000 sequences.

3. For each sequence in the ensemble compute the moments of the hit and miss run length distributions separately. Thus each moment has its own sampling distribution and there are separate sampling distributions for hits and misses.

4. Compute the z scores for the observed moments from their positions in the appropriate sampling distributions.

This procedure was followed for all sequences obtained in the various studies. The results were quite straightforward. In none of the four moments were there any systematic differences in z score between hits and misses. In particular, while the means of the observed hit and miss run length distributions tended to be located in the positive tails of their respective sampling distributions, their z scores did not differ. The equality in z score over the first four moments makes it quite clear that hits and misses cluster in an identical fashion relative to the frequency of their occurrence. Consequently, while our results demonstrate the existence of streaky performance, they do not provide evidence for flow states, "being in the zone," or any assessment of performance that focuses on hits.

The cognitive assessment that performance is "hot" usually follows an extended period of hits in the context of observer knowledge of the background hit rate. In our studies, a subject could be hot in this sense without it being detected. There are two reasons for this. First, each sequence is analyzed by comparing it only with other sequences that share its hit rate. A sequence of trials where there are an excessive number of hits is referred to the ensemble defined by that hit rate, and the question that is posed is: Relative to the number of hits in this sequence, are the runs unusually long and sparse? This question is relevant to streakiness in the context of the sequence taken in isolation, but ignores the possibility that this sequence is unusual in itself by virtue of its hit rate. Secondly, in our studies we have essentially precluded the possibility of unusual performance by calibrating subjects to perform at a specific hit rate. If a subject is having a good day for signal detection, the task presented to them will be correspondingly more difficult. In this way a subject could be streaky in the sense of having a few days of unusual sensitivity, but this sensitivity would not necessarily be manifest in sequence structure nor in their calibrated hit rate.

THEORETICAL ANALYSIS

The formation of streaks appears to be a natural outcome of making discriminations in signal detection. The implication of this finding is that the processes governing basic perceptual sensitivities are nonstationary. The probability of a hit or miss on a given trial is in some way related to (1) the outcome on earlier trials or to (2) where the trial is located in the overall context of the activity. These two modes of nonstationarity are distinct. Although the statistical signature of streakiness is positive sequential dependency in hits (and misses), this does not imply that hits cause hits. Hits could induce hits if the subject is aware of when a hit occurred and if this awareness led to enhanced performance. However, it may be that positive sequential dependency arises from nonstationarity in the operator that is unrelated to prior outcome. Learning is a candidate in this regard. If the subject learns as much from successful trials as from failures, then outcome *per se* is not the relevant variable. Rather the sequential dependency would be related to trial number; in early stages misses would generally follow misses, and in the latter stages hits would generally follow hits. Intermittent episodes of boredom or fatigue could also cause nonstationary performance that would not necessarily be linked to outcome. In this section we differentiate between these two forms of nonstationarity and attempt to characterize the underlying cause that is creating the clustering of outcome that pervades our data.

We have considered four models of nonstationarity in hit rate that could plausibly account for streaky performance. Briefly, these are:

Learning

A secular improvement (or worsening) over the course of a trial block can create clustering of hits and misses. For example, a sequence of trials where there is secular improvement will initially have relatively long runs of misses followed by relatively long runs of hits. Such a sequence could appear to have anomalously few runs if the expectation is derived from the average hit rate.

Wave Modulation

Streaks may arise if hit rate is modulated by wave-like fluctuations in attention, ability, or effort. In such a model, runs of hits will preferentially occur in wave crests, and runs of misses will preferentially occur in wave troughs.

Intermittent Effort

This is a discretized and stochastic version of the wave modulation model. In this model, there are two states of effort or ability that are distinguished by hit rate. Transitions between the two states are conceived to be probabilistic. This model is intended to capture the notion that boredom or ennui may influence run structure. Hits will tend to concentrate when the subject is paying attention relative to moments when the subject is off-task. An equivalent interpretation is that mundane performance is punctuated by periods of inspiration.

Markov Process

Part of the folklore of the hot hand is that increased confidence plays an important role in shooting performance. In informal conversations with a number of basketball coaches and varsity athletes, the notion that success breeds success was often used as an explanation of the hot hand phenomenon. We have therefore created a model where performance is presumed to be conditional: people try harder, have more confidence, or pay more attention following successful trials. This model naturally generates positive sequential dependency by building in correlation between successful trials. The Markov model may be thought of as a multi-state model where transitions between the states is a random variable depending on outcome.

The manner in which the different models were evaluated depended on the role of chance. The learning and wave modulation models are deterministic in that the hit rate can be specified exactly for all trials once the parameters of the model are fixed. This feature permits a regression analysis where the serial correlation in outcome between trials may be evaluated with the model factored out. In contrast, the intermittent effort and Markov models are inherently stochastic because it is not possible to predict what the hit rate will be on any given trial, even for specified model parameters. Evaluation of these models is less straightforward and requires that they be rendered through Monte-Carlo simulation. In what follows we describe the procedures that were developed to ascertain which model provided the best fit to the data.

Analysis by Part Serial Correlation: Learning and Wave Modulation

In this section we will use the serial correlation between successive trials as a statistical measure of streakiness. The serial correlation is absolutely equivalent to the runs *z* score that has been used as a measure of clustering in the analysis of our data. Calculation of the Fisher *Z* score for serial correlation and the *z* score for runs within all sequences and in all experiments has shown that these two measures of structure have nearly unit correlation.

In the learning and wave modulation models, where the hit rate can be specified in advance for all trials, it is possible to compute the part serial correlation with the model explicitly factored out. To the extent that the model is a correct description of what is causing nonstationarity in hit rate, the part serial correlation will be smaller than the serial correlation. In the limit that the model is a complete description of the hit rate structure, the distribution of Fisher *Z* scores for the part serial correlation will be normal with zero mean (this was explicitly checked by Monte-Carlo simulation of sequences with hit rate specified by learning and wave models). The goal of this analysis is therefore to ascertain how much smaller the part correlations are than the raw serial correlations. We will illustrate this technique with a simple example and then generalize it to allow for the fitting of model parameters.

A natural way to evaluate the role of learning is to factor out trial number from the serial correlations. The correlation coefficient between trial number and outcome is a measure of the difference between the average trial number for hits and the average trial number for misses. If learning is occurring during trial blocks, then the average trial number for hits should be larger than that for misses, leading to a positive correlation between the sequence of outcomes and trial number. We refer to this type of learning as linear because hit rate is conceived to be proportional to trial number. Let r_{12} = the serial correlation, r_{1M} = the correlation between the sequence and the model, and r_{2M} = the correlation between the sequence lagged by one trial and the model. Then the part correlation is

$$r_{\text{part}} = \frac{r_{12} - r_{1M}r_{2M}}{\sqrt{1 - r_{2M}^2}} .$$

If learning is occurring then both r_{1M} and $r_{2M} > 0$, and $r_{\text{part}} < r_{12}$.

We have computed the part serial correlations for those studies where the greatest streakiness was observed; discriminations of flash, brightness, orientation, and missing side. The average serial correlations for those sequences with negative runs *z* scores in the four studies are shown in the second column of Table 1. Although the magnitudes of these cor-

TABLE 1
Correlation Analysis

		Model for partial correlations		
Experiment	Serial correlation	Learning/linear	Learning/power law	Wave
Flash	.078	.070	.067	.061
Brightness	.058	.051	.047	.043
Orientation	.064	.056	.054	.047
Missing side	.085	.074	.070	.060
<Decrement>		12%	17%	26%

relations may not appear large, it should be noted that in this context, serial correlations of order .10 correspond to runs z scores more negative than -2.0. Column 3 shows the average part correlations with trial number (linear learning) factored out. The reductions in correlation achieved by factoring out trial number are of order 10% and are certainly not large enough to be conclusive. In order to place these magnitudes into context, we have generalized the learning model and compared it with a general wave modulation model.

The linear learning model just described is but one possible form of monotone increase in hit rate. A generalized learning model can be constructed by considering the family of power functions of hit rate:

$$\text{hit rate} \propto (\text{trial number})^{\beta},\ \beta > 0.$$

This class of learning models spans the range of learning curves that are everywhere convex or concave. In Table 1 we refer to these models collectively as power law learning models. For each sequence with positive serial correlation in each of the four relevant studies, we have determined the optimum β for the smallest part serial correlation in absolute value. Average part correlations for power law learning models with optimum β are shown in Column 4. The difference between the optimum model and the linear model considered above is not expected to be large because of the high correlation between power laws and a linear function.

This analysis was repeated with a wave modulation model. In this model, the hit rate is conceived to vary as

$$\text{hit rate} \propto \sin(2\pi k/\text{L} + \theta),$$

where k is the trial number, L is the wave period, and θ is the phase. Here both L and θ entered as free parameters in fitting to a minimum part correlation. Part correlations for optimized wave models are shown in Column 5. The key point to be derived from Table 1 is that the wave modulation model makes greater reductions in the serial correlations than do the learning models. In further analyses, the learning models were abandoned in favor of the wave modulation model.

Analysis by Monte-Carlo Simulation: Wave Modulation, Intermittent Effort, and Markov Chains

Models that conceive of performance in terms of stochastic occupation of discrete levels of hit rate states require an indirect procedure of evaluation. It is not possible to factor the model out of the serial correlations in those cases where the model is only a recipe for constructing an ensemble of realizations. Analysis must proceed by actually simulating the ensemble of sequences that can arise in the model and then comparing these sequences with data. This procedure can also be adopted for deterministic models since the hit rate on a given trial, even if it can be specified exactly, only determines the probability of hit—not the exact locations of hits within a sequence. In this section we develop criteria for evaluating models of nonstationarity within the framework of Monte-Carlo simulation.

Models and Associated Algorithms

The wave modulation, intermittent effort, and Markov models were simulated in order to create ensembles of sequences consistent with their respective logics and then to create sampling distributions of statistical variables that could be compared with data. The rules describing the algorithms are given below and summarized in Table 2.

Intermittent effort. In this model there are two states of effort that have corresponding hit rates b_0 and b_1. If the algorithm is in the low effort state, then with probability p_{high} there is a transition to the high effort state. Alternatively, if the algorithm is in the high effort state, then with probability p_{low} there is a transition to the low effort state. The hit rate in the high effort state is $b_1 = b_0(1 + \Delta)$. Trials were simulated by picking a uniform random deviate r on the interval (0,1). In the high effort state the trial is a success if $r < b_1$. In the low effort state the trial is a success if $r < b_0$. After each trial, the algorithm then decides which state it is going to be in on the next trial.

Markov process. We have constructed a gain function that takes into account whether there have been successes on the previous one or two trials. In these second order Markov models, the algorithm looks back two trials and there is a gain $G = \Delta$ if only the previous trial was successful, and a gain of $G = \Delta(1 + \delta)$ if both trials were successful. If the previous trial was a miss, then $G = 0$. On any given trial the instantaneous hit rate is $b_0(1 + G)$. It would be possible to elaborate this model still further by allowing a graded gain function that took into account the previous N trials, or by discriminating failure on the two previous trials

TABLE 2
Theoretical Models

Model	Hit rate function	Parameters
Intermittent effort	with probability p_{low} make transition into low state: $b = b_0$ with probability p_{high} make transition into high state: $b = b_1$	$p_{\text{low}}, p_{\text{high}}, \Delta$
Increment rule:	$b_1 = b_0(1 + \Delta)$	
Markov 2nd order	$[s_{k-1} = 1 \;\&\; s_{k-2} = 1] \supset [b_k = b_2]$ $[s_{k-1} = 1 \;\&\; s_{k-2} = 0] \supset [b_k = b_1]$ $[s_{k-1} = 0] \supset [b_k = b_0]$	Δ, δ
Increment rule:	$b_2 = b_0[1 + \Delta(1 + \delta)]$, $b_1 = b_0(1 + \Delta)$	
Wave continuous	$b_k = b_0[1 + \Delta\sin(2\pi k/\mathrm{L} + \theta)]$	Δ, L
Wave two-state	$b_k = b_0(1 + \Delta)$ $0 < [2\pi k/\mathrm{L} + \theta] \pmod{2\pi} \leq \pi$ $b_k = b_0(1 - \Delta)$ $\pi < [2\pi k/\mathrm{L} + \theta] \pmod{2\pi} \leq 2\pi$	

Note. b_0 is the nominal hit rate; b_k is the probability of success on trial k; s_k is the outcome of trial k: 1 if success, 0 if failure; θ is a constant uniform random deviate on the interval $[0,2\pi]$; and L is measured as number of trials.

from failure on only the previous trial. Either of these elaborations requires the addition of free parameters which is not justified at this exploratory stage of analysis. Note that this class of models essentially regards performance as arising from three states; a low state with hit rate b_0, and two high states where there is a gain depending on the outcome of the previous two trials. It is not necessary in these models therefore to have specific rules for trials on which the previous one was a failure. Failure breeds failure by virtue of maintaining the algorithm in the low state. Trials were simulated as above, with success if a random uniform deviate $r < b_k$, where b_k is the hit rate for the state appropriate to the kth trial.

Wave modulation. Two forms of wave modulation were simulated; continuous sinusoidal variation and a telegraph signal based on transitions between two states. Conceived as a continuous wave, the hit rate on trial k is

$$b_k = b_0(1 + \Delta\sin(2\pi k/L + \theta)).$$

Alternatively, a discretized wave has the form

$$b_k = b_0(1 + \Delta)\ 0 < [2\pi k/L + \theta](\mathrm{mod}\ 2\pi) \leq \pi$$
$$b_k = b_0(1 - \Delta)\ \pi < [2\pi k/L + \theta](\mathrm{mod}\ 2\pi) \leq 2\pi.$$

In these expressions L is the period of the wave and Δ is the amplitude. These two parameters are constant and are regarded as fixed properties of the task. θ determines the initial phase of the wave and is also constant. θ varies randomly across an ensemble of sequences. Trials were simulated as above with success if $r < b_k$.

Window Structure in Observed Sequences

Evaluation of the models requires the development of a sensitive measure of the way in which hits and misses are distributed in the data sequences. We adopt here a local measure of sequence structure that was used by Gilovich et al. (1985) in tests of nonstationarity. The motivation for the introduction of this measure in Gilovich et al. was that one signature of streakiness might be an excessive number of time intervals that are dominated by successful trials. One way to formalize the degree of domination is to take a window of size N and to count the number of such windows containing all hits. The null hypothesis of independent trials gives a predicted frequency of such pure windows, and the difference between the observed and predicted frequencies is a measure of streakiness.

We have used window counts not as an index of streakiness, but as a measure of local sequence structure. Specifically, we take nonoverlapping windows of size N and count all instances where there were $K = 0, 1, 2, 3, \ldots, N$ hits within a window. Following Gilovich et al., we perform this sum over each sequence N times; the first window commences at either $s_1, s_2, s_3, \ldots$ or s_N, where s_i is the ith member of the sequence. From the average hit rate of the sequence, b, we compute the expected number of windows of size N that contain K hits using the binomial distribution. The difference between the expected number and the observed number is the frequency excess for each hit number K. Finally, we transform the frequency excess to a probability excess by normalizing by the number of windows. The probability excess is positive or negative depending on whether the number of windows of size N with K hits was more or less numerous than expected by chance.

Models were fit to the data from all experiments that showed a large excess of individually anomalous sequences. This set includes the flash, brightness, orientation, missing side, distance ratio, and fractal studies (we present results only for the left/right 2AFC method of presentation as the sequential presentation generated similar models). As examples of the window statistics that will be used to assess models, Fig. 6 illustrates the excess probabilities for encountering K hits in windows of sizes 4, 6, and 8 in the flash detection experiment. Each data point is an ensemble average over the entire collection of sequences observed in the experiment.

There are several features in the patterns of excess probability that are of interest. The structure of the window pattern is driven by the large

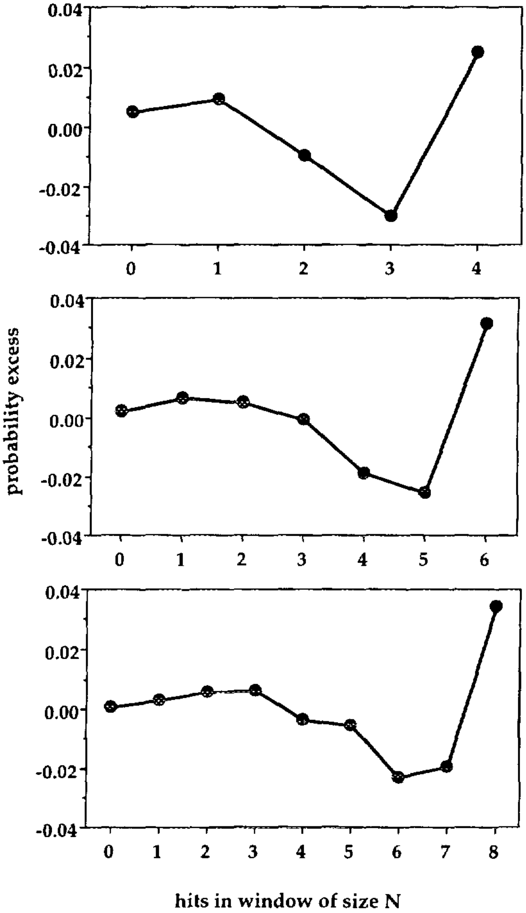


FIG. 6. The probability excess of encountering K hits in windows sizes 4, 6, and 8 in sequences generated by flash detection. The probability excess is computed relative to the expected probability under the null hypothesis of Bernoulli trials.

excess of windows with *N* hits. Since the total number of windows is fixed, excesses must come at the expense of other windows. The *N*-hit windows are drawing primarily from windows that would have had *N*-1 or *N*-2 hits. On the other hand, there is very little excess at windows with 0 hits and this creates a large asymmetry. In signal detection, this pattern is enforced by the placement of a hit rate floor at .5. For hit rates in excess of .5, permutations of hits and misses will naturally tend to preferentially make larger aggregates of hits; the failures tend to be too sparse to generate many windows of *N* misses.

It must be emphasized that the magnitudes of the probabilities are computed as *excesses* relative to the expectation under the null hypothesis of statistical independence. A sequence with a large hit rate does not make a positive contribution to the excess probability unless more windows of a given type are encountered than expected. In testing the codes for computing these statistics it was necessary to ascertain that the excess probabilities were zero for random Bernoulli trials independent of hit rate.

Independence of Local and Global Measures of Structure

The runs z score is a global measure of streakiness that imposes few constraints on the local sequence structure. There are many ways of assembling strings of hits and misses that will have a single value of run number but will vary in the number of windows of size N with K hits. For example, consider the following two sequences:

A 0 0 1 1 0 0 1 1 0 0 1 1
B 0 0 0 0 1 1 1 1 0 1 0 1

Both sequences have the same hit rate (.5) and the same number of runs (6). Limiting the comparison to windows of size 4, sequence A has 9 windows with 2 hits and no window has 0, 1, 3, or 4 hits (remember that windows are counted in 4 passes through the sequence, each pass staggers the initial window position by 1). Sequence B has 1 window with 0 hits, 1 window with 1 hit, 3 windows with 2 hits, 3 windows with 3 hits, and 1 window with 4 hits. The lack of correlation between runs z score and the window counts was also observed in our data. As an example consider a window of size 4 in the flash detection experiment. For this study the correlations were $r^2 = .26, .17, .05, .29$, and $.54$ for 0, 1, 2, 3, and 4 hits, respectively. Substantial correlations existed only with the probability of encountering windows with all hits. This lack of correlation will be further demonstrated below when we compare different models in their ability to reproduce the window structure.

The window probabilities are also fairly independent of each other. In general, substantial correlations ($r > .85$) existed only between the probability of encountering a window with N hits and a window with N-1 hits. The average level of unsigned correlation between windows was of order $|r| \sim 0.3$. We will consider the sequences from the flash experiment for $N = 8$ in detail as an example. For this study, the maximum correlation existed between $K = 8$ and $K = 7$ and was $r = -.87$. The next largest correlation existed between $K = 4$ and $K = 2$ and was $r = -.67$. 32 of the 36 possible correlations between the 9 window types (K = 0, 1, . . . , 8) had $|r| < .54$. The mean unsigned correlation was .3 and the median was .28. These results are typical of the other experiments. Although the window probabilities are not strictly independent of each other, only 1 or 2 degrees of freedom are lost in fitting models to the window statistics. In any case, the nonindependence of the window probabilities exists equally for all models, and does not affect the basic strategy of using these probabilities to differentiate between models.

The relative independence of the different statistical measures can be capitalized upon to provide rigorous tests of the theoretical models. Models are constructed by selecting parameters that yield ensembles with runs z scores that are similar to the data. These restricted models are then evaluated in terms of the observed window probabilities. In our analyses we have used windows of size $N = 4, 5, 6, 7,$ and 8. This range was sufficient to distinguish between the performances of the theoretical models. These models have only 2 (wave, Markov) or 3 (intermittent effort) degrees of freedom. The statistical measures together have of order N-1 degrees of freedom. The disparity between the number of free parameters in the models and the number of statistical measures they are required to fit, makes these tests challenging. As will be shown, it is a trivial exercise to fit the runs z score distributions, but models rarely produce window statistics that resemble the data.

Ensemble Construction

The three algorithms under consideration each have a number of free parameters in addition to requiring specification of the nominal hit rate b_0. Definition of these parameters is given in Table 2. For each choice of parameters and hit rate there is an associated ensemble of sequences. Denote these ensembles as $E(P{:}b_0)$, where P is the set of parameters. The experimental data does not have sufficient resolution in hit rate to support detailed comparisons as a function of hit rate. The experiments were designed, however, to have uniform coverage of hit rate within specific limits. These limits are given in Table 3. The ensembles of sequences that were collected in the experiments are modeled by collapsing over b_0 in $E(P{:}b_0)$ to generate the collections $E(P{:}l \leq b_0 \leq u)$, where l is the lower limit in observed hit rate, and u is the observed upper limit.

Each simulated ensemble $E(P{:}l \leq b_0 \leq u)$ consisted of 500 sequences of 500 trials that uniformly sampled the appropriate hit rate interval. The number of sequences and their length were chosen to generate probabilities for simulated window counts that had a minimum of noise for a reasonable amount of computation time. Extended computations using 1000 sequences did not show significant differences. Each algorithm has its own idiosyncrasies that required further specification. For concreteness we spell these out in detail.

Intermittent Effort: select p_{low}, p_{high}, and Δ. The simulation is initialized in the low state with $b = b_0$.

Markov second order: select Δ, δ. The simulation is initialized in the low state with $\mathrm{b} = \mathrm{b}_0$.

TABLE 3
Monte-Carlo Simulations

			Model parameters						
			Wave[a]		Markov[b]		Intermittent effort[c]		
Experiment	<runs z score>	Hit rate range	L	Δ	δ	Δ	p_{high}	p_{low}	Δ
Flash	−1.18	$.50 < b < .90$	21	.20	.12	.07	.10	.22	.40
Brightness	−1.15	$.65 < b < .85$	21	.19	.10	.07	.08	.22	.33
Orientation	−1.04	$.65 < b < .85$	24	.16	.10	.07	.10	.24	.35
Missing side	−1.08	$.50 < b < .90$	20	.20	.14	.05	.16	.28	.45
Distance ratio	−.24	$.50 < b < .90$	17	.10	.12	.02	.12	.18	.15
Fractal: 2AFC	−.30	$.60 < b < .80$	20	.12	.10	.03	.08	.22	.20

[a] L = wave period in units of number of trials, Δ = wave amplitude.

[b] Δ = gain for hit on previous trial, $\Delta(1 + \delta)$ = gain for hits on two previous trials.

[c] p_{high} = transition probability into high state, p_{low} = transition probability into low state, Δ = difference in hit rate between the high and low states.

Wave: select L, Δ. Each sequence is assigned a random phase $\theta \in [0,2\pi]$.

Evaluation of Theoretical Models

Ensembles of sequences were constructed using the wave, intermittent effort, and Markov algorithms. Parameters were initially varied to construct ensembles that yielded the same average runs z scores that were found in the flash, brightness, orientation, missing side, distance ratio, and fractal discrimination tasks. In all cases it was found that there were many sets of parameters, P, that yielded the same average runs z score. Further search was required to find parameter values that also optimally reproduced the window statistics in the individual experiments. Optimal parameter values are shown in Table 3.

The three models were evaluated in terms of a coefficient of fit that reflected how accurately they reproduced the excess probabilities for K hits in windows of size N. This measure of goodness-of-fit is given by

$$c_N = 1 - \frac{\sum_{K=1}^{K=N} (p_K^{\text{model}} - p_K^{\text{obs}})^2}{\sum_{K=1}^{K=N} (p_K^{\text{obs}} - \langle p \rangle^{\text{obs}})^2},$$

where p_K is the probability excess. The average excess, $\langle p \rangle^{\text{obs}}$, is necessarily identically equal to zero by virtue of the fact that the sum over the number of windows with K hits must equal the total number of windows.

The results for the seven experiments that were modeled are shown in Figs. 7, 8, 9, and 10. Figures 7 and 8 depict the behavior of the coefficient of fit as a function of the window size for the preattentive (flash, brightness, orientation, missing side) and effortful attention (distance ratio, fractal) tasks respectively. Figures 9 and 10 give detailed comparisons of the fits for windows of size 8 for both sets of tasks. We found that for matched (L,Δ) parameters, the two-state and continuous forms of wave modulation were virtually indistinguishable in terms of fits to the window probabilities. Results are shown for the continuous variation models.

The most salient result from the Monte-Carlo simulations is that the wave model performed unexpectedly well in the preattentive regime: the coefficient of fit exceeded .95 for all window sizes in the flash, brightness, and missing side studies and exceeded .90 for all window sizes in the orientation study. This is a strong result and should be placed into perspective. First, we stress that the window statistics are relatively independent of the runs z score as a measure of sequence structure and provide a separate assessment of model performance. As the graphs of the other models and the fits displayed in Figs. 9 and 10 make amply clear, there is absolutely no reason to expect that the window statistics could ever be fit simultaneously with the mean runs z score. Second, *it is not*

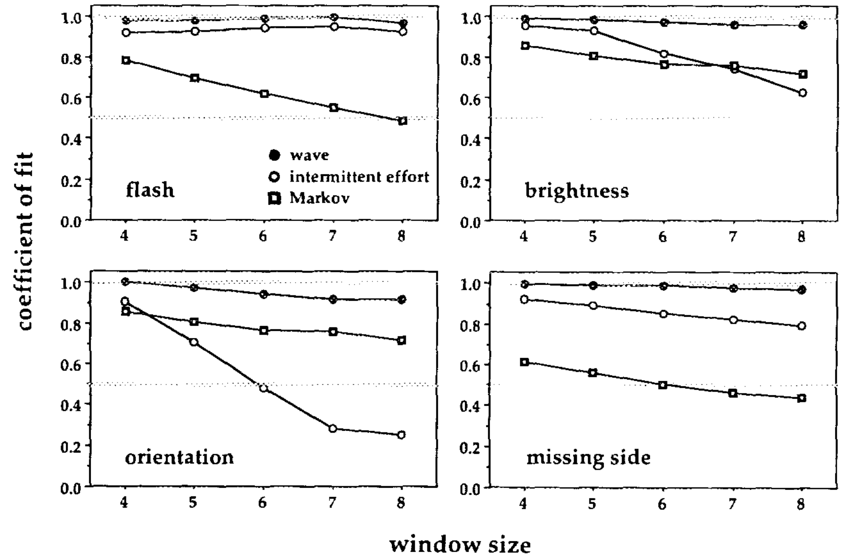


FIG. 7. The window statistics coefficient of fit for three theoretical models is shown as a function of window size. Results are given for the four preattentive tasks. The coefficient of fit is defined in the text. The wave process provides a nearly perfect fit in each case.

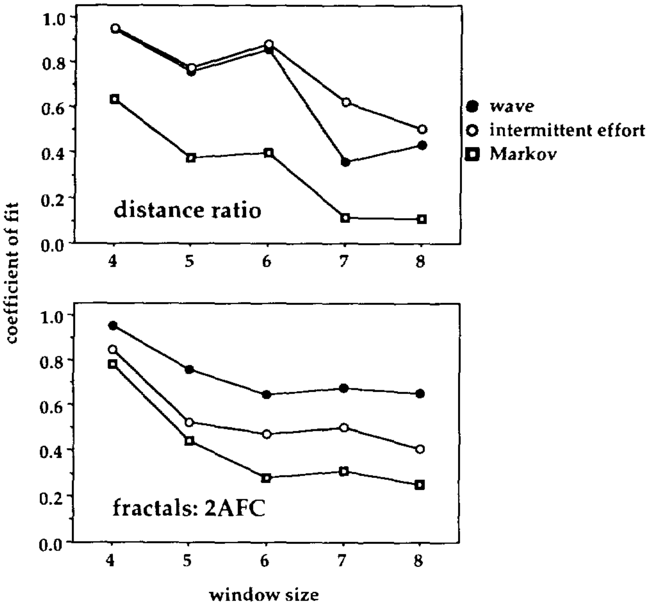
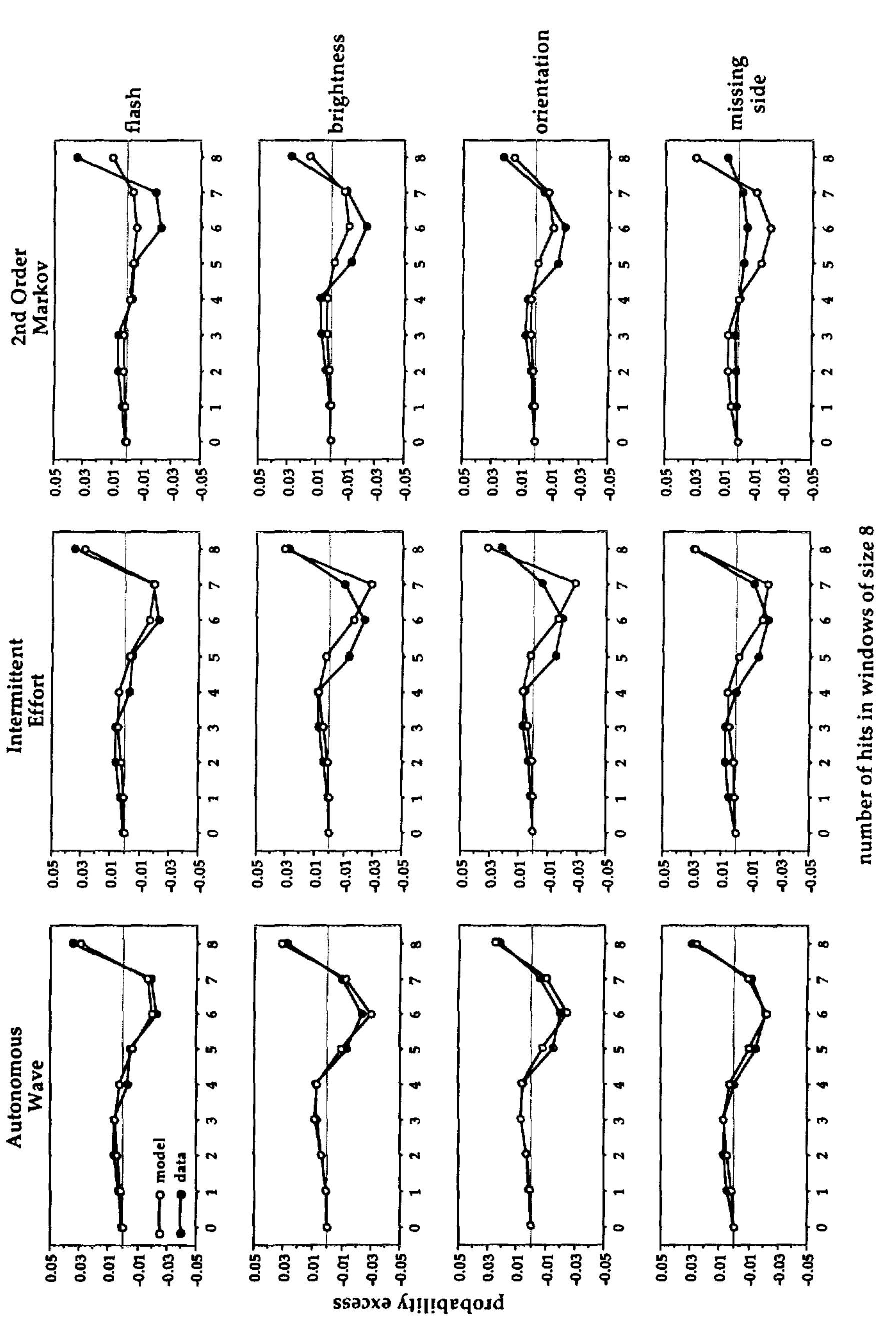


FIG. 8. The window statistics coefficient of fit for three theoretical models is shown as a function of wijndow size. Results are given for two focused attention tasks. The coefficient of fit is defined in the text.

obvious that patterns of hits and misses in human performance could be modeled by a simple algorithm. If we had found that human performance was like independent Bernoulli trials, then it could have been trivially modeled. However, the empirical studies make it manifestly clear that human performance is not Bernoulli-like and there is no *a priori* expectation that the algorithmic instantiation of a simple notion would generate anything that looks like data. The degree of fit to the window statistics produced by the wave modulation algorithm is unexpected, especially in a context where the model does not specify outcome but only the probability of an outcome. Despite the rhetorical advantage that might be gained by inventing a rationale for the inclusion of wave modulation as a model for consideration, it is nevertheless the case that its inclusion was a serendipity.

The intermittent effort model also performed well in the flash experiment, but progressively failed with increasing window size in the brightness, orientation, and missing side experiments. It is not surprising that the intermittent effort model should have some success. For the transition probabilities that were simulated, the model behaves as an incoherent wave with a period of about 10 trials. Despite the fact that this model has an additional degree of freedom, it still does not perform as well as the


Autonomous Wave
Intermittent Effort
2nd Order Markov
flash
brightness
orientation
missing side
model
data
probability excess
number of hits in windows of size 8
0.05
0.03
0.01
-0.01
-0.03
-0.05
0
1
2
3
4
5
6
7
8

wave model. The relatively poorer performance of the intermittent effort model is also evident in models of tasks requiring focused attention. The decrement is particularly large for the fractal discrimination experiment.

The Markov model generates the same pattern of failure across the four preattentive experiments. As window size increases, the coefficient of fit decreases. The failure of this model is not unexpected as none of these experiments incorporated feedback. However, the fits were no better for the ovateness discrimination study which did contain feedback. Furthermore, if the Markov model is correct in its motivation that success breeds success, then we should have observed a greater tendency for streaks in ovateness detection than in flash detection. Yet ovateness detection generated sequences that conformed to the expectation of a Bernoulli process. The Markov process is evidently not a viable model.

We have attempted to find converging evidence for wave modulation by submitting the outcome sequences to Fourier analysis. Ideally, the existence of a wave train would be evident by enhanced power at the frequency (inverse period) at which the part correlations were minimized; i.e. at that frequency where the wave model fits optimally to the data. This effort was frustrated by the fact that we are dealing with a rather small effect. The level of streakiness that we have observed in our studies is sufficiently large to isolate statistically, but the serial correlations at best account for only about 4% of the sequence variability. The wave model gives the best accounting for hit rate nonstationarity, but the absolute magnitude of the wave amplitude is too small to isolate through Fourier techniques.

The Origin of Streaky Performance

There are four results that have been obtained in these studies that together provide some insight into what causes streaky performance; two empirical, and two derived from numerical simulation. The empirical results are that streakiness is graded by the amount of attention that is brought to the task, and that preattentive discriminations form a special class defined by maximal streak production. The first analytic result is that the local sequence structure derived from tasks that are associated with preattentive discrimination are naturally fit by a wave modulation of hit rate. The second analytic result is that the fits are poorer when the task clearly requires effortful attention.

These results point to theory of performance that isolates two different systems with distinctive time histories. The first system is independent of

Fig. 9. Probability excesses in windows of size 8 produced by simulations of theoretical models are compared with data derived from the flash, brightness, orientation, and missing side experiments.

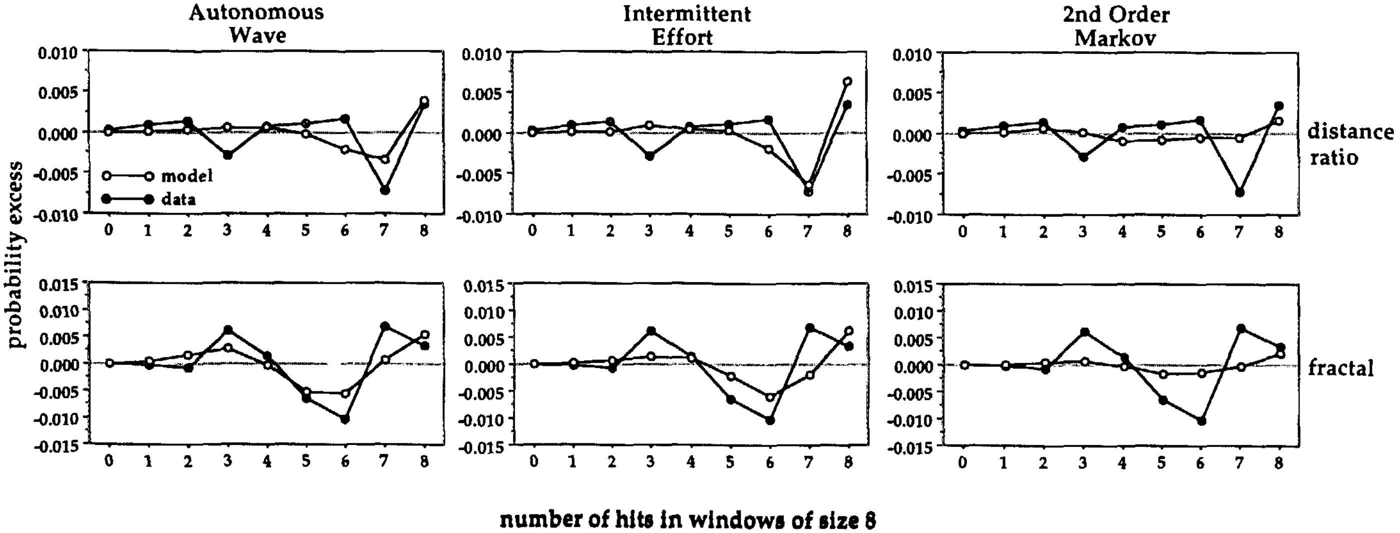


FIG. 10. Probability excesses in windows of size 8 produced by simulations of theoretical models are compared with data derived from the distance ratio and fractal (2AFC presentation) experiments.

the controlled aspects of attention and appears to be associated with early sensory and perceptual processing. In this system we posit small wave-like variations. The second system is logically central and is associated with controlled attention. Controlled attention is conceived here as a filter that does not have invariant fidelity over time, but rather fluctuates randomly depending on a myriad of cognitive and somatic factors. These two systems are capable of accounting for both the empirical and analytic results reported in this article.

In this two-system model, streaks are most clearly seen in tasks incorporating preattentive stimuli because it is in this regime that there is a minimal level of activity in the usage of attentional resources. Essentially, the stochastic filter is bypassed in preattentive discrimination, and the intrinsic wave-like modulation impressed by sensory/perceptual systems is visible as nonstationarity in hit rate. In discriminations requiring focused attention, random fluctuations in attentional resources mask the background wave variations at the level of response. Note that mask here does not mean eliminate; wave-like variations are present albeit at relatively reduced amplitude and residual streakiness is observed.

A formalization of this theory is given by the expression:

$$\text{hit rate}(t) = F(\text{attention},t) + \eta\sin(2\pi t/\Lambda) + c,$$

where F is the attentional filter and is a monotonically increasing function of attentional demand, c is a constant that characterizes the distal level of task difficulty, and t is time. The sine component reflects the contribution from sensory/perceptual processes not under direction of controlled attention. The wave amplitude, η, and its period, Λ, are considered to be constants that are set by intrinsic operator characteristics and are independent of task. As a function of time, F is considered to have stochastic fluctuations which are on the order of its mean magnitude. Where attention must be focused, F is large and consequently the wave component adds into a background that is varying on scales larger than its intrinsic amplitude. In the regime of preattention, F is small, and the wave component is manifest.

The intrinsic wave parameters (η,Λ) are distinguished from the derived model parameters (Δ,L). Only in the case when F is small, and the wave part of the hit rate function alone modulates performance, are Δ and L measures of η and Λ. This condition is met for all preattentive discriminations, and for this account to be meaningful, it is necessary that the simulated wave parameters be invariant over experiments in this class. In this regard it is significant that the flash, brightness, missing side, and orientation discrimination studies were all fit by similar model parameters; $L \sim 20$, $\Delta \sim .2$. The interpretation of Δ and L is different when F is large and the hit rate fluctuates around a constant value of c on top of a small wave variation. As the wave model does not have algorithmic instructions that take into account intermittent and large scale fluctuations, it treats the entire variation as if it were wave-like, and can only respond to a reduction in streakiness by decreasing Δ. The values of Δ and L are simply artifacts of the model in the simulations of tasks requiring effortful attention.

The existence of a definite wave amplitude in this theory has important consequences for streak production. η sets the limit of streakiness that can be produced. Tasks which permit the use of preattentive discrimination express the full amplitude of the wave. This theory makes the important prediction that tasks will not be found that have a level of streakiness greater than that found in the preattentive signal detection experiments. A corollary prediction is that all preattentively performed tasks will be fit by a wave model with $\Delta \sim .2$.

This work is hardly the first to suggest that there are oscillations in early sensory and perceptual systems. Psychologists in the late 19th and early 20th centuries (Urbantschitsch, Lange, Münsterberg, Eckener are a few) were concerned with issues that bear a remarkable resemblance to those raised here. In the early days of sensory psychophysics a number of investigators were puzzling over the fact that faint sounds, gentle touches, and dim lights appeared to wax and wane in intensity over time. Although the language, theories, and stimuli that derive from this epoch are somewhat foreign, these early findings are quite relevant to our account of streaks in threshold signal detection. In particular, the oscillations in perceived intensity exist primarily at threshold and their periods are of order a few tens of seconds—the period implicated by our theoretical models. The issue at the turn of the century was whether the oscillations were associated with fluctuations in attention or with peripheral sensory characteristics such as adaptation. Guilford (1927) presents an overview to this field and a number of experiments which point to a peripheral etiology; eye movements and retinal adaptation are argued to cause the oscillations in vision. This observation suggests that streaks may be also be explained as artifacts of adaptation.

There are a number of reasons to reject retinal adaptation as an adequate account for the existence of streaks. In the first place, the stimuli in the earlier vision studies were all presented continuously, and the phenomenal experience is a waxing and waning of intensity when fixation is maintained over several seconds. Our stimuli were not presented continuously (for only a few tens of milliseconds in the preattentive studies) and fixation was constantly changing. Second, we also find streaks in the complex tasks of judging relative length ratio and contour roughness. It is not obvious how retinal mechanisms or eye movements could influence the outcomes of such judgments to produce sequential dependencies.

In spite of the fact that retinal adaptation does not appear to be the relevant construct for the origin of streaks, we still argue for a sensory/perceptual locus of fluctuation that is logically and possibly physiologically distinct from mechanisms associated with focused attention. The motivations for this claim are the paired results (1) that the preattentive tasks were the streakiest, and (2) that extended practice had no influence on streak formation. On the other hand, the universality of the streak phenomenon is evidence for a central origin. If virtually all discrimination activity is streaky, then it is arguable that streaks arise in a mechanism that is globally accessed. Attention is a construct that is conceived to be global and our results may simply point to the existence of two attentional systems; one that generates coherent oscillations and is accessed in preattention, and one that generates white noise and is controlled. The identification of two types of attention, one associated with preattention and one controlled and effortful, has been made by Shiffrin (1988) and by Weichselgartner and Sperling (1987) on more general grounds. The issue of whether streaks arise centrally as an aspect of attention, or more peripherally in sensory/perceptual systems will not be decided here. In fact we note with some concern that Guilford remarks that the central/peripheral issue had been settled and unsettled four times in the four generations prior to his review.

SUMMARY AND CONCLUSION

In this article we have presented experimental evidence that iterated trials in signal detection generally results in sequences of outcomes that are streaky when compared to the expectation of a stationary Bernoulli process. Using time delays and a range of stimuli we were able to construct tasks that generated sequences that were highly streaky (flash, brightness, orientation, and missing side discrimination), that were intermediate in streakiness (distance ratio discrimination, fractal discrimination, tone detection, and consistent mapping in letter search) as well as one that was indistinguishable from a stationary Bernoulli process (ovateness discrimination). Our experiments suggest that the magnitude of the runs deficit is a decreasing function of the attentional resources demanded by the task. We found that it was possible to predict the level of streakiness in a given task by evaluation of the attentional demands.

The demonstration that discrimination performance does not derive from a stationary Bernoulli process raised the issue of what makes performance streaky. We considered four processes; learning, wave modulation, intermittent variations in effort, and conditionalization upon prior outcome. All processes were able to yield sequences that had the observed run deficits, but they differed in their ability to generate sequences that had the observed window statistics. Only the wave process was able to simultaneously produce the observed runs deficits and the window statistics for a class of experiments. The wave process simulated the data in the preattentive experiments with such precision that it motivated a theory of streak formation.

The conjecture that there is a wave-like entrainment of hit rate arises from Monte-Carlo simulations of data and so is extremely model dependent. However, the clear distinction between preattention and effortful attention that was found in both simulation and in the empirical studies suggests that this theory is on the right track. Wave-like structures have not been reported as a property of vigilance (see reviews by Davies & Parasuraman, 1982; Parasuraman, 1985), especially on the short timescales (20–100 s) that are relevant here. However, such variations might not be seen except in experiments such as we have conducted and would not have been recognized without the power of analysis that Monte-Carlo simulation affords.

REFERENCES

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review*, **70**, 91–106.

Bergen, J. R., & Julesz, B. (1983). Rapid discrimination of visual patterns. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 857–863.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. New York: Academic Press.

Eriksen, C. W., & Yeh, Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Performance and Perception*, **11**, 583–597.

Falmagne, J.-C. (1985). *Elements in psychological theory*. New York: Oxford University Press.

Fernberger, S. W. (1920). Interdependence of judgments within the series for the method of constant stimuli. *Journal of Experimental Psychology*, **3**, 126–150.

Fisher, D. L. (1982). Limited channel models of automatic detection: Capacity and scanning in visual search. *Psychological Review*, **89**, 662–692.

Fisher, D. L. (1984). Central capacity limits in consistent mapping, visual search tasks: Four channels or more? *Cognitive Psychology*, **16**, 449–484.

Gilden, D. L., Gray, S., & MacDonald, K. (1990). *The hot human*. Presented at the 1990 meeting of the Psychonomic Society.

Gilden, D. L., & Schmuckler, M. (1989). *The perception of fractal structure*. Presented at the 1989 meeting of the Association for Research in Vision and Ophthalmology.

Gilden, D. L., Schmuckler, M. A., & Clayton, K. (1993). The perception of natural contour. *Psychological Review*, **100**, 460–478.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, **17**, 295–314.

Guilford, J. P. (1927). Fluctuations of attention with weak visual stimuli. *American Journal of Psychology*, **38**, 534–583.

Hays, W. L. (1988). *Statistics* (4th ed., p. 824). New York: Holt, Rinehart and Winston, Inc.

Howarth, C. I., & Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology*, **8**, 163–171.

Jonides, J. (1980a). Towards a model of the mind's eye. *Canadian Journal of Psychology*, **34**, 103–112.

Jonides, J. (1983). Further toward a model of the mind's eye's movement. *Bulletin of the Psychonomic Society*, **21**, 247–250.

Julesz, B. (1975). Experiments in the visual perception of texture. *Scientific American*, **232**, 34–43.

Julesz, B. (1981). Figure and ground perception in isodipole textures. In M. Kubovy and J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 27–54). Hillsdale, NJ: Erlbaum.

Keller, J. M., Crownover, R. M., & Chen, R. U. (1987). Characteristics of natural scenes related to the fractal dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**, 621–627.

Kubovy, M., & Gilden, D. L. (1989). Apparent randomness is not always the complement of apparent order. In G. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure* (pp. 115–127). Washington, DC: American Psychological Association.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492–527.

Logan, G. D. (1992). Attention and preattention in theories of automaticity. *American Journal of Psychology*, **105**, 317–339.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception and Psychophysics*, **32**, 397–408.

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. New York: Freeman.

Palmer, J., Ames, C., & Lindsey, D. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human Perception and Performance*, **19**, 108–130.

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. II). New York: Wiley.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, **84**, 1–66.

Shiffrin, R. M. (1988). Attention. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., Vol. 2, pp. 739–811). New York: Wiley.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing. II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127–190.

Staddon, J. E., King, M., & Lockhead, G. R. (1980). On sequential effects in absolute judgment experiments. *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 290–301.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science*, **1**, 46–54.

Townsend, J. T., & Ashby, F. G. (1983). *The Stochastic modeling of elementary psychological processes*. New York: Cambridge Univ. Press.

Treisman, A. M. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 194–214.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, **12**, 97–136.

Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and preattentive processing. *American Journal of Psychology*, **105**, 341–362.

Ullman, S. (1984). Visual routines. *Cognition*, **18**, 97–159.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, **44**, 273–282.

Verplanck, W. S., Cotton, J. W., & Collier, G. H. (1953). Previous training as a determinant of response dependency at the threshold. *Journal of Experimental Psychology*, **53**, 37–47.

Verplanck, W. S., & Blough, D. S. (1958). *Journal of Experimental Psychology*, **59**, 263–272.

Voss, R. F. (1985). Random fractal forgeries. In R. A. Earnshaw (Ed.), *Fundamental algorithms for computer graphics* (NATO ASI Series, Vol. F17) Berlin Heidelberg: Springer-Verlag. Pp. 805–835.

Voss, R. F. (1988). Fractals in nature: From characterization to simulation. In H.-O. Peitgen & D. Saupe (Eds.), *The science of fractal images*. New York: Springer-Verlag.

Weichselgartner, E., & Sperling, G. (1987). Dynamics of automatic and controlled visual attention. *Science*, **238**, 778–780.

Wertheimer, M. (1953). An investigation of the randomness of threshold measurements. *Journal of Experimental Psychology*, **45**, 294–303.

(Accepted April 20, 1994)