ACTA NEUROBIOL. EXP. 1980, 40: 291-311

Lecture prepared for the Warsaw Colloquium on Instrumental
Conditioning and Brain Research
May 1979

PROTECTION FROM EXTINCTION BY A CONDITIONED INHIBITOR

Stefan SOŁTYSIK and George WOLFE

Mental Retardation Research Center, School of Medicine
University of California at Los Angeles
Los Angeles, California, USA

Abstract. The phenomenon of protection from extinction (PFE) of a conditioned stimulus (CS) by a conditioned inhibitor (CI) has not been yet unequivocally demonstrated for the CS–CI compound in which the CS precedes the onset of the CI. Preliminary data from a project addressed to this problem strongly indicate that PFE is a real and robust phenomenon. Moreover, the protection is demonstrated not only for the CS duration overlapping with the CI but also for the early part of the CS which is not prevented by the CI from eliciting a conditioned response. The review of a few theories of conditioning suggests that the phenomenon of PFE is theoretically acceptable and predicted within the framework of any hypothetical mechanism which allows for post-trial "processing" or "consolidation" of information acquired during the trial.

INTRODUCTION

Jerzy Konorski was first to propose a causal link between a conditioned inhibitor (CI, a signal of nonreinforcement in a Pavlovian conditioning paradigm) and the protection from extinction (PFE) of a nonreinforced conditioned stimulus (CS). In 1948 he posed the following question: "So why is it that the character of this stimulus (a CS previously paired with the injection of acid into the mouth and presently eliciting a leg flexion avoidance response, S.S.) as a conditioned defensive one is somehow 'preserved' when it is accompanied by a movement constituting a conditioned inhibitor? One has the impression that this very movement somehow protects the stimulus against extinction, and thus the conditioned reflex second type (i.e., instrumental, S. S.) is maintained perpetually. But we are unable to explain the causes of this phenomenon" (10, p. 231). Since then, there were a few attempts to verify the idea that there really is a prevention or retardation of the extinction of the conditioned responses to the nonreinforced CS, if it is presented in compound with the CI.

In dogs trained to press a bar for food in response to a CS, Chorążyna (3, 4) observed that after multiple presentation of a *simultaneous* compound of a CS and CI, the CS retained its capacity to elicit a conditioned response. This finding was not applicable for the avoidance situation where the CS precedes the CI and *elicits*, presumably, at least some suprathreshold fear CR necessary to motivate the avoidance instrumental response.

Sołtysik (17) described an experiment on one dog which had been trained to salivate and press a bar for food in response to a 12 s CS, but not to respond to this same CS if it was preceded by a CI. To simulate more closely the situation of avoidance learning, the author presented 120 times (in 10 daily sessions) a CS–CI compound in which the CS onset preceded for 3 s the onset of a CI and both stimuli coterminated after 12 s from the onset of the CS. During the first 3 s of the CS acting alone the initial stages of the CR were observed: orienting towards the CS, acceleration of heart rate, occasionally the onset of salivation or raising of the paw. After 120 trials of nonreinforced CS–CI compound, the CS alone elicited the full conditioned response. The evidence of PFE was unquestionable, though somewhat limited to the first few trials of CS alone, because the testing trials were reinforced with food and some rapid reinstatement of a CR could be assumed for the consecutive trials. In concordance with the view, prevalent at that time, that there is considerable generality of learning laws, the author confidently assumed that his finding applied also to defensive conditioning, the more so, that food CRs extinguish more easily than defensive CRs.

The next experimental work to address the question of protection against extinction was provided by LoLordo and Rescorla (12). Ten dogs were trained in a Sidman avoidance procedure and in a Pavlovian paradigm, on alternate days. The Pavlovian paradigm included aversive CSs and CIs; CIs, when presented, followed the CSs to simulate the avoidance response. The testing compared the course of extinction of two CSs, one of which was presented nonreinforced but protected by a CI, while the other was simply nonreinforced. No evidence for protection from extinction was found in this study.

Similarly, no evidence for a protective role of a CI was obtained in the study of Johnston, Clayton and Seligman (unpublished 1972, quoted by Seligman and Johnston, 16, p. 83) on rats. Seligman and Johnston felt justified to make a strong statement: "To summarize, our results and those of LoLordo and Rescorla suggest that protection from extinction is nonexistent when the inhibitor follows the conditioned fear stimulus as it must in avoidance paradigms." and later: "There was no reason to have hoped that protection could occur under these circumstances" (16, p. 83).

There are a few reasons, mostly procedural, why these two studies are not very convincing. The first reason for the failure in obtaining protection from extinction is the short training of a conditioned inhibitor in both LoLordo and Rescorla's and Johnston et al. papers. This brief training period probably did not allow the stimulus intended as a CI to become an inhibitor. No independent evidence of the inhibitory properties of the "CI" was presented. The second reason is the relative salience of CSs and the CI. It is generally believed that the CI has to be a strong or salient stimulus. It is in a double disadvantage in matching the CSs: firstly, it elicits an opposing (inhibitory in respect to the CR) process or response, which is known to be a weaker, more slowly trained, and easily disturbed one; and secondly, it is trained after the CR to the CSs is acquired, so it is a later addition to subject's behavioral repertoire. The strength of acquired responses is believed to reflect, besides other factors, their order of acquisition. In both studies which yielded evidence against the protection from extinction, the CSs were auditory stimuli while the conditioned inhibitor was a visual cue: turning off the light in LoLordo and Rescorl's study, and a light stimulus in Jonston's et al. In the latter experiment the "CI" was introduced only after the CS–US pairing was completed, so the "CI" was more like a cue for extinction than a real CI, which normally is acquired by long training including both CS–US and CS–CI trials mixed in each daily session.

The recent study of Hendersen and Harris (7) provides weak support for PFE, limited to the first trial of testing. This result is furthermore undermined by the use of a compound of fully overlapping CS and CI, so that no resemblance to the signalled avoidance paradigm was attempted. The authors were able, however, to provide an independent estimate of the inhibitory role of the CI. It was rather mediocre and the weak protective effect might simply reflect the weak inhibitory potential of their CI.

In summary, the interesting idea of Konorski that the conditioned inhibitor may *somehow* interfere with the extinction process, has produced very little experimental effort so far. Supporting evidence of Chorążyna (3, 4) and Sołtysik (17) are in the realm of food conditioning and need not apply to aversive situations. Hendersen and Harris' (7) data provided some evidence of PFE in aversive conditioning, but the effect was very weak and short lived, probably because the CI was not sufficiently established in its role as an inhibitor. The negative data of LoLordo and Rescorla (12) and of Johnston et al. (cited in Seligman and Johnston 16) should not be accepted because of the lack of evidence that an actual CI was used in their studies.

This dearth of experimental data on the PFE gives us an excuse to report the preliminary results from the first subject in an ongoing study on cats.

THE GENERAL STRATEGY AND THE EXPERIMENTAL DESIGN

Forewarned by the difficulties encountered in the aforementioned studies, the experiment, started a few months ago, was designed in such a way that only well trained and behaviorally clearly defined conditioned stimuli and conditioned inhibitor were used. Moreover, in contrast to all previous studies, all stages of our experiment were carried out without change in the density of shock-reinforced trials. Thus any sudden transitions from an overall situation of shock to no-shock were avoided. As was postulated by Capaldi (2) discrimination between US and noUS situations, and by extrapolation, between different densities of US trials, may contribute to the extinction phenomenon. In this study we wanted the extinction to be cued to particular stimuli and not to the entire situation. Besides, we expected the extinction, on such an "excitatory background", to procede more slowly enabling us to make more precise comparisons between conditioned responses elicited by differently treated CSs.

All stages of the experiment were performed with the animal secured in a modified Pavlovian stand. Although its head was fixed through a cranial implant (to enable physiological recording of heart rate, respiratory movements, etc.) it could walk or run freely on a treadbelt and the distances of locomotor responses were recorded. The unconditioned stimulus (electric shock) was given through two electrodes: one on the left foot (hindleg) and the other on the tail, about 4 cm from its base. The US reliably elicited high leg flexion, a vocal response, changes in respiration and heart rate, and occasionally a locomotor response. The movements of both hindlegs were recorded using light attachements connected to potentiometers, so that raising (flexing) the leg produced a change in electric resistence allowing the onset and the amplitude of flexion to be recorded.

The subject was first trained using Pavlovian leg flexion procedure to respond to three CSs: L_{CS} = continuous light from the panel placed 20 cm in front of the cat's head; A_{CS} = a continuous air blow to the sacral region about 5 cm forward from the tail base; and R_{CS} = a rotating cylinder painted in black and white stripes, placed above the light panel in front of the subject. The duration of these CSs was 5,200 ms. The shock US (2.5 mA, 60 Hz) was delivered 5,000 ms after the onset of a CS and lasted 300 ms, i.e., it terminated 100 ms after the termination of the CS, overlapping with it for 200 ms.

Following the establishment of the CRs to these CSs a conditioned inhibitor, a clicking sound (10 clicks/s) delivered from a laudspeaker situated behind the animal was added. In CS–CI compound trials, the CI was presented 2,000 ms after the onset of the CS and both stimuli were terminated simultaneously 3.2 s later; thus the duration of the CS was the same as in the CS–US trials, and the CI was always presented for 3.2 s.

There were three stages of the experiment. In the first stage (CONTROL) each session consisted of 18 trials: each of the CSs were paired with a shock US twice and the remaining 12 trials consisted of six L_{CS}–CI and six A_{CS}–CI presentations. The trials were randomly mixed, with one constraint, that the same trials did not occur three times in a row. The second stage (PROTECTION/EXTINCTION) started when the subject performed nearly 100% correctly, exhibiting the vocal and leg flexion CRs on the positive trials but not on inhibitory trials. During the Stage 2 the sessions consisted also of 18 trials, six of which were R_{CS}–US trials and the remaining 12 trials were nonreinforcement trials. Six times per session the A_{CS} was presented in compound with the clicker CI: this was the "protected from extinction" CS. And, six times the L_{CS} was presented alone. The selection of the protected and unprotected stimuli was done by a coin flip. The CRs to L_{CS} and A_{CS} were of comparable intensity, with slightly stronger vocal response and shorter latencies of the CRs to the L_{CS}. But this would work against the protection hypothesis. Also the location of the stimuli (L_{CS} in the proximity of always reinforced R_{CS}, and A_{CS} close to the CI) was such that could only work against the evidence for PFE. Still, one more adjustment was made to make the comparison even fairer (or more demanding for the protection hypothesis): the unprotected L_{CS} was presented during stage 2 of the experiment for only 2 s. This was done to control for the possibility that only the first 2 s of the protected A_{CS} really underwent extinction, because the remaining three seconds were "masked" by the CI. Had the unprotected L_{CS} not have been shortened, a differential rate of extinction might have resulted from such uneven treatment. Of course, there was a risk that the subject could learn to discriminate between short nonreinforced and long reinforced CS (using the early termination of a CS as a CI) but luckily this was not so in this first experiment[1]. The PROTECTION/EXTINCTION stage lasted 25 days and each of the CSs was presented 150 times: R_{CS} always reinforced, A_{CS} always with the CI, and L_{CS} always alone for 2 s. The third stage (TEST) of the experiment consisted of 5 sessions which were made up of 6 R_{CS}–US trials, six L_{CS} alone and six A_{CS} alone trials. Both L_{CS} and A_{CS} were presented for 5 s and nonreinforced, so this was extinction of the full duration CSs. The Table I summarizes the duration and trial composition of each stage for this experiment.

TABLE I

General outline of the experiment

Stage 1:10 sessions Control		Stage 2:25 sessions Protection/Extinction		Stage 3:5 sessions Test (extinction)	
Type of trial	No of trials in a session	Type of trial	No of trials in a session	Type of trial	No of trials in a session
R – US	2	R – US	6	R – US	6
L – US	2	L (2 s)	6	L (5s)	6
A – US	2	A – CI	6	A (5s)	6
L – CI	6				
A – CI	6				

Leg flexion data. Table II shows the conditioned leg flexion responses in each type of trial during all stages of the experiment. Percentages of leg flexion occurrence within 5 s after the onset of the CS are computed for 5 session blocks. Note the following facts. There is 100% correct responding to all three CSs during the last 5 control sessions. Responding to L_{CS}–CI and A_{CS}–CI is nearly absent during this period. There is no responding to A_{CS}–CI compound during the stage 2 of the experiment, but the L_{CS} (alone for 2 s) elicits 60% responses in the first 5 day block of the stage 2. In the stage 3, L_{CS} presented for

[1] It should be mentioned, however, that when the subject was later retrained and again exposed to nonreinforced CS of short duration, he was able to learn such a discrimination, and also a PFE was found with the 2 s CS alone, without additional CI.

Table II

Percent of conditioned leg flexions in 5 session blocks

Types of trials	Control		Protection/extinction					TEST
	Blocks of five sessions							
	1	2	3	4	5	6	7	8
R – US	100	100	97	100	100	97	97	100
L – US	100	100	–	–	–	–	–	–
A – US	90	100	–	–	–	–	–	–
L – CI	13	0	–	–	–	–	–	–
A – CI	17	3	0	0	0	0	3	–
L (2)	–	–	60	7	3	3	0	–
L (5)	–	–	–	–	–	–	–	0
A (5)	–	–	–	–	–	–	–	43

5 s does not elicit any responses while A_{CS} produces a considerable $43^{0}/_{0}$ CRs over the 5 days of testing. Since the table disregards the amplitude of responses and the distribution of responding over the 5 days of the TEST stage, Fig. 1 presents the record of the left hindleg flexions from the first 4 days of testing. The first responses to the A_{CS} are full size flexions. Even on the fourth day of testing (i.e., extinction) there is a noticeable tendency for the A_{CS} to elicit a CR.

Impressive as this result is, it has one potential weakness. It refers only to the response which comes late during the CS–US interval and, being of "consummatory" nature, need not characterize the emotional and motivating conditioned processes which are of greater interest for the theory of PFE as conceived for explaning avoidance behavior. Therefore the remaining data will be presented in such a way as to facilitate the discussion of the PFE of the initial part of the CS and/or CR.

Heart rate data. There is no point in rehearsing now the arguments for and against the notion that heart rate changes during the CS–US period reflect some central processes related to attentional and motivational machinery of the brain. But it seems worthwhile to present the data which show that in the cat, the heart rate change regularly accompanies the aversive CS and that this change is suppressed by the CI, protected from extinction by the CI, and, extinguished when the CS is presented without reinforcement. Figure 2 presents computer plots of averaged heart rate curves for the three sets of trials. The two top plots are responses to L_{CS}–US and A_{CS}–US during the first stage (CONTROL) of the experiment. Note the triphasic shape of the heart rate (HR henceforth) response: initial acceleration followed by a bradycardic wave that again reverts into late tachycardia. The US elicits a large tachycardic response which, in this well trained animal, looks like a natural continuation of the late "conditioned" acceleratory response. We postpone the discussion on what these three waves[2] may represent to the next paper (Sołtysik and Wolfe, in preparation) and draw the reader's attention to the middle plots. These are averaged L_{CS}–CI and A_{CS}–CI trials from the control sessions. The dotted line

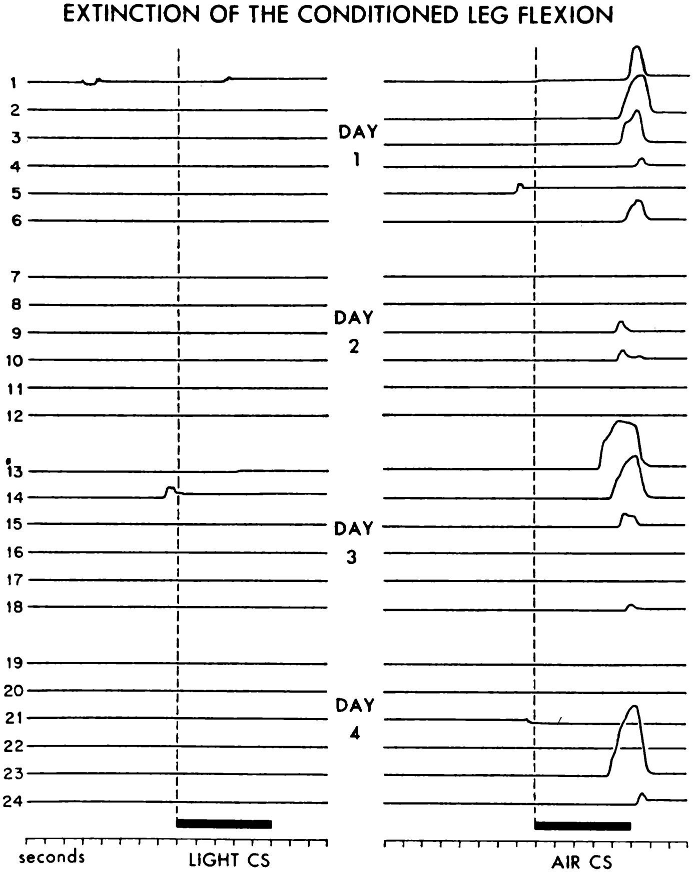

Fig. 1. Extinction of the conditioned leg flexion during first four test sessions. Each horizontal line is the new flexion record from an individual trial. Abscissae, time in seconds from the beginning of the trial to 3 s after CS termination.

[2] We refer to these deflections from the baseline as *waves* because we consider them as analogus to the evoked potential. If our recording was at the level of the heart's pacemaker membrane, the record (after substracting the steady shift of the pacemaker potential) would be very similar to our averaged HR plot, with the depolarization corresponding to tachycardia and hyperpolarization reflecting and, in fact, causing, bradycardia.

repeats the top plot from the CS–US trials for comparison. The initial acceleratory response is preserved on CS–CI trials, simply because for the first 2 s the CS is presented alone. But the addition of the CI erases the bradycardic wave and the late tachycardic part of the HR response. The bottom plots represent averaged HR responses to L_{CS} and A_{CS} alone during the first two days of testing in the third stage of the experiment. The response to L_{CS} is extinguished whereas the A_{CS} which was protected in the stage 2, still elicits the entire triphasic HR response. Of particular interest is the fact that not only the middle bradycardic and late tachycardic parts of the response are preserved, but that also the initial tachycardic response survived well 150 nonreinforcet (but assumedly "protected") trials of the stage 2 of the experiment. While the reason for the non-extinction of the middle and the late waves could be ascribed to the fact that they were not elicited on the

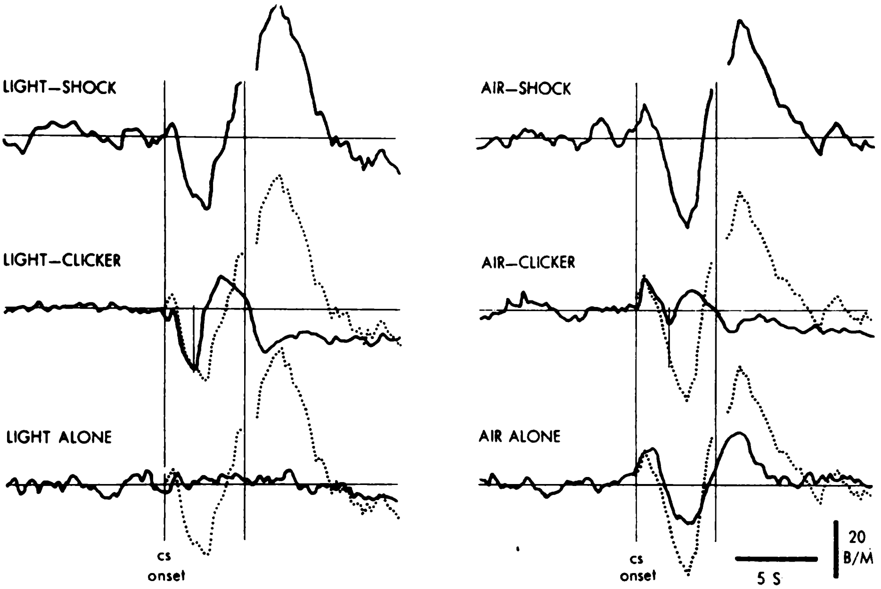

Fig. 2. Comparison of heart rate CRs during various light (left) and air (right) CSs. Each plot is the average instantaneous heart rate (four samples/s) beginning 10 s before CS. onset and ending 10 s after CS termination for a number of trials of the type indicated. All plots have the same vertical and horizontal scale. The vertical lines indicate the CS onset and termination; the short vertical lines indicate CI onset. Top plots are from the stage 1 of the experiment (CS–US trials). Middle plots are CS–CI trials from the same stage. Bottom plots are from the first 2 days of the Test stage. The L_{CS}–US, and A_{CS}–US plots (above) have been replotted on the middle and bottom plots to facilitate comparison (dotted lines). L_{CS}–US, $n = 10$; L_{CS}–CI, $n = 27$; L_{CS} alone, $n = 12$. A_{CS}–US, $n = 9$; A_{CS}–CI, $n = 30$; A_{CS} alone, $n = 12$.

CS–CI trials, the first acceleratory wave was preserved in spite of the fact that it *was elicited* and not reinforced by the US for 150 times during 25 days.

Respiration data. Equally interesting are our finding on the changes in respiration. Briefly, the recording was made from a thermistor placed in front of the external nares. Changes in both frequency and amplitude were recorded. A more extensive discussion on the method and analysis of data will be presented elsewhere (Sołtysik and Wolfe; Wolfe, in preparation). Frequency changes yielded interesting data but they seemed to be related more to attentional processes and will not be discussed now. On the other hand, the changes of the amplitude were found to be of great interest. The typical response to a CS signalling shock is a gradual reduction of the amplitude which drops to zero at the delivery of the shock. After we checked that this reduction is not due to redirecting of the breathing from the nose to the mouth,

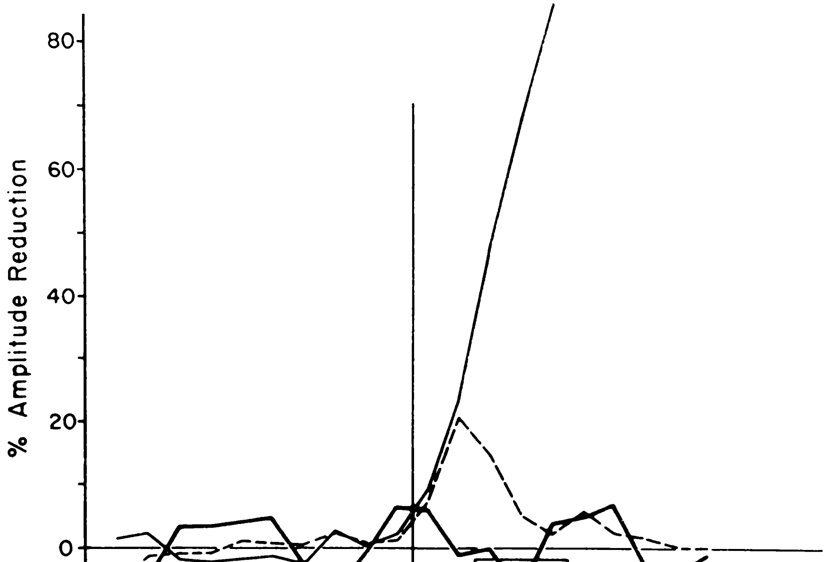

Fig. 3. Comparison of respiratory CRs during L_{CS}–US, L_{CS}–CI, and L_{CS} alone trials. Ordinate, percent reduction from the mean respiratory amplitude of the 10 s pre-CS period. Abscissae, time in seconds from the beginning of the trial to 10 s after the CS onset period. Thin solid line, averaged for 18 L_{CS}–US trials during the 10 control sessions. Dashed line, averaged CR for 45 L_{CS}–CI trials during the 10 control sessions. Thick solid line, averaged CR for 6 L_{CS} alone trials during the first Test session.

the reduction was assumed to be a real phenomenon, probably reflecting the increasing tension of the body musculature in anticipation of the upcoming noxious US. This, of course, "degrades" the amplitude reduction response as being extrinsic to the respiratory function, i.e., a sort of peripheral interference at the level of the "final common path", but it does not detract from its usefulness as an integrated index of emotional state, or preparatory readiness for the impending aversive US. In contrast to the complex HR response, this physiological concomitant of the CS was a monotonic and almost linear change from the baseline with the peak precisely timed at the moment of the delivery of the US. Figures 3 and 4 compare computer plots of respiratory amplitude changes to L_{CS} and A_{CS} from three sets of trials: CS–US during the CONTROL stage of the experiment (thin line), CS–CI trials during the same period (dashed line), and CS alone during the first day of

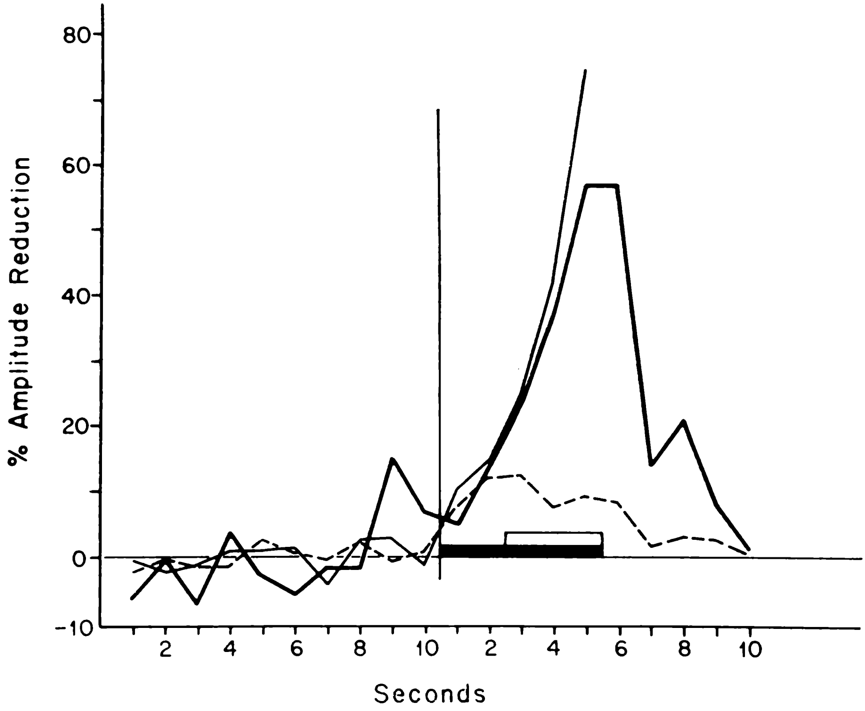

Fig. 4. Comparison of respiratory CRs during A_{CS}–US, A_{CS}–CI, and A_{CS} alone trials. Ordinate, percent reduction from the mean respiratory amplitude of the 10 s pre-CS period. Abscissae, time in seconds from the beginning of the trial to 10 s after the CS onset period. Thin solid line, averaged CR for 17 A_{CS}–US trials during the 10 control sessions. Dashed line, averaged CR for 50 A_{CS}–CI trials during the 10 control sessions. Thick solid line, averaged CR for 6 A_{CS} alone trials during the first Test session.

TEST stage (heavy line). The reduction in amplitude is expressed as upward deflection and scored as percentage change in relation to the mean amplitude during the 10 s prior to the CS onset[3]. The 100% response in our plot corresponds to zero amplitude, while the negative values denote an increase in amplitude. An example of the reliability and stability of these responses is provided by the next Fig. 5, in which the responses to a nonreinforced L_{CS} of 2 s duration are shown for the first 12 sessions of the stage 2 of the experiment. Each plot is an average of only six responses, but the shape of each plot is an inverted V

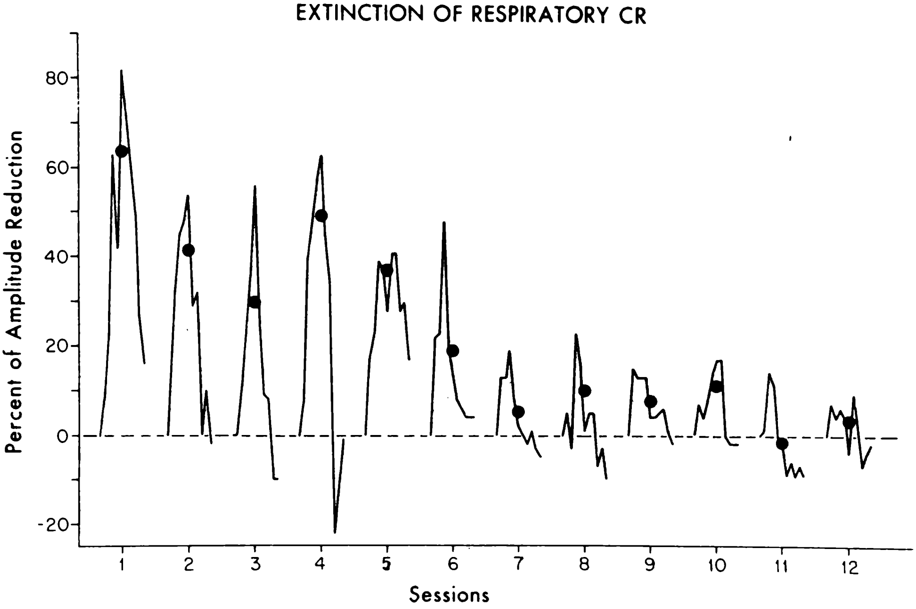

Fig. 5. Respiratory CRs during unreinforced 2 s light CS trials in PROTECTION/ EXTINCTION sessions one through twelve. Each curve, the average of six respiratory CRs, begins from a baseline of the average mean-pre-CS-amplitude and continues with ten values, for the 10 s after the CS onset, expressed as percent of the reduction relative to the pre-CS mean. The solid circles represent the average reduction throughout a five second period starting three seconds after the CS onset. Abscissa, percent amplitude reduction from mean pre-CS level.

[3] Expressing it as percentage of the pre-CS amplitude was indicated by the following considerations. Although the position of the thermistor should not change within a session, it varies between sessions, as does the gain setting of the amplifier. Occasionally the adjustment in gain has to be made between trials, so that the recorded quiet breathing in the intertrial intervals be of comparable size: 10–20 cm on the XY plotter or 1–2 V on the FM tape. Thus, the amplitude is arbitrary and the best expression of the change is in relation to some standard period, as in our case, the 10 s pre-CS period.

with the peak at about 5 s after the onset of the CS, i.e., precisely when the shock was presented on the CS–US trials in the earlier training. Even more astonishing is this timing of the peak when one realizes that the CS duration was only 2 s in this stage of training, so the timing of the peak could not be cued to the termination of the CS.

Returning to the Figs. 3 and 4, we find that on the CS–CI trials this respiratory response, already initiated in the first 2 s of CS, was promptly suppressed by the CI. The third, and most revealing plot, from the first day of the TEST stage, when the L_{CS} and A_{CS} were both presented six times nonreinforced for 5 s, fully supports the findings on the leg flexion and HR responses. The response to L_{CS} is extinguished, whereas the A_{CS} elicits a practically full size response. Again, the

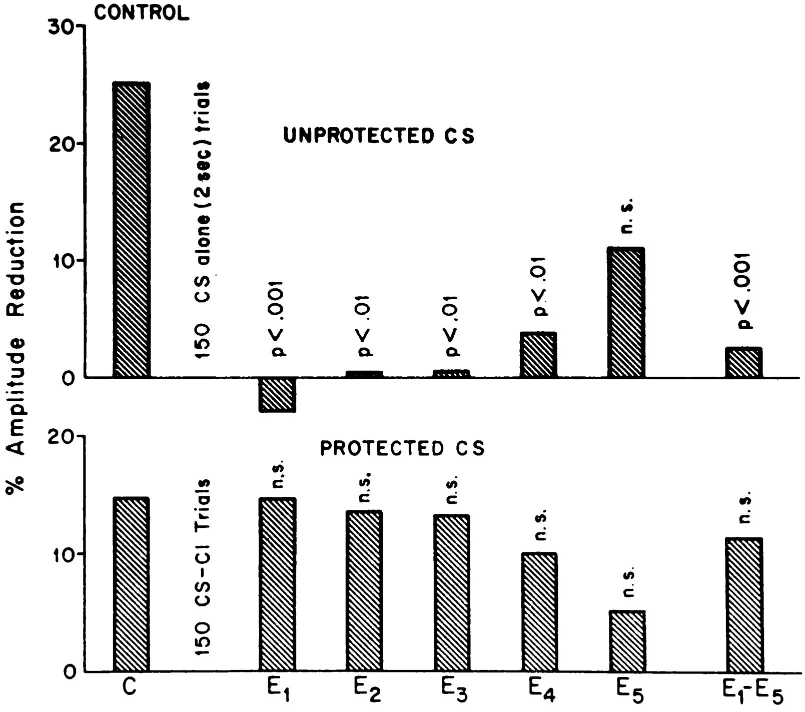

Fig. 6. Comparison of average respiratory CRs during the five Test sessions and the control sessions. Responses to Light-CS, above, and to Air CS, below. The columns represent averaged percent reducation computed in the following way: Respiration was sampled 1.5 s after the CS onset, the percent reduction of this sample from the mean pre-CS amplitude of that trial was determined, and the average percent reduction at 1.5 s was calculated. The columns represent these averages for all L_{CS}–US and A_{CS}–US trials in the control stage ($n = 12$), for the L_{CS} alone and A_{CS} alone trials in each of the five extinction sessions ($n = 6$), and for all five Test sessions combined ($n = 30$).

fact that the response was not extinguished during the last 3 s of the CS may be attributed to the non-elicitation when the CS–CI compound was presented 150 times in the second stage of the experiment. But, similarly to the HR response, the early part of the respiratory response, which *was* elicited on the CS–CI trials, also was preserved.

Figure 6 adds a little quantitative information about this early response. The scores from the 1.5 s point after the CS onset are presented from the CONTROL stage (from trials CS–US only) and compared by a t-test with the same response on each of the five TEST sessions, as well as with the average of all TEST sessions. In the case of the unprotected L_{CS} there was complete extinction on the first three days and only on the fifth day the difference is not statistically significant. No trace of extinction was found for the protected A_{CS} and only on the 4th and 5th day of testing (i.e., extinction) there was noticeable decrease in response, not attaining, however, statistical significance.

GENERAL DISCUSSION

The results presented in this admittedly preliminary report indicate that the phenomenon of PFE is fully demonstrable and powerful when a well trained CI (predictor of no US) is used. The protection from extinction was observed for all categories of conditioned responses studied in this experiment: leg flexion, heart rate change and change in the amplitude of respiration [4].

Thus, assuming tentatively that the PFE does in fact exist, let us briefly review a few theories of learning in order to find out if the PFE is a theoretically acceptable phenomenon.

Pavlov's hypothesis of "protective inhibition" (*14, lecture 22*; 1). This old-fashioned theory of the maintenance of conditioned "reflexes" satisfactorily predicts and "explains" the PFE. The CS–US pairing originally creates a new connection between CS and US cortical "centers". Afterwards, however, the conditioned reflex is maintained by the following mechanism. The CS activates the CS center. This activation, responsible for the CR, would linger and exhaust the neurons of this center if the US was not presented. The activation of the US center inhibits the CS center, or middle link of CS–US connection in Asratian's version of this hypothesis, and thus protects it from over-exhaustion; repeated omission of the US results in lowering reactivity (or "working capability" in Pavlov's words) which in overt behavior is ob-

[4] The PFE was aso observed in vocal CRs, not discussed in this paper.

served as extinction. The logical extension of this theory is that any other stimulus exerting inhibitory effects upon the center of the CS (or the CS–US bond) might also prevent extinction. Obviously the well-trained and specific CI is a prime candidate for this role. Unfortunately, this theory of Pavlov has many weaknesses, both empirical (it has never been verified) and logical; e.g., the inhibitory process is both a result of exhaustion and an active process preventing the exhaustion.

Consolidation hypothesis of associative learning. There are many versions of the hypothetical consolidation process which is initiated during the trial but takes place in the posttrial period. The simple version could be exemplified by the notion (6) of the coexistence of the traces of the stimuli which could interact and cause permanent changes in the common structural elements involved in the trace processes (e.g., neurons which participated in the reverberation of impulses, etc.). More elaborate accounts (post-trial "backward scanning" (9) or "rehearsal" (19) belong rather to the information processing and cognitive theories of learning. At any rate, the extension of the consolidation account to extinction learning (18) assumes the interaction between the trace of the CS and the trace of no US (assuming that the absence of the expected US is a perceived event equally capable of leaving a "trace")[5]. This interaction should lead to new learning, with the former CS becoming "extinguished", i.e., a predictor of no US. However, if the trace of a non-reinforced CS is terminated before it enters the post-trial phase of consolidation, no consolidation of a CS-no US association will occur. The CI might be such a terminator of a CS trace because it was trained to suppress the response elicited by the CS. Even more importantly, the CI, being a predictor of no US, may prevent also the initiation of the trace of the no US, which no longer comes as a surprise. This aspect of the presenting CI, is even better articulated in terms of Wagner's rehersal hypothesis of Mackintosh's áttentional theory (13).

Rescorla-Wagner model of classical conditioning. As pointed out by Henderson and Harris (7) the Rescorla-Wagner Model of conditioning (15) predicts PFE because there is no discrepancy between the V_A (current associative strength) and λ (limit of associative strength for a given reinforcer) on a CS/CI-noUS trial, when there is no response, and the value of λ for no US is assumed to be zero. However, the theory is not elaborated for evaluating the "net associative strength" for a compound of CS–CI when there is no full overlap of the eliciting and suppressing stimuli and the compound is arranged sequentially. It seems, however,

[5] Konorski (11) proposed elaborate arguments for such representation of no-stimuli in the brain.

that the "net" should refer either to the time instant when the reinforcer would normally be presented, or better to the onset of the posttrial period, when the "processing" of the, acquired during the trial, information begins. This would be in accord with Wagner's process-oriented development (19) of the original model. With such an amendment, the model is predictive of PFE, and, as will be discussed in the next section, allows also for the occurrence of the initial part of the CR to the CS prior to the CI onset.

Cognitive theory of avoidance conditioning (16). This theory includes a protection from extinction notion but applies it only to the expectancy of US, while allowing for the fear CR to be extinguished. In other words, CS in a CS–CI trial retains it signalling (predicting) informational role, because only CS alone would constitute a disconfirmation of a CS–US bond. On the other hand, the CS as an elicitor of emotional CR undergoes extinction on the CS–CI trial, because it is not reinforced by the US. Such a view of two levels of learning (cognitive and emotional) following different courses, preservation for expectancies, and extinction for emotions, was prompted by the data which seemingly disconfirmed the PFE and left unanswered questions of how the avoidance response is motivated in the lack of drive. If, however, the existence of PFE of the fear CR could be ascertained, the theory need not be abandoned. One solution would be to assume more interdependence between expectancies and CRs; it is rather displeasing in the present version of the theory that the elicitation of the expectation of shock which causes the subject to respond instrumentally to provide the expectancy of no US, could occur without an emotional CR, acquired originally by pairing the CS with the US. If any such discrepancy between the cognitive and conative levels of responding is possible, it should be rather in the opposite direction, when the subject is emotionally responding contrary to the better knowledge that nothing important is going to happen. It is the persistence of emotional CRs and not of expectancies that sends people to psychotherapists. And, contrarywise, expectancies alone seldom determine behavior if the emotions are absent or directed elsewhere. The other solution would come in the form of accepting both sources of motivation, emotional CRs and expectancies (in presence of preferences), interacting and summating. This adds to the complexity of the theory, but there is no reason to assume that a complex behavioral machinery would not be selected for in the evolution if it increases survival fitness.

This brief and, admittedly, very superficial review of a few theories of behavior, makes clear that PFE does not pose any threat or require major revision.

Retrograde protection from extinction of the CR elicited by the early portion of the CS, prior to the onset of the CI. If the PFE phenomenon is to be of help in explaining the resistance to extinction of avoidance responses the mechanism of protection must not depend upon prevention of the elicitation of the CR. As Seligman and Johnston (16) correctly point out, "protection of CS from extinction must occur while this very CS is eliciting a CR". Their theoretical objection to a role for PFE in avoidance centers on this issue. They argue that the reason a "conditioned inhibitor provides protection for a CS may be precisely because the CR is inhibited", and therefore that there is no cause to expect that the conditioned response elicited by the early portion of the CS, prior to the onset of the CI, could be protected. Our data fully support the possibility that PFE also applies to the initial part of the CR elicited prior to the onset of the CI, at least for the heart rate and respiratory responses.

Two issues are now raised: in what theoretical context can this retrograde protection from extinction (RPFE) be placed, and what might the mechanism of RPFE be, if it does not (as it logically must not) involve inhibition of the initial part of the CR? Neither Pavlov's nor consolidation accounts encounter any difficulties, because in both cases it is posttrial hypothetical aftereffects that determine the outcome of the trial, and as shown above, the CI may be assumed to prevent the normal course of the aftereffects of nonreinforcement.

The Rescorla-Wagner model encounters a problem, unless amended along Wagner's rehersal hypothesis. The original version of the model could be restated in the following way: What is elicited undergoes extinction on the nonreinforced trial. Therefore, the initial part of the CR elicited by the CS prior to the onset of the CI should be extinguished in the course of 150 CS–CI trials. There are even ways to explain RPFE without any amendment invoking posttrial processing of information. Assuming that the similarity and therefore generalization between the early and later portions of the CS is strong, one could hypothesize that protection of the CS_l (late part of the CS) somehow extends to the unprotected CS_e (early part of the CS). In other words, as long as the CS_l retains the capacity to elicit the CR, also the CS_e will remain functional elicitor of the CR. The experiment with a two-compound CS, where CS_e and CS_l are very different, may verify this hypothesis. Such a bipartite CS should exhibit much less or no RPFE if it is due to the generalization between CS_e and CS_l.

Another mechanism for RPFE might be derived from the ethological concept of a "preparatory-consummatory" sequencing of behavior.

If consummatory CR *must* be preceded by the preparatory response, then PFE of the former should automatically preserve the anteceding preparatory CR. In our subject, the consummatory CR was a leg flexion and the fear CR might be considered a preparatory CR. If HR and respiratory changes are in some way peripheral concomitant indices (not necessarily perfect correlates) of this preparatory CR, then their RPFE could be explained by the notion of the sequential structure of the behavior conditioned with the aversive US.

An interesting modification of the Rescorla-Wagner model by Frey and Sears (5) endows the CI with a property of a dynamic, acquired control of salience. If this property of reducing attention could be assumed as also reducing posttrial processing (according either to "consolidation of traces" or "rehearsal" accounts), the improved model would be predictive of the RPFE.

The hybrid theory of Seligman and Johnston (16) does not offer any solution for the RPFE. In the course of long training with CS–CI trials only, there should be a build-up of an expectancy that CS_e will be followed by the CS_1 + CI. So even the informational content (that of predicting US) of the cognitive response to CS_e should undergo diminution during the series of CS–CI trials. The only hope lies in the separation of cognitive response from the conditioned one. Even if the expectancies tend to change, due to the absence of CS–US trials, the PFE mechanism, operating on the CR level, may still prompt the organism to assume a conservative attitude and to respond, contrary to the "knowledge", with the fear CR. But in such a case the cognitive part of the Seligman theory becomes redundant, at least as an explanation of the resistence to extinction of the avoidance responses.

Finally, one can adopt a neuroethological inclination (i.e., a state of mind rather than theory), which, superficially trivial, may be most practical for the student of behaving organisms (as opposed to the study of behavior or learning as such). Firstly, within this broad orientation one may assume that if RPFE has any survival value for the organism, for instance, by preserving the avoidance response without requiring that the organism be periodically exposed to the aversive events (8), then there is a good chance that the mechanism providing the RPFE will be invented (by chance variability of neurobehavioral machinery) and selected for in the evolution of the species. And secondly, the final understanding of the PFE according to this theoretical bias should come from the studies on the neurobehavioral machinery, with or without the help of any formal theory of learning.

CONCLUDING REMARKS

In the final remarks, we would like to admit that we are aware of some weaknesses of our experiment, besides its preliminary character. Presenting a short CS in the stage 2 (Protection/Extinction) and a long CS in stages 1 (Control) and 3 (Test) makes it possible for the animal to learn to discriminate between the short and long CSs. This has not happened in our subject (although he learned that discrimination later) but it may happen in others and this may reduce the difference in responding to protected and unprotected CSs in the stage 3. One remedy would be using long CSs in all three stages of experiment, but this would put the unprotected CS in a disadvantaged position relative to the protected one, as discussed in the Introduction. The difference in the CR eliciting properties would be probably very large in the stage 3 but it would not be so convincing, because one could argue that the CI is masking the protected CS and therefore, the comparison is made between 5 s CS_1 and 2 s CS_2 being extinguished in stage 2. Of course, the extinction of a short duration CS should be less complete than extinction of the longer duration CS, especially if we compare their response eliciting properties in the stage 3 using 5 s duration for both of them. Another solution, using a neutral stimulus instead of CI in a CS_2-CI compound (instead of a short, non-reinforced CS) is also not a good solution because with our prolonged training, such a stimulus will acquire the role of a CI. A between-subject design would be more satisfactory, but that would require much larger number of subjects, which in a study lasting many months is a costly proposition.

A better solution for these paradigmatical and theoretical problems would be a trace conditioning design with all CSs of 2 s duration and the CS–US interval of 5 s. The 3 s CI would be presented in the "gap" between the termination of the CS and the time point where the US occurs on the CS–US trials. The CS would then remain unchanged in all stages of the experiment. This, however, would increase the generalization between the CSs (the traces of the CSs are characterized by stronger generalization than the actual stimuli) so the differences between the protected and unprotected CSs in the stage 3 might be diminished. And, more importantly, this would be analogous to using a two-component CS, with the CS_e being actual stimulus and CS_l its trace, so the discrimination between them might be easier than in the case of a 5 s CS, and as a result of this, the RPFE might be diminished if, as argued before, it depends on the generalization along the CS duration. Still it seemed to us that the trace conditioning is possibly as good a design as the one described in this paper and we are currently training one subject using only short CSs [6].

Thus, although each has it flaws, the present design and the trace conditioning design will be used in our study to add more weight to the evidence for both PFE and RPFE which we consider a real and important phenomenon in the realm of behavioral plasticity.

Finally, we should confess our belief that the PFE phenomena are as important in the maladaptive as they are in normal behavior. Many neurotic or persistent maladaptive responses may be explained by the operation of the PFE mechanism. Better understanding of this phenomenon may, therefore, be of considerable diagnostic and therapeutic value.

We wish to thank W. J. Wilson for reading the manuscript, and G. LeNae Boddie for helping prepare the manuscript for publication. This investigation was supported by USPHS HD-05958.

REFERENCES

1. ASRATYAN, E. A. 1969. Mechanism and localization of conditioned inhibition. Acta Biol. Exp. 29: 271–291.
2. CAPALDI, E. J. 1967. A sequential hypothesis of instrumental learning. *In* K. W. Spence and J. T. Spence (ed.), The psychology of learning and motivation, Vol. 1. Academic Press, New York, p. 67–156.
3. CHORĄŻYNA, H. 1957. Some data concerning the mechanism of conditioned inhibition. Bull. Acad. Pol. Sci. 5: 387–392.
4. CHORĄŻYNA, H. 1962. Some properties of conditioned inhibition. Acta Biol. Exp. 22: 5–13.
5. FREY, P. W. and SEARS, R. J. 1978. Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. Psychol. Rev. 85: 321–340.
6. HEBB, D. O. 1958. A textbook of psychology. W. B. Sounders Co., Philadelphia.
7. HENDERSEN, R. W. and HARRIS, K. 1979. Inhibitory protection of conditioned fear extinction. Acta Neurobiol. Exp. 39: in print.
8. HULL, C. L. 1929. A functional interpretation of the conditioned reflex. Psychol. Rev. 36: 495–511.
9. KAMIN, L. J. 1969. Selective association and conditioning. *In* N. J. Mackintosh and W. K. Honig (ed.), Fundamental issues in associative learning. Dalhousie University Press, Halifax.
10. KONORSKI, J. 1948. Conditioned reflexes and neuron organization. Cambridge University Press, London, 267 p.
11. KONORSKI, J. 1967. Integrative activity of the brain. An interdisciplinary approach. University of Chicago Press, Chicago, 531 p.

[6] The results of this experiment fully replicated the findings of this paper.

12. LOLORDO, V. M. and RESCORLA, R. A. 1966. Protection of the fear-eliciting capacity of a stimulus from extinction. Acta Biol. Exp. 26: 251–258.

13. MACKINTOSH, N. J. 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychol. Rev. 82: 276–298.

14. PAVLOV, I. P. 1928. Lectures on conditioned reflexes. International Publishers, New York.

15. RESCORLA, R. A. and WAGNER, A. R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *In* A. H. Black and W. F. Prokasy (ed.), Classical conditioning II: Current research and theory. Appleton-Century-Crofts, New York.

16. SELIGMAN, M. E. P. and JOHNSTON, J. C. 1973. A cognitive theory of avoidance learning. *In* F. J. McGuigan and D. B. Lumsden (ed.), Contemporary Approaches to conditioning and learning. John Wiley and Sons, New York, p. 69–110.

17. SOŁTYSIK, S. 1960. Studies on the avoidance conditioning: 3. Alimentary conditioned reflex model of the avoidance reflex. Acta Biol. Exp. 20: 183–192.

18. SOŁTYSIK, S. and ZIELIŃSKI, K. 1963. The role of afferent feedback in conditioned avoidance reflex. *In* Gutmann and P. Hnik (ed.), Central and peripheral mechanisms of motor functions. Publishing House of the Czechoslovak Academy of Sciences, Prague, p. 215–221.

19. WAGNER, A. R. 1976. Priming in STM: An Information processing mechanism for self-generated depression in performance. *In* T. J. Tighe and R. N. Leaton (ed.), Habituation: Perspectives from child development, animal behavior, and neurophysiology. Erlbaum, Hillsdale N. Y.

Stefan S. SOLTYSIK and George WOLFE, Mental Retardation Research Center, School of Medicine, University of California at Los Angeles, Los Angeles, California 90024, USA.