Request # 10165859

Email (PDF) To: sherman@cshl.edu

Cold Spring Harbor Laboratory (S-88)
Library ILL
1 Bungtown Road
Cold Spring Harbor, NY 11724-2203

Title: Psychological review.
Title Abbrev: Psychol Rev
Citation: 1975;82(4):276-298
Article: A Theory of Attention
Author: Mackintosh NJ
NLM Unique ID: 0376476 Verify: Unique Key
ISSN: 0033-295X (Print)
Collation: Serial: Printed language material
Publisher: Washington [etc.]
Copyright: Copyright Compliance Guidelines
Authorization: Gail Sherman
Need By: N/A
Maximum Cost: $11.00
Patron Name: C Prody, Freeman
Referral Reason: Not owned (title)
Library Groups: MLCNY,LILRC,MEDL1
Phone: 1.516.367-8352
Fax: 1.516.367-6845
Routing Reason: Routed to NYUMLC as Prefixed
Received: Apr 04, 2003 ( 07:04 AM EST )
Lender: Medical Library Center of New York/ New York/ NY USA (NYUMLC)

This material may be protected by copyright law (TITLE 17, U.S. CODE)

Bill to: NYUHHQ
Cold Spring Harbor Laboratory (S-88)
Library ILL
1 Bungtown Road
Cold Spring Harbor, NY 11724-2203
A Theory of Attention: Variations in the Associability of Stimuli with Reinforcement

N. J. Mackintosh
University of Sussex, Brighton, England

According to theories of selective attention, learning about a stimulus depends on attending to that stimulus; this is represented in two-stage models by saying that subjects switch in analyzers as well as learning stimulus-response associations. This assumption, however, is equally well represented in a formal model by the incorporation of a stimulus-specific learning-rate parameter, $\alpha$, into the equations describing changes in the associative strength of stimuli. Theories of selective attention have also assumed (a) that subjects learn to attend to and ignore relevant and irrelevant stimuli (i.e., that $\alpha$ may increase or decrease depending on the correlation of a stimulus with reinforcement) and (b) that there is an inverse relationship between the probabilities of attending to different stimuli (i.e., that an increase in $\alpha$ to one stimulus is accompanied by a decrease in $\alpha$ to others). The first assumption is justified by the data, the second is not: Overshadowing and blocking are better explained by the choice of an appropriate rule for changing $\alpha$, such that $\alpha$ decreases to stimuli that signal no change from the probability of reinforcement predicted by other stimuli.

In experiments on classical conditioning, the arrangement of a contingency between a stimulus and a reinforcer typically establishes that stimulus as an effective conditioned stimulus (CS), which reliably elicits a conditioned response (CR). In experiments on instrumental learning, the reinforcement of a particular response only in the presence of a specific stimulus typically establishes that stimulus as an effective discriminative stimulus, which reliably controls the occurrence of that response. Whatever differences there may be between these two procedures, it is not unreasonable to expect some similarity in the processes whereby the CS or discriminative stimulus are established as effective signals. The two stimuli may signal different events—the simple delivery of a reinforcer in one case, and the delivery of a reinforcer contingent on the occurrence of a response in the other—but one may hope that similar laws govern the acquisition of these signaling properties. My concern in this article, at any rate, is with the general question of how stimuli are established as signals or acquire associative strength, and no distinction will be made between classical and instrumental experiments.

Virtually all theories of associative learning have taken as their starting point the assumption that increments in the associative strength of a stimulus depend upon some degree of temporal contiguity between occurrences of that stimulus and of the event it signals. By those who wish to express these matters more precisely, it may be assumed that the magnitude of increments $\Delta V$ of a particular $i$ of the current strength, the assumption being a linearly decreasing. This assumption changes in the associative strength $V_A = c_i$ of a, where $\theta$ is a characteristic value of which the observer used $\lambda_i$ and $\lambda = \lim \lambda_i$ for the reinforcing used.

This may be clearly no mor it does not pr the most obvious and instrument in the observa more effective modality, $s$ of stimuli. Similar to which pigeon to gain cont the instrumental way of representing is conditioning not just the learning-rate also affected by the, which ma

\[ \Delta V \]

where $0 < \alpha$ represent the difference to stimulus salience as $d$ associated with...
of increments to the associative strength of a particular stimulus, $A$, is some function of the current value of $A$'s associative strength, the simplest version of this assumption being that such increments are a linearly decreasing function of that value. This assumption is expressed formally by the application of linear operators to changes in associative strength:

$$\Delta V_A = \theta(\lambda - V_A),$$  \hspace{1cm} (1)

where $V_A$ = current associative strength of $A$; $\theta$ is a learning-rate parameter, the value of which is determined by the reinforcer used within the limits $0 < \theta < 1$; and $\lambda$ = limit of $V$, also dependent on the reinforcer used.

This may be a starting point, but it is clearly no more than that. For example, it does not provide any account of one of the most obvious features of both classical and instrumental conditioning, namely, the observation that increments in associative strength depend on the nature of the stimulus. Pavlov (1927) noted that a more intense stimulus was a more effective CS than a less intense one, and that in the case of salivary conditioning with dogs, stimuli from some modalities, such as tactile and auditory stimuli, were more effective Cs than stimuli from other modalities, such as visual or thermal stimuli. Similarly, the wavelength of light to which pigeons are exposed is more likely to gain control over their food-reinforced instrumental responses than is the frequency of a tone sounded from a speaker (Forese & LoLordo, 1973). The simplest way of representing these and similar observations is to assume that the rate of conditioning to a particular stimulus is not just determined by some general learning-rate parameter, such as $\theta$, but is also affected by a stimulus-specific parameter, which may be referred to as $\alpha$. Thus,

$$\Delta V_A = \alpha\theta(\lambda - V_A),$$  \hspace{1cm} (2)

where $0 < \alpha < 1$. It is now possible to represent differences in the rate of conditioning to stimuli differing in intensity or salience as differences in the value of $\alpha$ associated with that stimulus. If conditioning proceeds faster with stimulus $A$ than with stimulus $B$, or if subjects, given conditioning trials with a compound stimulus consisting of $A + B$, respond strongly to $A$ but weakly to $B$, these facts may be represented by assigning a high value to $\alpha_A$, and a low value to $\alpha_B$.

To this point, the elementary ideas being advanced do not represent any significant departure from the central assumptions of traditional theories of learning, at least of theories such as those advanced by Hull (1943), Spence (1936, 1956), or Estes (1950, 1959). At least two recently proposed models in this tradition, for example, have included a stimulus-specific learning-rate parameter such as $\alpha$ (Kendler, 1971; Rescorla & Wagner, 1972).

**THEORIES OF SELECTIVE ATTENTION**

The framework expressed by Equation 2 is, in fact, sufficiently general to include another group of theories, superficially quite distinct. Theories of selective attention, of which several versions have been advanced in recent years (for example, Lovejoy, 1968; Sutherland, 1964; Sutherland & Mackintosh, 1971; Trabasso & Bower, 1968; Zeaman & House, 1963), appear to incorporate some quite radically new assumptions about the processes involved in even simple cases of instrumental learning. Nevertheless, I shall argue that their two central assumptions can easily be represented as variants on Equation 2.

Theories of selective attention have all been formalized as "two-stage" or "chaining" models, where the subject is envisaged as first observing or attending to a set of stimuli on each trial, and then making an overt response determined only by those stimuli attended to on that trial. In this form, they seem to differ sharply from the "single-stage" type of model characterized by Equation 2. Yet although the function of the observing or attentional stage may be described informally as that of initially exposing the subject to a particular set of discriminative stimuli so that those stimuli and no others are available for association with the outcome of the
trial, formally it is only this second point—whether or not a particular stimulus has its associative strength changed—that is of any substance.\footnote{This is of course an oversimplification, for the direction of the subject's attention is also assumed to affect which stimuli will actually determine performance at any given moment, but the question of performance may for the moment be deferred.}

This point has been represented in two different ways. In theories such as those of Zeaman and House (1963) and Lovejoy (1968), the subject is assumed to attend to only one set of stimuli on a trial. The probability of attending to a particular stimulus on a particular trial, therefore, determines the probability that the outcome of that trial will change the associative strength of that stimulus. In the theory of Sutherland and Mackintosh (1971), on the other hand, subjects are assumed to switch in several analyzers on a given trial. The strength of a particular analyzer, therefore, determines the amount by which the associative strength of a particular stimulus is changed. Over a series of trials, of course, these two formalizations amount to very much the same thing. More important, however, insofar as the strength of switching in a particular analyzer or the probability of attending to a particular set of stimuli simply determines which stimuli change their associative strengths (and by how much), this characteristic of two-stage models is equally well represented by a single-stage model that incorporates a learning rate parameter, \( \alpha \), as exemplified in Equation 2. Although theories of selective attention have, indeed, incorporated further assumptions, these assumptions are also easily represented in a single-stage model as determinants of the value of \( \alpha \).

The first of these assumptions, common to all theories of selective attention, is that \( \alpha \) is not a fixed consequence of such physical characteristics of a stimulus as its intensity or modality, but rather that \( \alpha \) may vary with the subject's experience. The experience in question may have occurred before the start of an experiment, but in the more interesting case, it is a consequence of the correlation between the stimuli and the reinforcement to which the subject is exposed during the course of an experiment. In the language normally used by such theorists, the probability of attending to relevant stimuli typically increases, while the probability of attending to irrelevant stimuli typically decreases. In other words, if variations in stimulus A are correlated with changes in reinforcement, \( \alpha_A \) may increase, while if variations in stimulus B are not correlated with changes in reinforcement, \( \alpha_B \) may decrease.

The second major assumption common to all of these theories is that the probability of attending to one set of stimuli is inversely related to the probability of attending to others. This is a necessary consequence of the "one-look" assumption incorporated into the models of Zeaman and House (1963) and Lovejoy (1968). Since subjects are assumed to attend to only one set of stimuli on a trial, it necessarily follows that as the probability of attending to one set increases, so the probability of attending to others must decrease. Moreover, even Sutherland and Mackintosh (1971), although allowing that learning can occur about all available stimuli on a single trial, assume that the strengths of all analyzers sum to 1.0. In general, therefore, theories of selective attention may be said to have assumed an inverse relationship between the values of \( \alpha \) specific to the population of stimuli sampled in any given experimental situation. If that situation can be conceptualized as containing only three sets of stimuli, A, B, and X, an increase in \( \alpha_A \) will be accompanied by corresponding decreases in either or both of \( \alpha_B \) and \( \alpha_X \). This assumption has been called the inverse hypothesis (Thomas, 1970).

Although these two assumptions (that the value of \( \alpha \) rather than being fixed may change with experience, and that the values of \( \alpha \) for different stimuli are inversely related to each other) have both been incorporated into essentially all theories of selective attention, and although both have als majority of ot learning, it is i are logically i fact, no reas or subjects might stimuli or igno carry with it attention is sel by the inver seceive of a the which while pe uli to change schedule of rei those stimuli, c in \( \alpha \) to one s accompanied i change in \( \alpha \) to a theory that tions common without the ot possibility, it suppositions relevant to the ed independent.

Changes in \( \alpha \)

Traces of the operation of attention in both Lashk (1932) early learning. Ac teets entered with a patic which stimuli irrelevant, bu reinforcement hypotheses a though hypo "docile," the from one hyj specified, and were better c of a sudden organization field, rather understood p. In this cont 1950) who quired distin a gradual k analysis in th
both have also been rejected by the majority of other theories of associative learning, it is important to see that they are logically independent. There is, in fact, no reason why the possibility that subjects might learn to attend to relevant stimuli or ignore irrelevant stimuli should carry with it the implication that such attention is *selective* in the sense required by the inverse hypothesis. One can conceive of a theory of associative learning, which while permitting $\alpha$ for various stimuli to change in accordance with the schedule of reinforcement associated with those stimuli, does not insist that a change in $\alpha$ to one set of stimuli is necessarily accompanied by an equal and opposite change in $\alpha$ to other sets of stimuli. Since a theory that accepts one of the assumptions common to most theories of attention without the other is thus clearly a logical possibility, it is important that the presuppositions and experimental evidence relevant to the two assumptions be examined independently.

**Changes in Attention with Experience**

Traces of the assumption that the direction of attention may change can be found in both Lashley’s (1929) and Krechovsky’s (1932) early writings on discrimination learning. According to Krechovsky, subjects entered the experimental situation with a particular set of hypotheses about which stimuli might be relevant and which irrelevant, but in the face of inconsistent reinforcement, would discard unsuccessful hypotheses and try out new ones. Although hypotheses were thus said to be “docile,” the actual rules governing changes from one hypothesis to another were not specified, and the implication was that they were better characterized as consequences of a sudden new insight, or of the reorganization of the subject’s perceptual field, rather than being due to any readily understood process of associative learning. In this context, it was Lawrence (1949, 1950) who first suggested that the “acquired distinctiveness of cues” might be a gradual learning process, amenable to analysis in traditional terms.

It was also Lawrence, of course, who first provided relatively convincing evidence that changes in $\alpha$ might occur. His experiments on simultaneous and successive discriminations (Lawrence, 1949, 1950) and on transfer along a continuum (Lawrence, 1952) showed transfer from one discrimination problem to another, which seemed to imply increases in the value of $\alpha$ for relevant stimuli and/or decreases in the value of $\alpha$ for irrelevant stimuli. This is not the place for a detailed review of these studies, but although other accounts of these transfer effects have been offered, there is reason to question their sufficiency (Sutherland & Macintosh, 1971). Similarly, although experiments on discrimination reversal undoubtedly involve a number of complex processes, it can be argued that one reason why overtraining sometimes facilitates reversal learning, and why exposure to a series of reversals of a single discrimination sometimes leads to eventually very rapid reversal, is that both overtraining and serial reversal training increase the value of $\alpha$ for the relevant discriminative stimuli (Macintosh, 1969; Macintosh & Holgate, 1969). Finally, studies of intra-dimensional and extradimensional shifts have shown that the magnitude of transfer from one discrimination problem to another depends on whether the stimulus dimensions that were relevant and irrelevant in the first problem retain their original significance in the second problem (Shepp & Eimas, 1964; Shepp & Schrier, 1969). It is very difficult to see how such a finding can be explained without assuming changes in $\alpha$ for relevant and/or irrelevant stimuli, which transfer to other stimuli falling along the same dimension.

It may be accepted that many of these findings imply that $\alpha$ changes during the course of discrimination learning; they do not, however, enable one to specify the direction of such changes. As was implicitly acknowledged in the preceding paragraph, many of these results could as well be a consequence of a decrease in $\alpha$ for irrelevant stimuli, as of an increase in $\alpha$ for relevant stimuli. It is somewhat
surprising, therefore, to find that formal theories of selective attention have concentrated on this second interpretation to the virtual exclusion of the first. It is true that Lawrence (1949) admitted that a decline in the distinctiveness of irrelevant stimuli was probably at least as important as an increase in the distinctiveness of relevant stimuli in generating the transfer effects he observed. But in the formal two-stage models of discrimination learning proposed by Zeaman and House (1963), Lovejoy (1968), and Sutherland and Mackintosh (1971), while differential reinforcement is assumed to increase attention to relevant stimuli, there is no comparable independent mechanism for reducing attention to irrelevant stimuli. Except in the sense that a formerly relevant analyzer will revert to its baseline strength if the stimuli it detects subsequently become irrelevant, the only mechanism for decreasing the strength of irrelevant analyzers is by competition with relevant analyzers. It is, in fact, only the inverse hypothesis that ensures that as attention to relevant stimuli increases, so attention to irrelevant stimuli must decrease.

There are good reasons to question this analysis. A substantial body of evidence suggests that exposure to a set of stimuli uncorrelated with reinforcement will specifically reduce the associability of those stimuli with subsequent changes in reinforcement, and that this is not just a consequence of any increase in attention to another set of stimuli. The simplest demonstration of this effect is provided by experiments on latent inhibition (Lubow, 1973; Lubow & Moore, 1959), in which nonreinforced preexposure to a particular stimulus interferes with subsequent conditioning to that stimulus. Experiments by Rescorla (1971) and Reiss and Wagner (1972) have established that such nonreinforced preexposure does not turn a stimulus into a conditioned inhibitor or signal for nonreinforcement and indeed may significantly interfere with the establishment of subsequent inhibitory conditioning. It is difficult, therefore, to see how such results can be explained except by postulating changes in a stimulus-specific learning-rate parameter such as \( \alpha \). Nor is it necessary to present a stimulus in the complete absence of reinforcement, as in typical experiments on latent inhibition, in order to retard subsequent conditioning. Indeed, sufficient exposure to uncorrelated presentations of a stimulus and reinforcer may be an even more effective procedure for reducing the associability of that stimulus with that reinforcer (Mackintosh, 1973).

Latent inhibition has also been reported in experiments on free-operant instrumental discrimination learning by Mellgren and Ost (1969) and Halgren (1974). Although a similar finding has not been reported in discrete-trial studies, there is evidence of what would seem to be an analogous effect. In experiments on simultaneous and successive discrimination learning, Waller (1970) and Hall (Note 1) obtained results that suggest that rats and pigeons may learn specifically to ignore irrelevant stimuli that vary from trial to trial without being correlated with reinforcement.

There is, therefore, a considerable amount of evidence, consistent in a general way with the first assumption of theories of attention, that the associability of a stimulus with reinforcement is not an immutable consequence of that stimulus’s physical characteristics. Exposure to a set of stimuli correlated with changes in reinforcement appears to increase the associability of those and similar stimuli with future changes in reinforcement. Conversely, exposure to variations in a set of stimuli in the absence of correlated changes in reinforcement appears to decrease the associability of those stimuli with future changes in reinforcement. The fact that such changes in \( \alpha \) occur, of course, does not of itself specify the nature of the rules for changing \( \alpha \) that a formal model would have to incorporate in order to provide a satisfactory account of the data. But there is already a strong suggestion that it is not enough to assume that attention to irrelevant stimuli decreases simply by virtue of an increase in attention to relevant stimuli. It seems probable that

a symmetrical change in \( \alpha \) would increase attention to irrelevant stimuli.

The Inverse Law

The second most theories of stimuli comp increase in a way will necessitate attention to stimulus should be clined in th to show that to a particular with experience on the validity. What then the inverse hypothesis that ensures that as attention to relevant stimuli increases, so attention to irrelevant stimuli must decrease.

One line of a necessary derivable from there is limit to the capacity to be understood of the limits of selective attention. Of the many experiments or learning in a dichotic list, human subject a limitation perhaps the former situation exposure to they are not rapidly, and pose, therefore information of their capacity in experimenter
a symmetrical set of rules will be required that will permit opposite but independent changes in $\alpha$ to relevant and irrelevant stimuli.

**The Inverse Hypothesis**

The second main assumption held by most theories of selective attention is that stimuli compete for attention, so that an increase in attention to one set of stimuli will necessarily involve a decrease in attention to some other set of stimuli. It should be clear that the experiments reviewed in the preceding section, designed to show that the probability of attending to a particular set of stimuli may change with experience, do not have any bearing on the validity of this second assumption. What then has been the justification for the inverse hypothesis?

One line of argument has been that it is a necessary corollary of the assumption, derivable from information theory, that there are limits to any organism’s capacity to process information. Stimuli must be assumed to compete for access to a channel of limited capacity, and the formation of associations between one set of stimuli and reinforcement will necessarily be at the expense of the formation of associations with other stimuli. Plausible as it may seem to insist on limitations on the capacity to process information, however, it may be questioned whether these concepts are of any substantial relevance to the sorts of situations to which theories of selective attention have been applied. Of the numerous differences between experiments on instrumental discrimination learning in animals, and those studies of dichotic listening or reaction times in human subjects on which the concept of a limitation on capacity has been based, perhaps the most important is that in the former situation subjects control their own exposure to the discriminative stimuli; they are under no constraints to respond rapidly, and there is little reason to suppose, therefore, that they are processing information at rates close to the limits of their capacity. Although it is true that in experiments on classical conditioning the experimenter controls the subject’s exposure to the CS, the stimuli used are typically salient, often unlocalized, and may sometimes last up to several minutes. It stretches credulity to suppose that the rat’s channel capacity is too limited to permit processing of both a loud tone and a bright light, when the two are switched on simultaneously for the duration of a 3-minute trial in an experiment on conditioned suppression. Whether or not strict simultaneity of stimulus processing is impossible, typical animal experiments have imposed no such requirement.

It is obvious, however, that the major justification for a particular theoretical assumption is the extent to which it permits the derivation of experimental data. What are the data that are typically explained by the inverse hypothesis? They may generally be described as instances of “stimulus selection,” data which suggest that the association of one stimulus with reinforcement does not depend simply on the characteristics of that stimulus, on its correlation with reinforcement, or even, apparently, on its past history, but may be affected by the nature of the other stimuli present at the same time, by their correlation with reinforcement, and by their past history. The most important instances of such data are provided by studies of overshadowing and blocking. The term *overshadowing* refers to the finding, first reported by Pavlov (1927), that the presence of an equally relevant, more salient stimulus may decrease or completely prevent conditioning to a less salient stimulus. A dog would associate a weak thermal stimulus with the delivery of food, provided it was presented on its own, but if it was presented only in conjunction with a more intense auditory stimulus, no conditioning would accrue to the weaker stimulus. In general, conditioning to B often proceeds more slowly if B is always presented in conjunction with A than if B is presented alone.

The term *blocking* was used by Kamin (1969) to refer to the finding that prior training on one element, A, of a compound stimulus, AB, might completely...
prevent conditioning to the second element, B. If rats received a series of trials on which a compound CS, consisting of a light and a noise, signaled the occurrence of a shock, they would show significant conditioned suppression to both stimuli. But if these compound trials were preceded by a series of trials on which, say, the light alone signaled shock, they would show little or no suppression to the noise.

Both overshadowing and blocking, it need hardly be pointed out, follow quite directly from the inverse hypothesis. An intense stimulus is precisely one that is associated with a high value of \( \alpha \), and its addition to the stimulus compound correlated with reinforcement will necessarily, according to the inverse hypothesis, decrease the value of \( \alpha \) associated with any other stimulus. Similarly, prior training with A relevant will increase \( \alpha_A \) and hence ensure that when B is added, \( \alpha_B \) will be low. In spite of this apparently happy agreement between data and theory, however, one can still ask whether blocking and overshadowing are explained more satisfactorily by a theory of selective attention than by any other theory, and whether they are observed on the occasions required, and for the reasons specified, by such a theory. There is reason to believe that they are not.

Rescorla and Wagner (1972) proposed an alternative explanation of these and other instances of stimulus selection. Kamin (1969) suggested that blocking occurs because on compound trials the subject already expects the reinforcer and only unexpected reinforcers are reinforcing. Rescorla and Wagner have formalized one aspect of this idea: The central assumption of their model is that changes in the associative strength of a stimulus are inversely related, not just to the current associative strength of that stimulus, but to the current strength of the entire stimulus complex of which it forms a part. Formally,

\[ \Delta V_A = \theta (\lambda - \bar{V}) \]  
(3)

where \( \bar{V} \) is the sum of the associative strengths of all stimuli (including A) present at the moment of reinforcement. In an experiment on blocking, a sufficient number of reinforced trials with A alone will ensure that \( V_A \) approaches \( \lambda \); at this point, reinforced trials with AB will result in little or no conditioning to B, since \( (\lambda - \bar{V}) \) will be nearly zero.

Rescorla and Wagner's model, although clearly quite distinct from any theory of selective attention, may still, without gross distortion, be said to explain stimulus selection by appeal to an inverse hypothesis. Stimulus are assumed to compete with one another, not for a limited attentional capacity but for a limited amount of associative strength conditionable by a given reinforcer. Other things being equal, the greater the associative strength accruing to one element of a compound CS, the less will be the associative strength of the other. Unlike theories of selective attention, however, Rescorla and Wagner's theory does not postulate what may seem to some implausible limits to an animal's capacity to attend to incoming stimuli; moreover, as we shall see, there are cases where its predictions differ from those derivable from theories of selective attention and cases where it is supported by the evidence. I shall argue, however, that in the final analysis the data on stimulus selection do not support any version of the inverse hypothesis: Blocking and overshadowing are not the consequence of any simple competition between stimuli for some limited resource.

**Experimental Analysis of Stimulus Selection**

**Blocking**

If A is established as a signal for reinforcement, and subjects are then exposed to AB, signaling the same reinforcer, the available data suggest that the failure of conditioning to B is not due to the fact that A preempts attention but is rather a consequence of the fact that B signals no change in reinforcement, predicting nothing that is not already predicted by the presence of A. When this condition is not satisfied, in Kamin's (1969) conditioned lished as a signal AB also signal relatively little if AB signals signaled by A then significant conditioning a case, Wagner reinforced trial by nonreinforcement increase in the to A, so far a amount learned that B is esta conditioned in.

While it may face of this ev of blocking of selective attention that the value conditions of r tainly seems r that any adeq must take as it observation th redundancy of ing occurs be reinforcer as. This is, of course adopted by K and Wagner (way of explai redundancy, o limitations of limitations on conditionable It is possible reduces condi reduces \( e_B \).

I have already theory of asso \( \alpha \) to change include a rule uncorrelatement. If it ceases when the prevailing reinforcement, decreasing att
is not satisfied, blocking may not occur. In Kamin's (1969) experiments on blocking in conditioned suppression, if A is established as a signal for a 1 mA shock, and AB also signals a 1 mA shock, B acquires relatively little associative strength. But if AB signals a stronger shock than that signaled by A alone, or no shock at all, then significant excitatory or inhibitory conditioning accrues to B. In the latter case, Wagner (1971) has shown that if reinforced trials with A alone are followed by nonreinforced trials with AB, then an increase in the number of reinforced trials to A, so far from decreasing $\alpha_B$ and the amount learned about B, actually ensures that B is established as an even stronger conditioned inhibitor.

While it might still be possible in the face of this evidence to save the analysis of blocking suggested by theories of selective attention, for example by assuming that the values of $\alpha$ change when the conditions of reinforcement change, it certainly seems more appropriate to assume that any adequate explanation of blocking must take as its starting point the cardinal observation that blocking depends on the redundancy of the added element: Blocking occurs because AB signals the same reinforcer as that signaled by A alone. This is, of course, precisely the position adopted by Kamin (1969) and Rescorla and Wagner (1972). But there is another way of explaining blocking in terms of redundancy, one which appeals neither to limitations of attentional capacity nor to limitations on the total associative strength conditionable by particular reinforcers. It is possible that the redundancy of B reduces conditioning to B because it reduces $\alpha_B$.

I have already argued that any adequate theory of associative learning must allow $\alpha$ to change and that this change must include a rule for decreasing $\alpha$ to stimuli uncorrelated with changes in reinforcement. If it were assumed that $\alpha_A$ increases whenever A signals a change from the prevailing or expected conditions of reinforcement, the symmetrical rule for decreasing attention to irrelevant stimuli would say that $\alpha_B$ decreases whenever B signals no change from the prevailing or expected rate of reinforcement. The expected rate of reinforcement, however, is precisely that which is predicted by other stimuli in the experimental situation. Thus the blocking experiment, in which AB signals no change in reinforcement from that predicted by A alone, is precisely the sort of situation that would result in a rapid decline in $\alpha_B$.

It is, in fact, possible to show that such a decline in $\alpha_B$ is one of the things that happens when AB signals the same reinforcement previously signaled by A alone. It will be recalled that if animals have learned that A signals a 1 mA shock, the presentation of AB, correlated either with a significant increase in the intensity of shock or with no shock at all, results in significant excitatory or inhibitory conditioning to B. Mackintosh and Turner (1971) found that a small number of trials on which AB signaled the same shock as A alone, interposed between initial conditioning to A and subsequent training with AB, significantly impaired this excitatory or inhibitory conditioning. The implication is that $\alpha_B$ decreased because when B was first introduced it signaled no change in the probability or magnitude of reinforcement.

It is, of course, one thing to say that if AB signals the same reinforcement previously signaled by A alone, $\alpha_B$ will decrease. It is another thing to say that this is the sole or even a major cause of blocking. This assertion has a number of further implications. If the blocking of conditioning to B depends on subjects learning to ignore B, for example, no blocking can be expected until they have done so. Where blocking is substantial, it must be assumed that changes in associative strength are slow relative to changes in $\alpha$. If conditioning proceeded more rapidly, one might expect to see relatively little blocking. But even under normal circumstances, one would expect to see some conditioning to B, since it must take some minimal number of trials (at least one) to reduce $\alpha_B$ to a point
where further changes in the associative strength of B are minimal.

There is, in fact, some evidence consistent with this final implication. Kamin (1969) found it possible to detect some conditioning to B in his standard experimental situation; more important, his data suggested that most if not all of this conditioning was a consequence of the first reinforced AB trial. This observation agrees with some results of an experiment on blocking in instrumental learning by Turner (Note 2). In this experiment, although pretraining on A resulted in some blocking of learning about B during AB trials, blocking was far from complete. Since animals receiving 100 AB trials learned no more about B than those receiving only 10 AB trials, however, the implication is that learning about B was confined to the first few trials after its introduction.

Although these results are consistent with the present argument, they do not bear on the strong prediction derivable from this analysis, that little or no blocking should occur on at least the first reinforced compound trial. A satisfactory test of this prediction requires a situation in which reliable conditioning occurs to the elements of a compound CS after a single reinforced trial. The conditioned suppression of licking in thirsty rats by a CS signaling shock provides one such situation, and we have attempted to see whether under these circumstances pretraining on A does result in significant blocking of conditioning to B when only one reinforced AB trial is given (Mackintosh, Note 3). In several experiments, no evidence of blocking could be found. These results suggest, therefore, that a substantial part of blocking observed in other situations may indeed be a consequence of a rapid decline in attention to the added, redundant element.

Overshadowing

The design of experiments on blocking is such that the added stimulus B predicts no change in the level of reinforcement from that predicted by A alone, and it was, of course, precisely this feature that made it possible to explain blocking in terms of a reduction in \( \alpha \), rather than in terms of the inverse hypothesis. But in experiments on overshadowing, animals may be trained with AB-signaling reinforcement from the outset. If A has not already been established as a signal for reinforcement, how can its presence detract from conditioning to B during reinforced compound trials, unless some appeal is made to the idea that stimuli compete with each other for association with reinforcement? If a subject simply receives reinforced trials with an AB compound, both A and B should be established as signals for reinforcement, and without invoking the inverse hypothesis it is hard to see why the presence of one should cause any decline in attention to the other.

There are, however, at least two circumstances that would appear to provide a sufficiently close parallel to the conditions of an experiment on blocking to make it possible to apply the principle of learned irrelevance to the case of overshadowing. If A is better correlated with reinforcement than B, or if A is a much more salient stimulus than B and therefore conditions more rapidly, then there is a sense in which B will be a redundant signal of reinforcement. In the former case, A, by virtue of its superior schedule of reinforcement, will acquire associative strength more rapidly than B. In the latter case, even though the schedule of reinforcement associated with A and B is identical, since A conditions more rapidly than B, A will be established as a reliable signal for reinforcement after a number of reinforced compound trials, while B will have acquired little associative strength. At this point, just as in the case where the schedule of reinforcement associated with A is better than that associated with B, the situation will be analogous to that obtaining in experiments on blocking: The presence of B predicts nothing that is not already predicted by A alone, and \( \alpha \) will decline.

Numerous studies of compound conditioning have established that of two equally salient stimuli, the more valid stimulus, associated with reinforcement, is the one that is overshadowed. The same result appears to obtain in experiments on blocking. Sch
A Theory of Attention

stimulus, associated with a better schedule of reinforcement, will overshadow the less valid (e.g., Egger & Miller, 1963; Wagner, 1969; Wagner, Logan, Haberlandt, & Price, 1968) and that of two equally valid stimuli, the more salient will overshadow the less salient (e.g., Kamin, 1969; Miles & Jenkins, 1973; Pavlov, 1927). What has not been recognized, however, is that these appear to be the only reliably established instances of overshadowing. Although the more salient element of a compound CS overshadows the less salient, there is no evidence that the less salient will detract from conditioning to the more salient, nor is there much evidence to suggest that if two equally salient stimuli are associated with the same schedule of reinforcement, one will overshadow the other.

If stimuli compete for association with reinforcement in the manner suggested by the inverse hypothesis, one would expect to observe a reciprocal interaction between all elements of a compound CS. If animals receive a series of reinforced trials with AB, then, other things being equal, any conditioning to B must always be at the expense of conditioning to A, and vice versa. The experimental evidence does not support this expectation. Pavlov (1927) reported that overshadowing was a matter of a reduction in conditioning to the weaker of two elements of a compound CS as a consequence of the presence of the stronger. The unspoken implication was that conditioning to the stronger was unaffected. Although Pavlov did not present the data to support this inference, the conclusion is entirely consistent with subsequent experiments.

Kamin (1969) and Mackintosh (1971), in studies of conditioned suppression, reported that a bright light would overshadow auditory stimuli of varying intensities, but neither found any evidence to suggest that the presence of the auditory stimulus detracted from conditioning to the light, nor that the amount of conditioning to the light was an inverse function of the intensity of the auditory stimulus. Schnur (1971), using relatively intense auditory and visual stimuli in another study of conditioned suppression, found no evidence of overshadowing at all. It is true that in some of these experiments ceiling effects may have obscured any evidence of overshadowing. In order to obtain further evidence on this point, we have studied overshadowing in conditioned suppression, using noise and light as CSs, and varying the intensity of the noise between 50 and 85 db re 20 μN/m². To counteract ceiling effects, subjects were tested over a series of trials in extinction, to the point where no group maintained complete suppression. Although there was some evidence of reciprocal overshadowing at intermediate intensities of the noise, in the extreme cases there was no evidence of any such interaction. The most intense noise significantly overshadowed the light, while the least intense noise was itself significantly overshadowed by the light. In neither instance, however, did the overshadowed element, even though itself acquiring significant associative strength, detract from conditioning to the stronger element.

In experiments on instrumental discrimination learning, significant overshadowing has been observed, provided that the overshadowed cue is either extremely difficult to discriminate (Lovejoy & Russell, 1967) or both relatively difficult and in a different location from the more salient, overshadowing cue (Sutherland & Andelman, 1967). When neither of these conditions is satisfied, overshadowing is not observed (Sutherland & Andelman, 1967; Turner, 1968), and it may be difficult to detect any evidence of an interaction between the stimuli (Sutherland & Holgate, 1966; Warren, Derdzinski, Hirayoshi, & Mumma, 1970). The most extensive study of overshadowing in instrumental discrimination learning is that by Miles and Jenkins (1973). Different groups of pigeons were trained on a successive discrimination with either the presence or absence of a tone, a difference in light intensity, or a combination of both cues serving to distinguish posi-
tive and negative trials. The discriminability of the light cue varied across groups. The most striking feature of their results is that the presence of the tone detracted from control by the light only with the least discriminable value of the light, but for this group the presence of the light did not significantly detract from control by the tone. Conversely, the more discriminable lights did overshadow the tone, but in these groups the presence of the tone did not detract from control by the light. No group, in other words, showed simultaneous overshadowing of both light by tone and tone by light.

It is too early to reach any unequivocal conclusion, but the weight of the evidence does not appear to provide much support for the inverse hypothesis. Overshadowing does not appear to be the general, reciprocal affair required by any theory that assumes stimuli necessarily compete for a limited total amount of associative strength or for access to a limited-capacity channel. To a first approximation, stimuli that are overshadowed appear to be only those with lower salience or inferior correlation with reinforcement, which results in a slower rate of conditioning than that accruing to concurrently presented, overshadowing stimuli. Any interaction between elements of a compound that is confined to this sort of situation may be quite satisfactorily explained by a theory that postulates reductions in $\alpha$ to stimuli that predict no unexpected changes in reinforcement.

**Specification of a New Theory of Attention**

If the data traditionally thought to support one of the assumptions adopted by theories of selective attention can be explained in other ways, it may be time to drop that assumption. The theory that emerges from the considerations advanced so far may thus be characterized as follows. Changes in the associative strength of a stimulus do not depend only on its correlation with reinforcement or on the magnitude of that reinforcement; they also depend on the nature of the stimulus.

Stimuli may condition at different rates, and at least one cause of these differences may be represented, as in Equation 2, by differences in a stimulus-specific learning-rate parameter, $\alpha$. The value of $\alpha$ is initially determined by the physical characteristics of the stimulus and the subject's sensory apparatus, but it may also change with experience. In particular, if stimulus A is correlated with changes in reinforcement, $\alpha_A$ will increase, and if stimulus B is not correlated with changes in reinforcement, $\alpha_B$ will decrease. The critical departure from traditional theories of selective attention is that in a situation where several stimuli, A, B, C, are presented, an increase in $\alpha_A$ does not necessarily or directly cause a decline in $\alpha_B$ or $\alpha_C$. Changes in $\alpha_B$ and $\alpha_C$ are independent of changes in $\alpha_A$. They are caused solely by the correlation of B and C with reinforcement and do not depend in any direct way on changes in $\alpha_A$.

**Rules for Changing $\alpha$**

Even if it is assumed that there may be changes in attention to particular stimuli, the assumption that such changes in attention are entirely independent at first sight seems to rule out the possibility of explaining the types of interaction between stimuli that are exemplified by the phenomena of blocking and overshadowing. This is the conclusion accepted by Fisher and Zeaman (1973), who have recently proposed a modification of the original Zeaman and House (1963) model, one of the new features of which is a rejection of the inverse hypothesis. They noted that their revised theory would not predict blocking or overshadowing under normal conditions. As the arguments of the preceding sections perhaps have shown, the validity of this conclusion depends on the nature of the rules proposed for changing attention. Fisher and Zeaman (1973) follow Sutherland and Mackintosh (1971) in assuming that the strength of a particular observing response will increase whenever the outcome of a trial confirms the subject's expectations about the stimulus detected by that observing response; similarly, disconnection results in of that observ as these will n decline in the response to the experiment on ceives a series by AB, the f: preceded by a signaled by A on any chang attending to F observing respi whether a part expected on ti neously present.

A theory such therefore, in change indepe rsonal association with the stimuli, while acquired distin learning, and i dimensiona phenomena of ging and oversh course, that s sense attention argue that th distinctiveness (and should) dif ferent theoretic theory of assoc that explained allowing $\alpha$ t epos by Fish that explained manner propos Rescorla and V.

Although a h hybrid theo that Rescorla blocking and case of the in creases in the component of $\alpha$ to be at the ex As we saw ear is vulnerable that can be a
similarly, disconfirmation of an expectation results in a decrease in the strength of that observing response. Rules such as these will not, of course, predict any decline in the strength of an observing response to the redundant stimulus in an experiment on blocking. If a subject receives a series of reinforcements signaled by AB, the fact that these trials were preceded by a series of reinforced trials signaled by A alone will have no effect on any changes in the probability of attending to B. The rules for changing observing responses make no reference to whether a particular reinforcer is or is not expected on the basis of other simultaneously presented stimuli.

A theory such as Fisher and Zeaman's, therefore, in which observing responses change independently and simply in accordance with the reinforcement schedule associated with each independent set of stimuli, while able to explain data on the acquired distinctiveness of cues, reversal learning, and intradimensional and extradimensional shifts, will not predict such phenomena of stimulus selection as blocking and overshadowing. It is possible, of course, that such phenomena are in no sense attentional in origin. One might argue that the phenomena of acquired distinctiveness and stimulus selection could (and should) be handled by entirely different theoretical constructs. An adequate theory of associative learning would be one that explained acquired distinctiveness by allowing \( \alpha \) to change in the manner proposed by Fisher and Zeaman (1973), and then explained stimulus selection in the manner proposed by Kamin (1969) and Rescorla and Wagner (1972).

Although a case could be made for such a hybrid theory, it must be recognized that Rescorla and Wagner's analysis of blocking and overshadowing is a special case of the inverse hypothesis, since increases in the associative strength of one component of a compound CS are assumed to be at the expense of other components. As we saw earlier, therefore, the analysis is vulnerable to some of the arguments that can be advanced against any such interpretation of stimulus selection. If overshadowing is not the general and reciprocal phenomenon postulated by the inverse hypothesis, and if little or no blocking occurs on the first trial on which a new component is added, this is evidence against Rescorla and Wagner's analysis just as much as against the version of the inverse hypothesis postulated by theories of selective attention. The force of the previous argument was precisely that overshadowing and blocking are consequences of reductions in the value of \( \alpha \) associated with overshadowed and blocked stimuli and are therefore to be explained by the postulation of appropriate rules for changes in \( \alpha \).

It remains, then, to propose such rules. Instead of saying, as do Sutherland and Mackintosh (1971) and Fisher and Zeaman (1973), that subjects learn to attend to and ignore stimuli to the extent that those stimuli successfully predict the outcome of a trial, we want to say that it depends on whether the stimuli are uniquely successful in their predictions. The intuition that we require to formalize is that \( \alpha_A \) should increase if A predicts an otherwise unexpected reinforcer, while \( \alpha_A \) should decrease if A signals no change in reinforcement from the level expected on the basis of other events. There are presumably, a number of ways in which this might be done, but possibly the simplest is as follows. The extent to which a reinforcer is predicted by A is represented by the absolute value of the term \( |\lambda - V_A| \), where, as usual, \( \lambda \) is the asymptotic associative strength conditional by that reinforcer, and \( V_A \) is the current associative strength of A. If we wish \( \alpha_A \) to increase whenever the outcome of a trial is predicted by A better than by all other events on that trial, we could say,

\[
\Delta \alpha_A \text{ is positive if } |\lambda - V_A| < |\lambda - V_X|, \tag{4}
\]

where \( V_X \) is the associative strength of all stimuli other than A present on that trial. Conversely, if we wish \( \alpha_A \) to decrease whenever the outcome of trial is predicted by other events at least as well as by A,
we could say,
\[ \Delta \alpha_A \text{ is negative if } |\lambda - V_A| \geq |\lambda - V_X|. \] (5)

Stimuli can, of course, be established as signals for nonreinforcement just as for
reinforcement. The simple assumption that the value of \( \lambda \) for nonreinforcement
is either zero or some negative number will permit appropriate changes in \( \alpha \) on
nonreinforced trials.2

It is easy to see how such rules will predict the occurrence of overshadowing and
blocking under appropriate circumstances; \( \alpha_B \) will decrease if B is always
presented in conjunction with A and A has previously been established as a signal for
the reinforcer in question, or if, by virtue of its greater salience, A is initially
associated with a higher value of \( \alpha \) and thus acquires associative strength faster than B.
If it is further assumed that the size of the change in \( \alpha \) is proportional to the
discrepancy between \( |\lambda - V_B| \) and \( |\lambda - V_X| \), then \( \alpha_B \) will decrease most
rapidly in the blocking situation, where on the first trial on which B is introduced,
\( V_B \) is near zero and \( V_X \) (which includes \( V_A \)) is near \( \lambda \). Conversely, \( \alpha_A \) will increase
faster if reinforcement is signaled by A alone than by a compound CS that includes
other stimuli as well as A. In principle, therefore, some reciprocal
overshadowing would be predicted, but not if the stimuli were sufficiently salient to be
associated with high values of \( \alpha \) at the outset of conditioning. The data on the
effects of salience on overshadowing presently available are not sufficient to say
whether such an assumption is justified; it might prove necessary to adopt rather
different rules in light of further research.

2 This should not be construed as unconditional acceptance of Rescorla and Wagner’s (1972)
assumption that reinforcement and nonreinforcement both result in changes in a single variable, \( V \).
For present purposes, one could equally well assume that the omission of an expected reinforcer
resulted in an increment in some specific inhibitory process, which grew to some asymptote, \( \lambda_f \), and
that the net associative strength, \( V \), of a stimulus was determined by subtracting this inhibitory
process from a separate excitatory process which was incremented on reinforced trials.

Overshadowing will also be predicted if A is a better predictor of reinforcement
than B, for example if reinforced trials signaled by AB are alternated with non-
reinforced trials signaled by B alone. Under these conditions, the different sched-
ules of reinforcement associated with the two stimuli will result in faster conditioning
to A than to B, and the consequent reduction in \( \alpha_B \) will interfere with further
conditioning to B. Similar arguments would apply to the other examples of relative

There is, however, one serious problem with this analysis. There is evidence that
a less valid stimulus will not only fail to gain associative strength but also may
apparently lose such strength as it had acquired during an earlier history of rein-
forcement. Wagner et al. (1968) showed that a light accompanied by a pair of
tones would acquire significant control over responding, but that when the cor-
relation of the tones with reinforcement was increased, the light lost control, even
though its own correlation with reinforcement was unchanged. Similarly, if B is
more salient than A, it may initially acquire
control, even though it is a less
valid predictor of reinforcement. As train-
ing continues, however, the difference in
validity outweighs the difference in sali-
ence, and B loses control while A gains
control (e.g., Jenkins, 1973; Rescorla,
1972b). An analysis that accounts for
overshadowing by saying that a less valid
stimulus fails to acquire associative
strength because of a decrease in \( \alpha \) is in
no position to explain how it can lose
strength that it has earlier acquired.

The simplest explanation of such results,
consistent with the present theory, would
be to assume that \( \alpha \) affects both learning
and performance. Even if its associative
strength remained high, then, a stimulus
might lose control over responding if a
decline in \( \alpha \) decreased the probability that
this associative strength would be trans-
lated into performance. If B acquires
control over responding, but is then pre-
sented only in conjunction with a second
stimulus (A) better correlated with rein-
forcement, any
may reflect th
brought abou
B’s validity. It
this is a satisfi
later, this gene
compelling reas
performance.

The present t
depends on a c
element; as su
be complete an
of change in \( \alpha \),
determining the
the extent to
signals some ch
If A alone sig
signals nonreinf
increase ratl
must be a be
forcement than
reinforcer than
then \( \alpha_B \) will t
rapidly than it
same reinforc
sumably be su
blocking observ
experiment in
especially wher
his experiment
and this woul
produce subtu
B in one or
ever, reason t
may be produ
any surprisin
presentation o
Kamin (1969)
of a second sh
ound trial re
tioning to the
this double sh
as a stronger
pignanesi (197
when they pre
visual stimulu
ound trial.
would suggest
simply due to
the added ele
strength, in
Rescorla and

forced, any decline in control by B may reflect the eventual decline in \( \alpha_B \) brought about by the relative change in B's validity. It is too early to say whether this is a satisfactory solution; as noted later, this general case provides the only compelling reason for assuming that \( \alpha \) affects performance as well as learning.

The present theory states that blocking depends on a decline in \( \alpha \) to the added element; as such, blocking should never be complete and will depend on the rate of change in \( \alpha \). The most obvious factor determining the degree of blocking will be the extent to which the added stimulus signals some change in reinforcement.

If A alone signals reinforcement and AB signals nonreinforcement, then \( \alpha_B \) will tend to increase rather than decrease, since B must be a better predictor of nonreinforcement than A. If AB signals a stronger reinforcer than that signaled by A alone, then \( \alpha_B \) will tend to decrease, but less rapidly than it would if AB signaled the same reinforcer as A. This should presumably be sufficient to predict the unblocking observed by Kamin (1969) in his experiment in varying shock intensity, especially when it is remembered that in his experiment AB signaled a 4 mA shock, and this would have been sufficient to produce substantial levels of conditioning to B in one or two trials. There is, however, reason to believe that unblocking may be produced by the occurrence of any surprising event shortly after the presentation of the compound stimulus. Kamin (1969) reported that the delivery of a second shock 5 sec after each compound trial resulted in significant conditioning to the added element, even though this double shock did not appear to act as a stronger reinforcer. Gray and Appignanesi (1973) observed a similar effect when they presented a brief auditory and visual stimulus shortly after each compound trial. If confirmed, such results would suggest that unblocking was not simply due to an increase in \( \lambda \) enabling the added element to acquire associative strength, in the manner suggested by Rescorla and Wagner (1972). Unblocking may occur whenever the added element predicts some event of consequence (such as the second shock in Kamin's experiment), whether or not that event is itself able to support conditioning. According to the present analysis, one could argue that the function of the surprising event was to prevent a decline in the value of \( \alpha \) associated with the added element, thus enabling conditioning between that element and the original unconditioned stimulus (UCS) to proceed normally.

The application of the present rules to the case of acquired distinctiveness and dimensional transfer in instrumental discrimination learning is relatively simple. Relevant stimuli in a discrimination problem predict the occurrence and omission of reinforcement more accurately than any other stimuli, and an increase in their \( \alpha \) values will follow from Equation 4. Similarly, irrelevant stimuli, uncorrelated with the delivery of reinforcement, are presented in conjunction with predictive relevant stimuli and should suffer a decline in \( \alpha \) value.

The prediction of the effects of nonreinforced preexposure to a CS and of uncorrelated presentations of a CS and UCS requires that such a CS be regarded as forming a compound with a set of background stimuli, C. The crucial point, then, is that the presentation of AC (CS plus background) predicts no change from the probability of reinforcement or nonreinforcement predicted by C alone, and a decline in \( \alpha_A \) should follow. There is, however, some question whether this is an ideal account. In the first place, it is only the phenomenon of latent inhibition that necessitates the otherwise rather unhappy assumption that \( \alpha_A \) declines even when \( |\lambda - V_A| \) is equal to \( |\lambda - V_X| \). For reasons of theoretical symmetry, if for no others, one would expect this equality to produce no change in \( \alpha_A \). In the second place, there is reason to believe that random presentations of a CS and UCS retard subsequent conditioning more seriously than does simple nonreinforced preexposure to the CS; this additional effect, moreover, may be relatively specific
to conditioning with that particular UCS (Mackintosh, 1973). One interpretation of this observation is that \( \alpha \) is both stimulus- and reinforcer-specific. Changes in the associability of a tone and shock, resulting from exposure to particular correlations between the tone and shock, will produce only generalized changes in the associability of that tone with other reinforcers.

If further research were to confirm this suggestion, the notion of \( \alpha \) as a stimulus-specific learning-rate parameter would need replacing with a parameter \( \alpha_{A,R} \), the value of which determined the magnitude of changes in the associative strength of \( A \) when it was paired with reinforcer \( R \) (or the omission of that reinforcer). If \( \alpha \) thus represented the associability of a particular stimulus with a particular reinforcer, changes in \( \alpha \) would be produced only by exposure to that stimulus and reinforcer (or similar ones), and the phenomenon of latent inhibition would lie outside the scope of the theory—to be explained, perhaps, in terms of some simpler mechanism of habituation.

In the absence of further evidence, further speculation along these lines would be idle. For the present, the rules suggested in Equations 4 and 5 may be regarded as one way of representing the idea that the associability of a stimulus with reinforcement will be affected by the predictiveness of that stimulus relative to that of other, concurrently presented stimuli. These rules are certainly not to be thought of as final, but rather as illustrating that it is possible to express this informal idea reasonably precisely. These particular rules, it will be recognized, bear a more than passing resemblance to Rescorla and Wagner's (1972) rules for changing \( V \): In their system, a change in \( V_A \) depends on the status of both \( V_A \) and \( V_X \); here a change in \( \alpha_A \) depends on the relative status of \( V_A \) and \( V_X \). Informally, they assume that only if a reinforcer is otherwise unpredicted will its presentation in conjunction with \( A \) increase the associative strength of \( A \); here it is assumed that only if a reinforcer is otherwise unpredicted will its presentation increase the probability of attending to \( A \).

**Starting Value of \( \alpha \): Representation of Stimulus Salience**

It is not enough to devise rules for changing \( \alpha \). A comprehensive theory must specify the limits within which \( \alpha \) may change and the conditions determining the starting value of \( \alpha \). The simplest solutions to these problems are to allow \( \alpha \), as a learning-rate parameter, to vary between 0 and 1 and to assume that the starting value of \( \alpha \) for any particular stimulus is positively correlated with the intensity or salience of that stimulus. While these are reasonable assumptions, they are not in fact sufficient to account for the effects of stimulus salience.

Lovejoy (1968), in his formalization of a two-stage theory of selective attention, distinguished between the fixed, base-level distinctiveness of a cue and its changeable or directable distinctiveness. Only the latter was subject to modification by learning. The implication was that changes in attention to various cues due to experience are "bounded, second-order effects which take place within definite limits that depend on the initial strength of the cues" (p. 60). The type of consideration that led Lovejoy to this solution was the observation that however much one might pretrain rats to attend to brightness cues in a jumping stand, they are unlikely to learn a simultaneous brightness discrimination as rapidly as a spatial discrimination. If the difference between spatial and brightness discriminations was represented only by differences in the starting value of a parameter such as \( \alpha \), and if sufficient exposure to a correlation between a set of stimuli and reinforcement resulted in the \( \alpha \) value for those stimuli tending to 1.0, it would follow that differences in the difficulty of visual and spatial discriminations would eventually disappear. It seems clear that differences in the difficulty of instrumental discriminations cannot be adequately represented in this simple way: animals may never perform with complete accuracy if the discrimination is hard enough (e.g., Hara & Warren, 1961), and Mackintosh (1969), for example, noted that differences in the speed of reversing an easy or tion after predicted l

It is equ account is social cond effect of d CS is a d tioning, K very weak of condi Asymptotic course, ca in a learni

Differen S+ and S or between the CS in seem to h performa a model t by assign which the vary betw perience. parameter set of st criminabil other disc more con: knowledge generalization (1954), fo authenticating process CS than consequen between t more inte ference be (the back presence forcement result in to the ba of inhibit CS than differenter: the pres tense CS, the great and S— greater t
m of

rules for
tory must

determine simplest	o allow α, vary be-
that the particular
with the stimulus.
ations, account

ization of attention, may be as-
level unchangeable. Only the
ration by t changes to ex-
er effects to limits th of the
ideration was the one might
ness cues likely to dis-
minimatical and present
ng value sufficient for a set
sulted in ending to
es in the dis-

It seems difficult to
not be simple way; complete
is hard
61), and
reversing

an easy or difficult brightness discrimination after extensive overtraining were not predicted by such a model.

It is equally easy to show that such an account is inadequate for the data of classical conditioning. Although the main effect of differences in the intensity of the CS is a difference in the rate of conditioning, Kamin (1965) has shown that a very weak CS may result in a lower level of conditioning even at asymptote. Asymptote differences in performance, of course, cannot be attributed to variations in a learning-rate parameter.

Differences in the discriminability of S+ and S− in a discrimination problem or between the presence and absence of the CS in classical conditioning, therefore, seem to have more permanent effects on performance than can be encompassed by a model that represents such differences by assigning different values to α, but which then permits α for any stimulus to vary between 0 and 1 as a result of experience. Instead of introducing a further parameter (such as a fixed α′) for each set of stimuli, to represent their discriminability from the background or from other discriminative stimuli, it would seem more consistent with other psychological knowledge to rely on the notion of generalization. Perkins (1953) and Logan (1954), for example, argued that the well-authenticated observation that conditioning proceeds more rapidly with an intense CS than with one less intense may be a consequence of differences in generalization between the CS and the background. The more intense a CS, the greater the difference between that CS and its absence (the background). Reinforcement in the presence of an intense CS and nonreinforcement in its absence, therefore, will result in less generalization of excitation to the background and less generalization of inhibition from the background to the CS than will a comparable amount of differential reinforcement correlated with the presence and absence of a less intense CS. By the same line of argument, the greater the difference between S+ and S− in a discrimination problem, the greater the generalization of excitation to S− that will result from each reinforcement of S+, and the greater the generalization of inhibition to S+ that will result from each nonreinforcement of S− (cf. Kendler, 1971).

These ideas can be formally represented in a relatively simple way by introducing a parameter, $S_{i,j}$, representing the similarity of the $i$th and $j$th stimuli, such that $0 \leq S \leq 1$. With two discriminative stimuli, $A_1$ and $A_2$, reinforcement of $A_1$ will result in an increment in the associative strength of $A_1$ according to the usual equation:

$$\Delta V_{A_1} = \alpha_{A_1} \theta (\lambda - V_{A_1}).$$

(2)

By generalization from $A_1$ to $A_2$, however, this trial will also result in an increment in the associative strength of $A_2$, according to the equation,

$$\Delta V_{A_2} = S_{A_1,A_2} \alpha_{A_1} \theta (\lambda - V_{A_1}).$$

(6)

Similarly, nonreinforcement of $A_2$ will decrease the associative strength of $A_2$, and will also result in a proportional decrease in the associative strength of $A_1$.

A sufficiently high value of $S_{A_1,A_2}$ will ensure that discriminative performance between $A_1$ and $A_2$ will never be perfect, just as a sufficiently high value of $S_{A_1,A}$ (strong generalization between A and its absence) will ensure that even at asymptote the level of conditioning maintained by a weak CS will be below that maintained by a more intense CS. With suitable performance rules, intermediate levels of generalization will not affect performance at asymptote, but the generalization parameter will still provide an additional factor (over and above α) determining rate of learning. It is possible, therefore, to allow that with sufficient training the value of α associated with any stimulus can approach 1, without thereby abolishing differences in rate of learning correlated with differences in intensity or discriminability.

**Specificity of α**

Most formal theories of selective attention have followed Zeaman and House (1963) and Sutherland (1964) in assuming
that subjects learn to attend not just to specific stimuli; such as a vertical line, black door, or red response key, but to stimulus dimensions, such as line orientation, brightness, or hue. Sutherland and Mackintosh (1971), for example, assumed that changes in attention could be represented by changes in the strengths of various analyzers, each of which detected variations along a particular stimulus dimension. The assumption is a rather natural one for theories designed to account for the data of experiments on simultaneous discrimination learning, where subjects are usually confronted on each trial with a pair of stimuli, such as vertical and horizontal lines or black and white doors, which can clearly be described as differing along such dimensions. As Jenkins and Sainsbury (1969) have noted, however, it may not seem quite so naturally suggested by a successive discrimination between the presence and absence of a specific stimulus, and it can be applied only with some difficulty and little success to certain of the data obtained in such a situation.

There are, of course, experimental data that seem to provide good evidence for the dimensionality of attention. In studies of intradimensional and extradimensional shifts, the specific stimuli to which the subject is exposed change from one stage of the experiment to another; the observation that learning of the shift problem is faster if the relevant stimuli of the first problem differed along the same dimension as those of the shift certainly suggests that the changes in attention resulting from original training were changes in attention to entire stimulus dimensions. Similarly, if the phenomenon of transfer along a continuum (Lawrence, 1952) is to be interpreted in attentional terms, the implication is that an increase in the probability of attending to stimuli differing widely along a particular dimension results in an increase in the probability of attending to another set of stimuli differing less widely along the same dimension.

It is questionable, however, whether data such as these require the assumption that analyzers appropriate to entire stimulus dimensions are strengthened and weakened. One could equally well assume that an increase in attention to one stimulus generalized to other stimuli in proportion to their similarity to the training stimulus. Thus an increase in $\alpha_A$, resulting from the correlation of reinforcement with the presence of $A_1$ and nonreinforcement with its absence, might result in graded increases in $\alpha_{A_2}$, $\alpha_{A_3}$, ..., $\alpha_{A_n}$. Similarly, if $A_1$ signaled reinforcement and $A_4$ signaled nonreinforcement, as in an experiment on discrimination learning, generalized changes in $\alpha$ would accrue to the intervening stimuli $A_2$ and $A_3$ from both $A_1$ and $A_4$.

This approach has the advantage of again narrowing the discrepancy between traditional, single-stage theories of associative learning and two-stage theories of selective attention. No appeal is made to hypothetical observing responses or analyzers detecting dimensions of stimulus change. A change in the state of one stimulus, whether of its associative strength, $V$, or of its learning rate parameter, $\alpha$, is said to generalize to other stimuli in accordance with entirely traditional (even if unspecified) assumptions. There is the further advantage that this revision may explain the relatively small differences often observed in comparisons of intradimensional and extradimensional shifts in animal subjects. If changes in $\alpha$ to one pair of stimuli differing along a particular dimension result in only generalized changes in $\alpha$ to another pair, it is entirely reasonable to expect these latter changes to be rather small.

**Performance Rules**

Theories of learning are inferences from observed behavior; some relationship must be postulated between the theorist’s terms and the experimenter’s subject’s responses. It will be sufficient, for present purposes, to suppose that the value of $V_A$ bears some monotonic relationship to the probability that $A$, as a CS, will elicit an appropriate CR, or as a discriminative stimulus, will result in the occurrence of an appropriate attention, necessity between the subjects’ traditional notion of th to affect as a con trial, but makes on a rat wil the white only on t of black active sten and other look mod the subje dimens that trial tendenc and Mac com pla analyz, on one set behav.

There that in tw can be co set of s condition equally conclusio by both Mackin case who an bot may co This is trained na.
A Theory of Attention

293

stimulus weakened, that an stimulus gen- portion to stimulus, ting from with the tent not readed in- similary, if a signaled experiment eneralized the inter- both A1, an appropriate instrumental response. Al- though deliberately vague, this is neither contentious nor problematic. Theories of attention, however, are faced with the necessity of specifying further relations between their theoretical terms and their subjects' responses. As noted earlier, in traditional two-stage theories the direction of the subject's attention is assumed to affect not only what the subject learns as a consequence of the outcome of the trial, but also what response the subject makes on that trial. The probability that a rat will choose the black rather than the white arm of a T maze depends not only on the relative associative strengths of black and white, but also on the relative strengths of attention to brightness and other stimuli. In the original onelook model of Zeaman and House (1963), the subject attends to only one stimulus dimension on each trial, and behavior on that trial is controlled only by the response tendencies to those stimuli. Sutherland and Mackintosh (1971) adopted a more complex solution, assuming that behavior was sometimes controlled by only one analyzer, but that sometimes more than one set of stimuli would determine behavior.

There is, in fact, definitive evidence that in two-choice discriminations behavior can be controlled simultaneously by several sets of stimuli; rats can, after all, solve conditional discriminations. But there is equally good evidence pointing to this conclusion from studies of simple simultaneous discriminations, where solution could in principle be based on only one set of stimuli at a time. When more than one set of stimuli is simultaneously (but redundantly) relevant, it is possible to show that performance may be controlled by both on a single trial (Sutherland & Mackintosh, 1971, p. 142). Even in the case where one set of stimuli is relevant and another irrelevant, there is evidence that both relevant and irrelevant stimuli may combine to control performance. This is suggested by the fact that rats, trained on a simultaneous visual discrimi- nation, having responded consistently to one position over a long series of trials, may make no errors once they break their position habit (Sutherland & Mackintosh, 1971, p. 91). Analysis of the performance of animals shifted from one discrimination to a second, in which the originally relevant stimuli become irrelevant, has shown that few if any errors occur on those trials when the new S+ is combined with the former, but now irrelevant, S+. Such results suggest that both stimuli contribute to the subjects' choice behavior (Medin, 1973; Tighe, 1973).

These conclusions are quite consistent with the general trend of the present argument. It may be assumed, therefore, that all stimuli contribute to performance in accordance with their current associative strengths. The question remains whether one should assume that changes in a not only determine the rate of learning about specific stimuli but also affect the extent to which a stimulus' associative strength is actually translated into performance. One could certainly assume that the contribution of any stimulus to performance was a product of its associative strength and a value. Thus as aA declined, even if V_A had some moderate value, the contribution of A to performance would also decline.

There is actually very little evidence bearing on this question. Since, in gen- eral, a stimulus with a low value of a will also be one with little associative strength, it may be very difficult to decide whether the reason why a particular stimulus fails to affect behavior is because it is not attended to (i.e., has a low a value) or because its associative strength is low. There is evidence from studies of discrimination learning that differences in the associative strength of S+ and S− from an initial discrimination are preserved while the animal learns another problem with the initial stimuli no longer correlated with reinforcement—even though there may be no detectable evidence that such stimuli affect behavior (e.g., latency of responding) on the intervening problem (Mackintosh, 1963; Stettnner, 1965). Similarly, experiments on generalization in
pigeons have established that in the presence of one stimulus (e.g., a colored background on the response key), another set of stimuli, such as the orientation of a line on the key, may exert little or no control over responding, in spite of the fact that a particular line has been established as a signal for reinforcement, and reliably sloping gradients may be observed if different lines are presented on a black background (Newman & Benefield, 1968).

Such instances of masking of control, however, can be explained without recourse to the assumption of competition between stimuli for control. The associative strength of the masking stimuli may be sufficiently great that ceiling effects obscure the control being exercised by the masked stimuli. These results, therefore, cannot really be said to provide conclusive evidence of the influence of $\alpha$ on control. Until such evidence is provided, it would seem reasonable to suggest that $\alpha$ may simply be a learning-rate parameter, with no effect on the control of behavior. Behavior would then be determined by a combination of the associative strengths of all stimuli present and impinging on the subject's receptors.

The only set of results that may require a revision of this assumption is the observation, noted earlier, that a stimulus imperfectly correlated with reinforcement may initially acquire associative strength but will eventually lose control if it is accompanied by other, more valid stimuli. The explanation provided by Rescorla and Wagner (1972) is that such a stimulus loses associative strength due to competition with more valid stimuli for a limited total amount of associative strength. The present theory is better able to explain why an invalid stimulus should not acquire associative strength in the first place than to explain why, having acquired strength, it should then lose it. It may be necessary to assume that although such a stimulus maintains its associative strength, a decline in $\alpha$ will decrease the probability that it will control responding.

### Conclusions

The theoretical ideas proposed here may be summarized briefly. Changes in the associative strength of a stimulus are partly determined by a learning-rate parameter, $\alpha$, specific to that stimulus. This parameter is itself subject to change, increasing when a stimulus predicts a change in reinforcement, decreasing when it does not. This idea is formally equivalent to one of the main tenets of two-stage, attentional theories of learning, namely, the assumption that the probability of attending to a stimulus determines the probability of learning about that stimulus and may itself change with experience. This equivalence may justify characterizing the present set of ideas as a theory of attention, but since that term has a number of connotations, it might be better to stress that what I am proposing is a theory about the associability of stimuli with reinforcement. At the risk of belaboring the obvious, I must reiterate that $\alpha$ is a learning-rate parameter, and possibly a determinant of performance: There is no implication that changes in $\alpha$ correspond to changes in perception.

Theories of selective attention have all incorporated a second assumption—that the probability of attending to one set of stimuli is inversely related to the probability of attending to others. The empirical justification for this assumption has been that it predicts the occurrence of such instances of stimulus selection as overshadowing and blocking. A closer analysis suggests, however, that these phenomena may not be due to any such competition for attention, but may rather be a consequence of subjects learning to ignore stimuli that signal only the occurrence of a reinforcer already predicted by other stimuli. This idea can be expressed by the choice of appropriate rules for changing $\alpha$. Thus the single assumption that $\alpha$ increases and decreases in particular ways is sufficient to explain many of the data that have been thought to require a theory of selective attention.

It is worth concluding with some qualifications. It is obvious enough that the ideas propos for a theory model of col learning. In edge, there is this. It is some factors of the prese...
ideas proposed here are more a program for a theory than a fully elaborated formal model of conditioning and discrimination learning. In our present state of knowledge, there is little need to apologize for this. It is more important to point to some factors that may limit the generality of the present arguments.

First, I have assumed that it is possible to predict the behavior of a subject toward a compound stimulus, AB, solely from knowledge of the status of its elements, A and B. The possibility of "configural" conditioning or "compounding" shows that this is not always true: With sufficient training, animals can learn to respond appropriately when AB is consistently reinforced but A and B separately are consistently not reinforced (e.g., Rescorla, 1972a; Woodbury, 1943). There is good reason to believe, moreover, that configural learning occurs even when it is not explicitly reinforced. Booth and Hammond (1971), for example, showed that continued reinforcement of AB might be sufficient to reduce significantly the level of responding maintained by A and B alone, even without nonreinforcement of the component stimuli in isolation.

Similar configural learning is a fairly common observation in operant discrimination learning by pigeons. If pigeons are rewarded for pecking an illuminated response key in the presence of a 1,000 Hz tone but not in its absence, they will learn not to respond to the key light alone but will equally refuse to peck in the presence of the tone if the key light is turned off (Jenkins & Harrison, 1960). This case is of particular interest in the present context, for it is an example of the main unresolved problem for the analysis proposed here—how a salient but relatively invalid stimulus comes to lose control over responding after it has initially acquired such control. If pigeons stop pecking when the key light is turned off, it is obviously misleading to suggest that the light has lost control over responding. Perhaps similar configural effects occur in other, similar situations. There are, in fact, very few data available to show how far the apparent absence of control by B following reinforcement of AB reflects configural conditioning to the AB compound.

A second factor ignored by the present analysis may also contribute to at least some cases where salient irrelevant stimuli lose control of responding. From the time of their earliest discovery, results showing apparently selective effects in conditioning and discrimination learning have been attributed by some investigators to the occurrence of overt orienting responses. If B fails to acquire control over responding following reinforcement of AB, it is said to be because the subject oriented towards A and thus failed adequately to observe B. One of the few instances where there is actually good reason to suppose that such effects might operate is in the experiments of Jenkins and Sainsbury (1969), in which pigeons were trained to peck at visual displays containing discrete elements. In this situation, A and B would refer, for example, to small red and green dots or to a circle and a star; these elements were displayed on a large, specially constructed, response key, which enabled the experimenters to record which element was pecked at on a given trial. If birds were trained on a successive discrimination between AB+ and B−, Jenkins and Sainsbury found that they learned the discrimination by concentrating their pecks at A on reinforced trials. This suggested the possibility that when pecking at (orienting towards) A on reinforced trials, the pigeon associated only A with reinforcement. Although nominally reinforced on 50% of trials, therefore, B lost control of responding because it failed to gain strength on reinforced trials. Although later finding that pigeons could, to some extent, associate B with the outcome of a trial even though pecking at A, Jenkins (1973) also obtained results that suggest that the (partial) selectivity of orienting responses in this type of situation may indeed play an important role in enabling birds to learn the discrimination between AB+ and B−. If in such a problem reinforcement was forfeited on positive trials whenever the birds pecked at A,
they continued to peck at B on both positive and negative trials and thus failed to learn the discrimination. Because pecks were directed at B, it gained strength on reinforced trials and, thus continued to control responding. It is possible, therefore, that in this situation at least, the loss of control by the relatively invalid stimulus, B, depends on a mechanism of selective orientation that lies outside the scope of the present analysis.

Finally, it is necessary to acknowledge the possibility that there may be situations in which selective effects occur for the reasons postulated by traditional theories of selective attention. Mackintosh (1971), for example, found that significant overshadowing of a weak tone by a strong light might occur on the very first trial of conditioning. Such overshadowing clearly cannot depend on subjects learning the redundancy of the tone nor on the mechanism proposed by Rescorla and Wagner (1972); it suggests a perceptual or attentional interaction. It is possible that similar instances of attentional interactions would be observed more frequently if the learning were sought under more appropriate conditions. I earlier argued that the inverse hypothesis of theories of selective attention did not rest on any very secure rationale: Animals may have a limited capacity for processing information, but it is hard to believe that this limitation prevents the simultaneous analysis of the relatively salient stimuli typically used in studies of conditioning. If there is a limit, however, its effects should be apparent under suitable conditions, such as the brief controlled presentations of complex stimuli.

This argument has also been advanced by Riley and Leith (in press), and its force is suggested by the results of an experiment, undertaken by Turner (Note 2), on delayed matching to sample in pigeons. Given unlimited exposure to the sample key, subjects performed accurately even when they did not know until after the sample had been turned off whether they were going to be required to match the color or the line displayed on the sample. Under these conditions, therefore, they were clearly able to attend to both features of the sample. But as soon as exposure to the sample was reduced to a fixed, very brief period of time, performance deteriorated sharply, unless the conditions were given an additional, conditional cue that signaled whether the ensuing trial was going to require a line match or a color match. The implication is that pigeons were unable to analyze (or remember) both features of the compound sample when it was presented for only a brief interval. Provided the conditional cue signified which feature they needed to attend to, however, they maintained accurate performance. There was evidence, therefore, that stimuli were competing for access to a limited capacity system, and the inverse hypothesis was fully supported.

Although it may seem unpleasantly complex, it is possible that a complete analysis of conditioning and discrimination learning will require the assumption that the degree of competition between stimuli may vary from one extreme, where all available stimuli are analyzed on a single trial, to the other, where something like a one-look model may apply.

REFERENCE NOTES


REFERENCES


Jenkins, H. M. Noticing and responding in a discrimination based on a distinguishing element. Learning and Motivation, 1973, 4, 115–137.


(Received October 7, 1974)