PSYCHOMETRIKA—VOL. 8, NO. 1
MARCH, 1943

A GENERAL THEORY OF LEARNING AND CONDITIONING: PART I

WALTER PITTS
THE UNIVERSITY OF CHICAGO

In the first of two parts in which a general mathematical theory of non-symbolic learning and conditioning is constructed, the sections of the theory dealing with non-symbolic learning and conditioning are presented, and a number of its qualitative implications are compared with available experimental results. In general, the agreement is found to be rather close.

1. Introduction: the current theories.

The field of conditioning and learning has attained a development on the purely experimental side which renders it an excellent point for the entry of quantitative theory into psychology; and the exploitation of various aspects of this domain for theoretical purposes has already made a promising beginning. The work of this kind done so far has been chiefly from two standpoints: the first, exemplified in the researches of Householder, Landahl, and Rashevsky, as well as others of the same group (summarized in 39 and 40), attempts with some success to explain the phenomena directly upon a neurological basis, while the second, comprising studies such as those of Gullikson (15), Gullikson and Wolfle (16), Hull (27, 28, 29), and Spence (48), prefers to elaborate first a macroscopic account of behavior per se, while leaving the neurological foundations until a later stage. These two approaches are of course rather complementary than competitive: the development of theoretical neurology provides very many suggestions for macroscopic work, and the latter simplifies the neurological problem by requiring mechanisms to account for only a few general propositions instead of a multitude of facts in no obvious relation.

So far, the theoretical structures built according to this second viewpoint have shared the common defect of being in a certain sense incompletely quantified. To a considerable degree the stimuli and responses involved, and the successive trials in the learning process, are regarded as unanalysed units; and the laws of variation for the strengths of tendencies to make given responses are stated in terms of the *number* of previous learning trials belonging to specified categories. This procedure, to be sure, has advantages in intuitive and mathematical simplicity which are not to be despised; but it nevertheless conduces to a certain lack of generality. In particular, if we are to count the trials in the learning process, classifying them separately only into a few groups—such as temporal order, or status as success or failure—we must arrange them in a predetermined temporal order and make them as much similar *inter se* as is experimentally feasible, so that in even so slight a modification of the experimental routine as spacing the learning trials irregularly, or varying some of the characteristics of the learning situation among the trials, we can no longer make a theoretical prediction at all. Moreover, there is usually no obvious way of generalizing a system constructed on these lines to such slightly different cases, while for more radically altered situations the difficulty is correspondingly worse. To remedy this, we shall seek to analyze the learning process in a more detailed fashion than before, so that we may state the contribution of each trial to learning in a way depending upon its relevant characteristics and those of the previous trials; and then to essay a statement of our ideas in a form more completely accessible to the methods of mathematical analysis, which, for their most fruitful use, enjoin substantial continuity in the quantities considered. In this way we may use the same experimental generalizations upon which the preceding theories have been based to construct a new one, similar to them in many respects, but rather more extensive in the scope and detail of its considerations. More precisely, we shall consider the results of our discussion applicable to all aspects of learning and conditioning in which the effect of symbolic or verbal factors is not of great significance; and within this field we shall deal with all cases of learning and conditioning in which independent or related stimuli, with given original tendencies to produce specified types of response, are distributed over time in specified intensities in an arbitrary way, continuous or otherwise; and in which affective stimulation, if this form part of the experimental routine, is distributed in any given manner.

In partial confirmation of our hypotheses, we shall point out how most of the principal experimental generalizations can be inferred from the theory, at least as regards comparative order of magnitude; while a rigid quantitative test would require data in a detail not ordinarily given in experimental results. The theory does not seem too difficult to verify in most of its aspects, however, by a fairly extended and precise set of experiments, whose performance would also provide direct information upon a number of matters of considerable import upon which little data are available, and which, even if disconfirming our present system in some of its aspects, would assuredly make suggestions leading to a better one of comparable range and generality.

The presentation of the theory will be divided into two parts. Part I deals only with the case where the stimuli and responses are wholly independent, so that transfer and generalization do not occur, and proposes a law of variation for the reaction-tendency which takes into account all of classical conditioning and the various sorts of inhibition affecting it. Part II extends the discussion, still under the hypothesis of complete independence, to cases where reward and punishment are involved as motivating factors, then generalizes all the preceding results to the case where the stimuli and the responses are related psychophysically, thus providing a theory of transfer, generalization, and discrimination, and concludes with a more precise statement of the observational interpretation of the terms and quantities we shall employ.

2. *The determination of the simple response-tendency.*

As mentioned above, we shall suppose in this part that the stimuli are completely unlike, so that generalization and discrimination do not occur; that the reactions are very much different, so that transfer does not arise; and that the reaction-thresholds do not influence one another, so that there is no "generalization of inhibition." Although in most cases this is not entirely realized, in many types of experiment which will readily occur to the reader, such as classical conditioning for one, and serial learning of properly selected nonsense-syllables for another, it is perhaps sufficiently so for practice. In any case, a prior development of the theory for the independent case, before introducing the various complications involved in related quantities, has an heuristic and intuitive value which will presently become apparent.

We shall begin by introducing the notion of the *simple response tendency.* In the absence of inhibition, a reaction of the given sort will occur whenever this quantity exceeds its reaction-threshold; in more complicated circumstances the occurrence of the reaction depends upon the supraliminal character of some quantity which, regarded as a function of the simple response tendency alone, is linear, though in fact depending on other variables as well. A more complete definition of this term will be given in the last section: for the present, the foregoing may supply the reader with sufficient intuitive understanding to follow the subsequent arguments. If the different reaction-tendencies involved in a situation be assigned numbers from 1 to M, we shall denote the simple response-tendency, for the j-th reaction, evaluated at the time t, by $E_j(t)$. If there be N different types of stimulus involved in the situation, we may assign numbers to them also, say from 1 to N. We shall then denote by $P_i(t)$ $[i = 1, 2, \cdots, N]$ the logarithm of intensity of the type i stimulation being presented at the time t.

In general, the presentation of type i stimulation at a given time will generate an unconditioned simple response-tendency of type j over a certain period of time thereafter, for suitable i and j. If a unit intensity of type i stimulation be applied at a time t, then the resulting *unconditioned* response-tendency of type j, effective at a later time $t + \delta$, will be denoted by $T^0_{ij}(\delta, t)$. We may define this quantity more exactly as follows: let the average value of P_i throughout a small interval Δt about t be $P_i(t)$; let the average value of E_j during a small interval $\Delta' t$ about $t + \delta$ be $\bar{E}_j(t + \delta)$; let P_i vanish outside Δt, and P_k, $k \neq i$, vanish everywhere; and suppose that no conditioning of the stimulus-type i to the reaction j has occurred. Then the limit $\bar{E}_j(t + \delta)/P_i(t)$ as Δt and $\Delta' t$ approach zero about t and $t + \delta$, respectively, is $T^0_{ij}(\delta, t)$. If in this definition we omit the requirement that type i stimulation has not been conditioned to the reaction j, so that we consider the gross magnitude of the response-tendency E_j, we shall obtain a function to be denoted by $T_{ij}(\delta, t)$.

Supposing that stimuli of type i have been applied over the period of time $t = 0$ to $t = t$, we proceed to calculate the total unconditioned tendency to make a type j response at time t which results from this. If the various increments in E_j emanating from different previous intervals combine additively, and the relation between the strength of the stimulus and the magnitude of the elicited reaction-tendency is the familiar logarithmic one, as found, for example, by Kupalov and Gantt (34), this tendency will be, by definition,

$$\int_0^t P_i(\vartheta) T^0_{ij}(t - \vartheta, \vartheta)\, d\vartheta .$$

Combining such quantities for all stimulus-types, and remembering our independence assumption, we shall obtain the total unconditioned part of $E_j(t)$, which is

$$\underline{S_j(t) =} \int_0^t \sum_{i=1}^{n} P_i(\vartheta) T^0_{ij}(t - \vartheta, \vartheta)\, d\vartheta. \tag{1}$$

The total magnitude of $E_j(t)$ may be derived in a similar way from T_{ij}, yielding

$$E_j(t) = \int_0^t \sum_{i=1}^{n} P_i(\vartheta) T_{ij}(t - \vartheta, \vartheta) d\vartheta. \tag{2}$$

The portion of this due to previous conditioning is then

$$Q_j(t) = E_j(t) - S_j(t). \tag{3}$$

To derive an expression for the variation of $T_{ij}(\delta, t)$ with time, we may employ the following method. Suppose we have two time intervals, during the first of which, $\xi \rightarrow \xi + d\xi$, type i stimulation is being applied, and a later interval, $\eta \rightarrow \eta + d\eta$, during which there is both a positive conditioned and an unconditioned tendency to make a type j response. By the fundamental fact of conditioning, if conditioning has not already reached the saturation point, this will result in an increased tendency for future occurrences of type i stimuli to evoke type j responses after a time lag of magnitude $\eta - \xi$—that is to say, an increment of $T_{ij}(\delta, t)$ for $t \geqq \eta$ and $\delta \equiv \eta - \xi$ will be set up. Let us denote this increment by $\Delta_{\xi\eta} T_{ij}(\delta, t) d\xi d\eta$, supposing that to a sufficient approximation this quantity is constant for all points ξ and η in $d\xi$ and $d\eta$ respectively. Since we shall ultimately allow these intervals to approach zero as a limit, this constitutes no real loss in generality.

First, it is a well-established experimental fact that

$$\Delta_{\xi\eta} T_{ij}(\eta - \xi, t)$$

decreases with an increase in the separation $\eta - \xi$ of the stimulus and the response-tendency. This has been observed directly, in the form of rapidly augmenting difficulty of producing trace-conditioning with greater distance, by Kappauf and Schlosberg (32), H. M. Wolfle (54), Rodnick, (43), Switzer (50), and Pavlov (37). In some of these studies, it is true, it is found that there is an initial increase in facility of conditioning with distance between the conditioned and the unconditioned stimuli, but this may be explained on the hypothesis of a diminished contribution to conditioning from temporal generalization, which as an effect is a consequence of the theory. Pavlov (38, 39) supposes that this gradient with trace-conditioning is due to the fact that the reaction-tendency is here conditioned, not to the stimulus itself, but to the arrival of a certain stage in the decay of a condition set up in the subject by the stimulus, and called its *stimulus-trace*; apart from the obscurity of the notion itself, (which has been used extensively by Hull, especially in 27), it is sometimes assumed that this trace has a numerical magnitude, which declines with time at a rate proportional to its own magnitude, or according to a negative exponential law; and then that the increment in a trace-conditioning resulting from a given reinforcement is proportional to the strength at the time of the response of the trace originating in the stimulus—an assumption which seems to be in gross agreement with the data cited. We shall consequently make this supposition in its quantitative form, although without making any ontological assumptions as to the nature of the "stimulus trace" which furnishes its rationale. We shall suppose that $\Delta_{\xi\eta}\, T_{ij}(\eta - \xi , \eta)$ is proportional (1), to $P_i(\xi)e^{-\alpha(\eta-\xi)}$, for a suitable constant α—this is the strength at $d\eta$ of the "trace" from $d\xi$— and (2), to $S_j(\eta) - Q_j(\eta)$, or the amount whereby the unconditioned part of the type j response-tendency exceeds the conditioned part, both evaluated at η. This last assumption is in agreement with the studies of Campbell and Hilgard (7), who found a correlation of .55 ± .06 between strengths of conditioned and unconditioned responses; of Schlosberg (45), and Campbell (6), who found less significant correlations, as to the use of $S_j(\eta)$: for the subtraction of $Q_j(\eta)$ we have considerable evidence which shows that the increment in conditioning which results from a single reinforcement grows less and less as conditioning proceeds, and finally ceases—i.e., the conditioning curve is negatively accelerated. Among the curves of this sort we may mention those of Hovland (25) especially, together with others reproduced in (18). The *S*-shape of some conditioning curves will be discussed in a later study, together with its possible relation to temporal generalization. If we choose units for P_i properly, our assumptions yield

$$\begin{aligned}&\underline{\Delta_{\xi\eta}}\, T_{ij}(\eta - \xi , \eta)\, d\xi d\eta \\ &\qquad = P_i(\xi)\,[S_j(\eta) - Q_j(\eta)]\, e^{-\alpha(\eta-\xi)}\, d\xi d\eta\, .\end{aligned} \tag{4}$$

Now suppose $\delta \neq \eta - \xi$. We may expect that $\Delta_{\xi\eta}\, T_{ij}(\delta , \eta)$ will be smaller than before, but scarcely that it vanishes; the reinforced interval between the conditioned stimulus and the eduction of its response is very far from being so rigidly maintained in experimental trials (See E. J. Rodnick, 43). Perhaps the principal source of this lack of constancy may be sought in more or less accidental physiological fluctuations. If we make the usual assumption, that these fluctuations are large in number and independent, and hence are distributed in an approximation to normality about η, say with a precision β, we shall obtain

$$\Delta_{\xi\eta} T_{ij}(\delta, \eta) d\xi d\eta = P_i(\xi)[S_j(\eta) - Q_j(\eta)] e^{-\alpha\delta - \beta^2(\eta-\xi-\delta)^2} d\xi d\eta, \tag{5}$$

where we of course now evaluate the "stimulus trace" at $\xi + \delta$ instead of at η as before. It may be remarked that on account of the presence of the "stimulus trace" from $\xi \rightarrow \xi + d\xi$ in (5), in the factor $e^{-\alpha\delta}$, the actually observed variations in the interval between a stimulus-presentation and the evocation of its conditioned response will be skewed positively, perhaps even settling about a new mean, much closer to ξ than is η. This is the expression of a fact well-known experimentally (44, 50, and 45), a circumstance which lends additional plausibility to our choice of assumptions.

Finally, this increment in $T_{ij}(\delta, t)$, set up for $t = \eta$, may be expected to decay with time, from the accumulation of adventitious retroactive inhibition (whose effect, incidentally, our theory will enable us to predict, if it be specified in detail, in accord with the hypothesis which explains such inhibition in terms of counter-conditioning; see section 6) or from some other cause, in such a way as to result in its slow but eventually almost complete loss. Here again, there is evidence which leads us to suppose an exponential law of decay (20, and the work of Ebbinghaus), and in doing so, we shall be in harmony with previous theories (Hull, 27, 28). Combining this with (5), we shall finally obtain

$$\Delta_{\xi\eta} T_{ij}(\delta, t) d\xi d\eta = P_i(\xi)[S_j(\eta) - Q_j(\eta)] e^{-\alpha\delta - \beta^2(\eta-\xi-\delta)^2 - \gamma(t-\eta)} d\xi d\eta, \tag{6}$$

for a suitable decay-constant γ.

Now to obtain the total value of $T_{ij}(\delta, t)$ at any time, we must consider (1), the unconditioned part $T^0_{ij}(\delta, t)$, and (2), contributions of magnitude (6) from all pairs of intervals $d\xi$, $d\eta$, $d\xi$ preceding $d\eta$, which occur before t. If these intervals are still considered as finite in extension, this result will be only approximate; if, however, we allow them to approach zero in extension, while increasing indefinitely in number, so that they still cover the whole region $0 \rightarrow t$, then, taking the limit, our result will become exact; and we shall have

$$T_{ij}(\delta, t) = T^0_{ij}(\delta, t) + \int_0^t d\eta \int_0^\eta P_i(\xi)[S_j(\eta) - Q_j(\eta)] \operatorname{Exp}\{-\alpha\delta - \beta^2(\eta - \xi - \delta)^2 - \gamma(t - \eta)\} d\xi. \tag{7}$$

This expression may be substituted in (2); taking account of (1) and

(3), and changing the order of integration by Dirichlet's rule, we shall find

$$Q_j(t) = \int_0^t S_j(\eta) \int_\eta^t d\vartheta \int_0^\eta F(\vartheta, \xi) \operatorname{Exp}\{-\alpha(t-\vartheta) - \beta^2(\eta - \xi - t + \vartheta)^2 - \gamma(\vartheta - \eta)\}\, d\xi - \int_0^t Q_j(\eta) \int_\eta^t d\vartheta \int_0^\eta F(\vartheta, \xi) \operatorname{Exp}\{-\alpha(t-\vartheta) - \beta^2(\eta - \xi - t + \vartheta)^2 - \gamma(\vartheta - \eta)\}\, d\xi\,, \tag{8}$$

where we have put

$$F(\vartheta, \xi) = \sum_{i=0}^{N} P_i(\xi)\, P_i(\vartheta)\,,$$

and, of course, $\operatorname{Exp}(x)$ is a more convenient form of e^x. Upon abbreviating

$$\left\| \begin{matrix} \alpha, \beta, \gamma \\ \eta, t \end{matrix} \right\| = \int_\eta^t d\vartheta \int F(\vartheta, \xi) \operatorname{Exp}[-\alpha(t-\vartheta) - \beta^2(\eta - \xi - t + \vartheta)^2 - \gamma(\vartheta - \eta)]\, d\xi\,, \tag{9}$$

this may be written

$$Q_j(t) = \int_0^t \left\| \begin{matrix} \alpha, \beta, \gamma \\ \eta, t \end{matrix} \right\| S_j(\eta)\, d\eta - \int_0^t \left\| \begin{matrix} \alpha, \beta, \gamma \\ \eta, t \end{matrix} \right\| \times Q_j(\eta)\, d\eta\,. \tag{10}$$

(10) is a Volterra integral equation of the second kind for the determination of $Q_j(t)$, with a kernel

$$K(t, \eta) = -\left\| \begin{matrix} \alpha, \beta, \gamma \\ \eta, t \end{matrix} \right\|\,.$$

Since all other quantities occurring in (8) are known from the experimental situation, if we can solve (8) for $Q_j(t)$, this will constitute a solution to our problem. That this can be done follows from the facts, (1) that $F(\vartheta, \xi) = \sum_i P_i(\xi)\, P_i(\vartheta)$ may have at most a finite number of regularly-distributed discontinuities, owing to the fact that the presentation of some stimulus may have begun or ended abruptly; hence $\left\| \begin{matrix} \alpha, \beta, \gamma \\ \eta, t \end{matrix} \right\|$ is continuous; and (2), $S_j(\eta)$ is also continuous by (1), for any physically possible form of T^0_{ij}. We shall accordingly find it possible to find Q_j, and hence the total reaction tendency $E_j = Q_j + S_j$, in terms of $S_j(t)$, the unconditioned tendency to make a type j response, and the functions P_i, which measure the stimulation occurring up to t in the experimental situation; and both of the functions may be determined by observation, the latter directly, the former through equation (1). The explicit solution of (10) may be made fairly easily in series. Suppose we define $\left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\|^{(n)}$, where n is a positive integer, by the following recursion:

$$\left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\|^{(1)} = - \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\| ,$$

$$\left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\|^{(n+1)} = \int_{\eta}^{t} - \left\|\begin{matrix}\alpha,\beta,\gamma\\ t,\zeta\end{matrix}\right\| \cdot \left\|\begin{matrix}\alpha,\beta,\gamma\\ \zeta,t\end{matrix}\right\|^{(n)} d\zeta .$$

Now define the function $\Psi(n, t)$ called the *solvent kernel* of $-\left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\|$, by the infinite series

$$\Psi(\eta, t) = \sum_{n=1}^{\infty} \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\|^{(n)} , \tag{11}$$

which converges throughout the finite range. Then the solution of (10) may be written

$$\begin{aligned} Q_j(t) = \int_0^t \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\| S_j(\eta)\, d\eta + \int_0^t \Psi(\eta, t)\, d\eta \times \\ \int_0^{\eta} \left\|\begin{matrix}\alpha,\beta,\gamma\\ \zeta,\eta\end{matrix}\right\| S_j(\zeta)\, d\zeta , \end{aligned} \tag{12}$$

as may easily be verified by substituting the series (11) in (12) and the resulting value of Q_j in (10). This expression may be simplified somewhat: remembering that

$$\int_{\eta}^{t} \left\|\begin{matrix}\alpha,\beta,\gamma\\ t,\zeta\end{matrix}\right\| \Psi(\zeta, \eta)\, d\zeta = \Psi(\eta, t) - \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\| ,$$

which is an immediate consequence of (12), we shall have, applying Dirichlets' rule and reliterating,

$$\begin{aligned} Q_j(t) &= 2 \int_t^0 \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\| S_j(\eta)\, d\eta - \int_0^t \Psi(\eta, t) S_j(\eta)\, d\eta \\ &= \int_0^t \{ 2 \left\|\begin{matrix}\alpha,\beta,\gamma\\ \eta,t\end{matrix}\right\| - \Psi(\eta, t) \} S_j(\eta)\, d\eta . \end{aligned} \tag{13}$$

This solution, as mentioned before, is entirely general, subject of course to the assumptions we have made, and the independence requirements of the first three sections.

3. *Extinction and the reaction threshold.*

We shall partially follow Razran (42) and others in the present discussion, in attributing the effects of experimental extinction and allied forms of inhibition to the combined action of three groups of forces. The first of these, which involves setting up a counter-conditioning between the conditioned stimulus and some extraneous and more or less adventitious reaction, is covered by the considerations developed in the previous section, except as regards its interference with the original conditioned reaction; and this will be discussed in section 6. The second sort of influence, an "inverse conditioning" between the stimulus and the same reaction, is also a consequence of our previous discussion, as may be seen at once from (6). Here, under the circumstances of an extinction trial, in which the unconditioned stimulus is not presented, $S_j(\eta)$ vanishes; but on account of previous conditioning and the presentation of the conditioned stimulus, $Q_j(\eta) > 0$. Accordingly, (6) becomes negative, and the strength of conditioning suffers a decrement, proportional in amount to its own magnitude, rather than an increase as before.

The two components of extinction that we have mentioned are essentially like conditioning itself in the rapidity of their acquisition, rate of decay, and other properties. It is notorious, however, that most of the known properties of experimental extinction and allied inhibitions are as different from those of any form of conditioning as it is possible to be: extinction is obtained very much more rapidly than conditioning (Hovland, 25); it decays, to a partial extent, very quickly, a phenomenon known as "spontaneous recovery" (See Hovland, 25; Pavlov, 37, 38; Finch and Culler, 13; D. G. Ellson, 12; C. H. Coombs, 10; E. R. Hilgard and D. G. Marquis, 19, 21; for further references also Hilgard and Marquis, 18; and Razran, 43, the latter being an excellent summary of the facts regarding extinction); the drugs which accelerate it usually retard conditioning, and conversely; (see Hilgard and Marquis, 18, for a discussion of this with references) it can be destroyed temporarily by any sudden stimulus ("disinhibition"), a procedure which has no effect on conditioning at all; and, in general, all the influences which operate to favor extinction usually reduce conditioning; and conversely, agencies which aid conditioning tend to diminish the facility wherewith extinction is obtained. This last is evidenced by the correlations between extinction and conditioning which have been computed from experimental populations (summarized in 18, p. 119), and are predominantly high and negative.

This evidence tends to suggest very strongly that there is a component of experimental extinction which is quite different in its law of variation from ordinary conditioning: in our present discussion we shall take this into account by introducing a *variable reaction-threshold,* itself subject to conditioning, but of the relatively temporary and quick-decaying kind which is characteristic of the inhibitory phenomena. In previous macroscopic theories (e.g. Hull, 27), the reaction-threshold has been taken as constant or at most subject to random variations; but there is no necessity for this, and any apparent one probably results from a confusion between the "reaction-threshold" as we understand it, having reference to gross behavioral phenomena, and the "reaction-threshold" of neurology, which does not stand to the former in any simple relation. The latter is perhaps a physiological constant; but the former may be subjected to variation of any convenient kind. Indeed, a good quantity of evidence appears to suggest the hypothesis that experimental extinction and related inhibitions are due in large part to a temporary rise in the reaction-threshold, understanding this latter as is usual in psychology: as examples of this, of a qualitative sort, we may mention the following.

First, it is well known (25, 44, 18, p. 127) that after a conditioning has been completely extinguished, a sudden fear-evoking stimulus will effect a temporary resurgence of the conditioning; this phenomenon is generally called "disinhibition." If we suppose that a startling stimulus causes a suddenly heightened state of excitation in the organism, which may be interpreted as a temporary lowering of all reaction-thresholds, our hypothesis would lead us to expect just such a loss of inhibition.

Again, it is known that depressant drugs, such as bromides, retard conditioning while expediting extinction, whereas stimulants, such as benzedrine and caffeine, conversely accelerate conditioning and retard extinction. Here also, if we make the natural supposition that depressants raise all the reaction-thresholds, while excitants reduce them, then our hypothesis, which asserts that extinction is a conditioned rise in the threshold, would predict more rapid extinction in the former case, since less further net increase would be required to inhibit the reaction; but slower conditioning, since more reinforcement would be needed to enable the response-tendency to attain the augmented threshold. As remarked, this is just what occurs. A similar account can be given of many cases involving phenomena such as "sensitization," or "pseudoconditioning" (18, chap. I).

A more precise statement can be made readily, by proceeding in relation to the considerations of the previous section. Let us denote the reaction-threshold for type j responses by $R_j(t)$, and its initial, "normal" value by R_{0j}. Now, if we remember that experimental extinction appears to occur wherever there is conditioned reaction-tendency, whether there is unconditioned response-tendency there also or not, and hence is not likely to depend on S_j, and also that it appears to increase and decrease more rapidly than ordinary conditioning, the same reasoning which led to (6) will yield

$$\Delta_{\xi\eta}\, T^{R}{}_{ij}(\delta\,, t) = \nu\, P_i(\xi) Q_j(\eta)\; e^{-\alpha\delta-\beta'^2(\eta-\xi-\delta)^2-\gamma'(t-\eta)}, \tag{13'}$$

where the expression on the left denotes the increment in the strength of conditioning between P_i and R_j, of which $T^{R}{}_{ij}(\delta\,, t)$ is the total value. Precisely as before, we may integrate (13′) to obtain

$$T^{R}{}_{ij}(\delta\,, t) = T^{R0}{}_{ij}(\delta\,, t) + \nu \int_0^t Q_j(\eta)\; d\,\eta \int_0^\eta P_i(\xi) \times$$
$$e^{-\alpha\delta-\beta'^2(\eta-\xi-\delta)^2-\gamma'(t-\eta)}\; d\,\xi\,, \tag{14}$$

and substituting this in the analogue of (2) for R_j and T^R, and abbreviating by (9), we derive

$$R_j(t) = R_0 + \nu \int_0^t \left\| \begin{matrix} \alpha\,, \beta'\,, \gamma' \\ \eta\,, t \end{matrix} \right\| Q_j(\eta) d\,\eta\,, \tag{15}$$

which specifies R_j directly in terms of the conditioning process. Of the constants α, β', γ', γ in (15), α remains the same as in 9, since it represents the strength of the "stimulus trace" from ξ at η, which is the same in the two cases; the relation of β' to β is unknown, although the general liability of R_j would lead us to suppose $\beta' > \beta$; and the very much more rapid decay of experimental extinction as compared with ordinary conditioning implies $\gamma' > \gamma$. We have mentioned before that probably $\nu > 1$. Incidentally, the exponential law for spontaneous recovery, involved in the term $e^{-\gamma(t-\eta)}$ in (13′), is well attested by recent experiments: among these we may mention especially those of D. G. Ellson (12) and Hovland (25).

The hypothesis (15) is sufficient to explain, qualitatively at least, all the principal types of inhibition. First among these we may take the so-called "inhibition of delay," an observed effect in which, after trace-conditioning has taken place between a stimulus i and a reaction j, there is an active inhibition of the response between the presentation of i and the elicitation of j, so that stimuli which even unconditionally evoked j will fail to do so, or the strength of the response will undergo marked coarctation (see e. g. Pavlov, 37, and Rodnick, 44). For, suppose that in the routine of conditioning, i is ordinarily presented at a time, say t, and j appears, due to unconditioned stimulation, at $t + \delta$; but let it be the case, as in experimental trace-conditioning routines, that while $S_j(t + \delta) > R_j(t + \delta) > 0$, $S_j(t + \varepsilon)$, for $0 < \varepsilon < \delta$, is negligible. Nevertheless, a tendency to respond, $E_j(t + \varepsilon) = Q_j(t + \varepsilon)$, which is wholly the result of conditioning, will be built up at that point, owing to the generalization from $t + \delta$, mentioned in the derivation of (3) above. Now when this has occurred, the situation at $t + \varepsilon$ is exactly the same as in experimental extinction; by (15) the reaction-threshold R_j will be conditioned to increase at that point, and the net result will be a strong inhibition. As remarked, this is what occurs. This explanation in terms of our hypothesis, incidentally, has two further consequences of interest. First, the inhibition of delay will be subject to spontaneous recovery; this is remarked upon by Switzer (50), and Rodnick (44). Secondly, after the inhibition of delay has disappeared by spontaneous recovery, the accumulated Q_j at $t + \varepsilon$ will be active reaction-tendency, and if great enough will tend to shorten the reaction-latency or even elicit a second anticipatory reaction before the primarily conditioned one. Effects of this sort have been observed by Rodnick (43), in considerable detail.

A somewhat simpler application of (15) may be seen in the so-called "inhibition of reinforcement." A glance at (15) makes evident the fact that, as a series of reinforcements is traversed, Q_j grows larger and larger, and consequently R_j begins to increase also, the more rapidly the further conditioning proceeds. When Q_j reaches the saturation point, and no longer rises, the increment in R_j continues apace, and the net result is an apparent *reduction* in the strength of conditioning. This is especially marked when the later conditioning reinforcements are massed together in large numbers, so that R_j may increase faster than it can be lost by spontaneous recovery. The diminution from this source would be expected to disappear after a relatively short interval. This phenomenon does in fact occur, and has all the properties we have mentioned: it is observed by Hovland (26), and is discussed by Razran (42), and by Hilgard and Marquis (18, p. 124).

Another sort of phenomenon is that sometimes called "conditioned inhibition." In this effect, a given stimulus i is conditioned to a reaction j, and then paired (successively) with another stimulus k. i and k are then presented together, without the unconditioned stimulus, until the reaction j is extinguished. Then, by our hypothesis, i and k are both conditioned to evoke a rise in R_j. With respect to i, a few reinforcements are sufficient to remove its extinction, so that it again elicits j. When paired with k, however, to which an increment in R_j is still conditioned, we should predict that, under suitable conditions, the response due to i would be partially or completely inhibited. In this case we may call k a *conditioned inhibitor* of the reaction j. This part of conditioned inhibition is of course quite transitory; by more extended efforts, however, a more permanent variety can be established, according to the mechanisms of counter-conditioning and inverse conditioning remarked upon at the beginning of the present section. These expectations are all verified observationally; for an experimental study see (18) p. 126. It is upon these more permanent effects also that we rely to account for the significant part of experimental extinction which is relatively permanent in nature, and for the inculcation of discrimination by the "method of contrasts," which Hovland found (24, 26) to consist of two parts, one analogous to ordinary adaptation in its rapid spontaneous loss and susceptibility to disinhibition, the other being the relatively permanent inhibition found in previous studies (e.g., Hilgard and Humphries, 20), as would be required by a hypothesis such as ours.

The law of variation of the reaction-threshold which we have embodied in (15), however, must not be applied indiscriminately. There are a number of experimental indications which appear to suggest that with respect to many human reactions, which we should ordinarily say are to some extent under "voluntary control," the reaction-threshold, even more than the other variables of conditioning, is determined by factors of this sort, and is consequently not subject to conditioning, at any rate to the extent of being in large measure controlled by such processes. By this, of course, we mean that the threshold depends partially upon the exigencies of other influences involving the subject than are contained within our present theoretical framework—e.g., by the verbal instructions issued to the subject at the beginning of the experiment. Effects of this kind would tend to explain many empirical findings: for example, that experimental extinction fails to occur in conditioning situations with a frequency which appears to vary directly with the extent to which we should say that the function in question is "controlled voluntarily" (18). Again, we may adduce the experiment of J. Miller (35), who considered the effect on conditioning of the eyelid response when subjects were issued instructions at the beginning of the experiment (a) to inhibit actively, and (b) to refrain from control of the response entirely. In the group which was issued the instructions (a), conditioning was slower and extinction quicker than in the control group; whereas the instructions (b) had precisely the converse effect. This is what we should expect on the hypothesis of a raised reaction-threshold in the first case and a diminished one in the other. Further, in cases of serial learning of nonsense syllables, to reproduce each rote-syllable immediately upon presentation of its predecessor in the series, the eduction of reactions in general appears to be syntonic with the beginning of a syllable-presentation, even though the reaction-tendency, as determined from (12), may be greater at another point during the same presentation (27)—i.e., instructions are followed by modifying the reaction-threshold. In general, we may say that the major difference in experimental behavior between the conditioned response and human learning of non-symbolic material resides in the almost total absence in the latter of all the complicated forms of inhibition which are characteristic of conditioning in very many of its aspects—the same procedure which totally extinguishes conditioning will have no effect in a learning experiment—and the assumption of voluntary control of the reaction-threshold in the one case, and accordance with (15) in the other, would appear to reconcile this difference excellently. This difference in observational properties, it may be remarked, has proved a strong objection to the previous behavioristic theories which have attempted to find in learning only a more complex and multiform kind of conditioning. This hypothesis, of course, does not remove the reactions considered, which constitute nearly all of human learning, from the scope of our considerations: in the first place, the threshold cannot be determined wholly at will, and in cases of learning with the subjects' cooperation, where it is perhaps forced down to its minimum possible level, — except for a few regular differences of the kind appearing in the case of serial rote-learning—the course of learning will still be determined as in section 1. The only difference in fact, is that here the subject may refuse to cooperate, and fail voluntarily to learn, whereas in the conditioning of involuntary and autonomic functions this is not possible. Finally, if a criterion for the "voluntary" nature of a reaction is sought, we may perhaps reply in terms of (15): a response-function is involuntary to an extent depending directly upon the presence and regularity of experimental extinction.

REFERENCES

1. Bass, N. J. and C. L. Hull. The irradiation of a tactile conditioned reflex in man. *J. comp. Psychol.*, 1934, 17, 47-65.
2. Brogden, W. C. and E. Culler. Experimental extinction of higher order responses. *Amer. J. Psychol.*, 1935, 47, 663-669.
3. Brogden, W. C., E, A, Lippman, and E. Culler. The role of incentive in conditioning and extinction. *Amer. J. Psychol.*, 1938, 51, 107-117.

4. Brunswik, E. S. Probability as a determiner of rat behavior. *J. exp. Psychol.*, 1939, 75, 175-197.

5. Bugelski, R. Extinction with and without sub-goal reinforcement. *J. exp. Psychol.*, 1940, **26**, 121-133.

6. Campbell, A. A. The interrelations of two measures of conditioning in man. *J. exp. Psychol.*, 1938, **22**, 244-268.

7. Campbell, A. A. and E. R. Hilgard. Individual differences in ease of conditioning. *J. exp. Psychol.*, 1931, **14**, 561-571.

8. Carnap, R. Testability and meaning. *Phil. of Sci.*, 1935 & 1936, **3**, 419-471, and **4**, 1-40.

9. Carnap, R. Logical foundations of the unity of science. *Int. Encyc. of Unified Sci.*, 1939, vol. 1, no. 1, 42-62.

10. Coombs, C. H. Adaptation of the galvanic skin response to auditory stimuli. *J. exp. Psychol.*, 1938, **22**, 244-268.

11. Courant, R. and D. Hilbert. Methoden der Mathematischen Physik. Bd. 1. Berlin: Julius Springer, 1931.

12. Ellson, D. G. Spontaneous recovery of the galvanic skin response as a function of the recovery interval. *J. exp. Psychol.*, 1939, **25**, 586-600.

13. Finch, G. and E. Culler. The relation of forgetting to experimental extinction. *Amer. J. Psychol.*, 1933, **47**, 656-662.

14. Gullikson, H. Studies of transfer of responses: I. Relative versus absolute factors in the discrimination of size by the white rat. *J. genet. Psychol.*, 1932, **40**, 37-51.

15. Gullikson, H. A rational equation of the learning curve based on Thorndike's law of effect. *J. gen. Psychol.*, 1934, **11**, 395-439.

16. Gullikson, H. and D. L. Wolfle. A theory of learning and transfer. *Psychometrika*, 1938, **3**, 127-449 and 225-251.

17. Hilgard, E. R. An algebraic analysis of conditioned discrimination in man. *Psychol. Rev.*, 1938, **45**, 472-496.

18. Hilgard, E. R. and D. G. Marquis. Conditioning and learning. New York: D. Appleton-Century Company, 1940.

19. Hilgard, E. R. and D. G. Marquis. The course of acquisition and retention of conditioned eyelid responses in man. *J. exp. Psychol.*, 1936, **19**, 227-247.

20. Hilgard, E. R. and L. G. Humphries. The retention of conditioned discrimination in man. *J. gen. Psychol.*, 1938, **19**, 111-125.

21. Hilgard, E. R. and D. G. Marquis. Acquisition, extinction, and retention of conditioned eyelid responses to light in dogs. *J. comp. Psychol.*, 1935, **19**, 29-38.

22. Horst, A. P. A new method for determining the absolute affective value of a series of stimulus situations. Thesis, University of Chicago, 1931; published in part in *J. educ. Psychol.*, 23, no. 6, Sept. 1932.

23. Hovland, C. I. The generalization of conditioned responses: I. The sensory generalization of conditioned responses with ranging frequencies of tone. *J. gen. Psychol.*, 1937, **17**, 125-148.

24. Hovland, C. I. The generalization of conditioned responses: II. The gen-

eralization of conditioned responses with varying intensities of tone. *J. genet. Psychol.*, 1937, **30**, 279-291.

25. Hovland, C. I. The generalization of conditioned responses: III. Extinction, spontaneous recovery, and disinhibition of conditioned and of generalized responses. *J. exp. Psychol.*, 1937, **21**, 47-61.
26. Hovland, C. I. The generalization of conditioned responses: IV. The effect of varying amounts of reinforcement upon the degree of generalization of conditioned responses. *J. exp. Psychol.*, 1937, **21**, 261-276.
27. Hull, C. L. et al. Mathematico-deductive theory of rote-learning. New Haven: Yale University Press, 1940.
28. Hull, C. L. The problem of stimulus-equivalence in behavior theory. *Psychol. Rev.*, 1939, **46**, 9-30.
29. Hull, C. L. Mind, mechanism, and adaptive behavior. *Psychol. Rev.*, 1937, **44**, 1-32.
30. Humphries, L. H. The effect of random alternations of reinforcement and extinction on the acquisition and extinction of conditioned eyelid responses. *J. exp. Psychol.*, 1939, **25**, 141-157.
31. Humphrey, G. The nature of learning in its relation to the living system. New York: D. Appleton Century, 1938.
32. Kappauf, W. S. and H. Schlosberg. Conditioned responses in the white rat: III. Conditioning as a function of the length of the period of delay. *J. genet. Psychol.*, 1937, **50**, 27-45.
33. Kellogg, W. N. and I. S. Wolf. The nature of the reaction retained after several varieties of conditioning in the same subjects. *J. exp. Psychol.*, 1939, **24**, 366-383.
34. Kupalov, P. S. and W. H. Gantt. The relation between the strength of the conditioned stimulus and the size of the resulting conditioned reflex. *Brain*, 1927, **54**, 83-98.
35. Miller, J. The effect of inhibitory and facilitory attitudes on eyelid conditioning. Ph.D. Dissertation. Yale University. Abstract in *Psychol. Bull.*, 1939, **36**, 377-378.
36 Münzinger, K. F. and C. C. Dove. Serial learning: I. Gradients of uniformity and variability produced by success and failure. *J. gen. Psychol.*, 1937, **16**, 403-413.
37. Pavlov, I. P. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. London: Oxford University Press, 1927.
38. Pavlov, I. P. Lectures on conditioned reflexes. (trans. and ed. by W. H. Gantt) New York: International Publishers, 1928.
39. Rashevsky, N. Mathematical biophysics, physicomathematical foundations of biology. Chicago: University of Chicago Press, 1938.
40. Rashevsky, N. Advances and applications of mathematical biology. Chicago: University of Chicago Press, 1940.
41. Razran, G. H. S. Transposition of relational responses and generalization of conditioned responses. *Psychol. Rev.*, 1938, **45**, 532-537.
42. Razran, G. H. S. The nature of the extinctive process. *Psychol. Rev.*, 1939, **46**, 264-297.

43. Rodnick, E. H. Characteristics of delayed and trace-conditioned responses. *J. exp. Psychol.*, 1937, 20, 409-425.

44. Rodnick, E. H. Does the interval of delay of conditioned responses possess inhibitory properties? *J. exp. Psychol.*, 1937, 20, 507-527.

45. Schlosberg, H. An investigation of certain factors related to ease of conditioning. *J. gen. Psychol.*, 1932, 7, 328-342.

46. Skinner, B. F. The behavior of organisms: an experimental analysis. New York: D. Appleton-Century, 1938.

47. Spence, K. W. Analysis of the formation of visual discrimination habits in the chimpanzee. *J. comp. Psychol.*, 1937, 22, 77-111.

48. Spence, K. W. The differential response in animals to stimuli ranging within a single dimension. *Psychol. Rev.*, 1937, 44, 430-444.

49. Switzer, S. A. Anticipatory and inhibitory characteristics of delayed conditioned reactions. *J. exp. Psychol.*, 1937, 20, 409-425.

50. Switzer, S. A. Anticipatory and inhibitory characteristics of delayed conditioned reactions. *J. exp. Psychol.*, 1934, 17, 603-620.

51. Thorndike, E. L. The psychology of wants, interests, and attitudes. New York: D. Appleton-Century, 1935.

52. Volterra, V. Theory of functionals. London: Blackie and Son, Limited, 1931. Trans. by Miss M. Long.

53. Williams, S. B. Resistance to extinction as a function of the number of reinforcements. *J. exp. Psychol.*, 1938, 23, 506-522.

54. Wolfle, H. M. Conditioning as a function of the interval between the conditioned and the original stimulus. *J. gen. Psychol.*, 1932, 7, 80-103.

55. Youtz, R. E. P. Reinforcement, extinction, and spontaneous recovery in a non-Pavlovian reaction. *J. exp. Psychol.*, 1938, 22, 305-318.

56. Youtz, R. E. P. The change with time of a Thorndikian response in the rat. *J. exp. Psychol.*, 1938, 23, 128-134.

57. Youtz, R.E.P. The weakening of one Thorndikian response following the extinction of another. *J. exp. Psychol.*, 1939, 24, 294-304.