TWO-FACTOR THEORY HAS STRONG EMPIRICAL EVIDENCE OF VALIDITY

BEN A. WILLIAMS

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Traditional two-factor theory is supported by parallels in the clinical literature. Theoretical problems with two-factor theory are obviated by the role of safety signals, which serve as positive conditioned reinforcers and retard the extinction of conditioned fear.

Key words: avoidance learning, safety signals, conditioned reinforcement, learned helplessness

One of my first courses as a graduate student was a seminar entitled ''Motivation and Action'' taught by Richard Herrnstein. This was at the time that Herrnstein was writing his 1969 paper on avoidance theory, which serves as one of the major antagonists to Dinsmoor's (2001) cogent conceptual critique of shock-frequency reduction as an account of avoidance data. Even as a graduate student I shared Dinsmoor's current assessment because it was not apparent how shock-frequency reduction could make contact with the rat's behavior unless a discrimination was formed between the stimulus complex prior to bar pressing versus that after bar pressing. Given such a discrimination, there seemed no valid reason to reject the assumption of two-factor theory that it was the reduction in aversiveness caused by the transition in the stimulus complex that served as the reinforcer that maintained the avoidance response. Whether one made the additional assumption that conditioned fear had developed due to Pavlovian conditioning to the stimulus complex preceding a bar press seemed to be more an issue of theoretical style than empirical substance.

Although I was an agnostic about this issue of theoretical style when I was a graduate student, I now believe that there are significant empirical reasons for the utility of conditioned fear as a theoretical construct, both in terms of the details of the data and in terms of its theoretical consistency with other areas of research that seemed initially to be only tangentially related. Two-factor theory also provides one of the most persuasive behavioral accounts of important clinical phenomena.

Perhaps the most compelling clinical example is obsessive-compulsive behavior, which follows the pattern predicted by two-factor theory in remarkable detail. For patients who engage in compulsive hand washing, for example, the behavior is triggered by the exposure to some anxiety-provoking stimulus, such as medical scenes, accident sites, or newspaper pictures of mayhem, with the anxiety that is produced persisting until the hand-washing behavior has occurred, at which time the anxiety immediately dissipates. There is a very close correspondence between the reported urge to do the compulsive behavior and the experienced anxiety (Hodgson & Rachman, 1972). Furthermore, the behavioral treatment procedure that has been found most effective to eliminate the compulsive behavior is the same that has been found to be most effective in extinguishing avoidance behavior in the laboratory: exposure to the fear-producing stimulus while the avoidance response is prevented (Meyer, 1966).

The parallel between avoidance behavior in the laboratory and human neurotic behavior that has generated the most attention to two-factor theory is the extreme resistance to extinction that is characteristic of both. The failure of avoidance behavior to be extinguished in the laboratory has been viewed as problematic for two-factor theory, because continuation of avoidance behavior without presentation of the aversive stimulus should cause conditioned fear to be extinguished, which should abolish the negative reinforcer that two-factor theory postulates to maintain the avoidance response. Because of this prob-

Address correspondence to the author at the Department of Psychology, University of California, San Diego, La Jolla, California 92093-0109 (E-mail:bawilliams@ucsd.edu).

lem, Solomon and Wynne (1954) postulated the additional ''processes'' of ''partial irreversibility'' and ''conservation of anxiety.'' Such fudge factors now seem unnecessary because a significant part of the difficulty in extinguishing avoidance responding appears to be due to higher order conditioning that occurs within the different segments of the conditional stimulus (CS) for fear. Levis and Boyd (1979) provided important evidence for the role of higher order conditioning by comparing the rate of extinction when the CS was a single 15-s stimulus and when the CS was composed of a serial compound of three different 5-s stimuli (CS1-CS2-CS3). Extinction was much slower with the serial compound, because responding to CS1 was reinvigorated whenever the latency to respond was greater than the duration of CS1, thus causing CS1-CS2 pairing to be renewed. Similarly, when responding to CS2 finally also decreased sufficiently for CS3 to occur, the CS2-CS3 pairing reinvigorated both CS2 responding and CS1 responding on subsequent trials. For complete extinction to occur, all three elements of the compound had to be extinguished, which required a much greater amount of training than when the CS was a single stimulus. Stampfl (1987) has argued that a similar pattern of extinction also occurs for human neurotics.

It is important to recognize that the problem of resistance to extinction completely disappears if one entertains the ''safety-signal'' version of two-factor theory. There is substantial evidence that a CS− for an aversive unconditional stimulus (US) is functionally equivalent to a CS+ for a positive US (e.g., Fowler, Goodman, & DeVito, 1977; Rescorla, 1969). As argued by Dinsmoor (2001), the stimulus complex associated with the termination of a shock, or with a period of shock-free time, should serve as a positive conditioned reinforcer for avoidance responding. Given that these safety signals have acquired positive reinforcement value, avoidance responding should persist as long as that value is maintained.

An adaptation of the Rescorla–Wagner (1972) theory of Pavlovian conditioning to operant behavior in fact predicts that avoidance behavior should persist indefinitely during extinction unless there are changes in the conditioning context. If one assumes that any response associated with a negative discriminative stimulus for shock will be energized, the dynamics of such an explanation are as follows: Assume that the effect of an aversive stimulus is represented as a negative number, and the effect of a positive reinforcer is represented as a positive number. Initially, before avoidance responding occurs, situational cues, including the CS, acquire conditioned aversiveness (e.g., −100 value units). When an avoidance response occurs, the shock is not presented. Thus, the level of conditioned value appropriate to the absence of a reinforcer is zero, so a discrepancy exists between this zero value and the existing level of conditioned fear to the situation. The result is that the compound stimulus of the situational cues and response-produced cues will gain value in order to reduce this discrepancy. The situational cues begin with a negative value and so become less negative (e.g., −50 units of value), whereas the response-produced cues begin with zero value and thus acquire positive value (e.g., +50 value units). When the subject responds again, with no shock presented, again the value expectation appropriate to the absence of shock is zero. But now the sum of the values of the situational cues and response-produced cues is also zero because the negative value of the situational cues is canceled by the positive value of the response-produced cues. This produces an equilibrium that causes further learning to cease, but at that time the response-produced cues have positive value so that the response should persist indefinitely.

The reason that extinction of avoidance behavior does eventually occur can then be explained by random variation in the situational cues (e.g., the time since the chamber was last cleaned). Using the values from the preceding example, this would mean that the negative value of the situation would be reduced by generalization decrement, let us assume, to −30 units. Then the sum of this negative value with the positive value of response-produced cues would be discrepant with the zero value appropriate to the absence of a reinforcer. To reach an equilibrium, the values of the situational cues and the response-produced cues would both be reduced, so that the response would decrease in strength. Further trial-to-trial random variations in the experimental context would

eventually result in the response-produced cues having zero value, so operant responding should no longer occur.

The preceding analysis depends upon the assumption that conditioned inhibitors with respect to shock (safety signals) are equivalent to stimuli paired with positive reinforcers. Substantial evidence exists for such equivalence (see Dickinson & Pearce, 1977, and Mackintosh, 1983, chap. 7, for discussions).

The concept of a safety signal also elucidates other important phenomena. Noted by Dinsmoor (2001) is its importance for understanding the finding of preference for signaled over unsignaled response-independent shock. According to the standard two-factor theory, this preference seems paradoxical because the signaled shock alternative presumably involves not only the aversiveness of the shock but also the conditioned aversiveness of the signal. However, when it is appreciated that the situation in the absence of shock is a safety signal, this paradox disappears.

A second application of the theoretical utility of the concept of the safety signal is the phenomenon of learned helplessness. When animals are subjected to uncontrollable aversive stimulation, their ability to learn subsequent operant behavior using avoidance contingencies is severely impaired, relative both to untrained controls and to subjects that have received the same amount of aversive stimulation in the initial phase of training but with escape or avoidance contingencies that allow termination of the shock. The usual procedure is to yoke the inescapable-shock subjects to the subjects that have the response contingency, so that the duration and spacing of shocks are made the same for the two groups. Although the initial interpretation of this finding was that the learned helplessness effect was due to the learning of a ''generalized expectancy'' of there being no relation between responding and aversive stimulation, the addition of safety signals to the procedure has shown this interpretation to be unnecessary. Jackson and Minor (1988) demonstrated that the deficit normally due to pretraining with inescapable shock could be eliminated when safety signals were presented at the time of shock-free periods that were produced by the responding of the master subjects to terminate the shock. Analysis of this finding suggested that the effect of the safety signal was to reduce fear of the experimental context (also see Mineka, Cook, & Miller, 1984). For the master subjects the stimuli correlated with having just made a response presumably served as the safety signal, whereas for the yoked subjects with inescapable shock it was necessary to explicitly signal the times when shock was not likely to occur. Thus, learned helplessness apparently results from the subject being in a highly aversive environment in which there are no signaled periods of relief from the aversiveness. In other words, learned helplessness is the result of the predictability of the absence of shock, rather than the availability of a response contingency that allows the subject to control the shock's presentation. The fact that the master subjects with the response contingency never develop helplessness even without explicit safety signals has the further implication that safety signals are generated by the animal's own behavior being predictive of shock-free times.

Although two-factor theory and safety-signal theory are sometimes portrayed as competing accounts, there is no reason why their separate underlying processes cannot operate simultaneously. In fact the safety signals produced by the avoidance response may not only provide positive reinforcement for that response but also protect the fear conditioned to the CS from extinguishing. An important but relatively unexplored finding in the Pavlovian conditioning literature (e.g., Soltysik, Wolfe, Nicholas, Wilson, & Garcia-Sanchez, 1983) is that presentation of a previously established conditioned inhibitor following the CS presentation during extinction greatly retards the rate of extinction. If a similar effect operates in avoidance procedures, classical two-factor theory and safety-signal theory are complementary rather than in conflict. Taken together, they appear to provide a comprehensive explanation of all known aspects of avoidance behavior.

REFERENCES

Dickinson, A., & Pearce, J. M. (1977). Inhibitory interactions between appetitive and aversive stimuli. *Psychological Bulletin, 84,* 690–711.

Dinsmoor, J. A. (2001). Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing. *Journal of the Experimental Analysis of Behavior, 75,* 311–333.

Fowler, H., Goodman, J. H., & DeVito, P. L. (1977). Across-reinforcement blocking effects in a mediational test of the CS's general signalling property. *Learning and Motivation, 8,* 507–519.

Herrnstein, R. J. (1969). Method and theory in the study of avoidance. *Psychological Review, 76,* 49–69.

Hodgson, R. J., & Rachman, S. (1972). The effect of contamination and washing in obsessional patients. *Behaviour Research and Therapy, 10,* 111–117.

Jackson, R. L., & Minor, T. R. (1988). Effects of signaling inescapable shock on subsequent escape learning: Implications for theories of coping and "learned helplessness." *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 390–400.

Levis, D. J., & Boyd, T. L. (1979). Symptom maintenance: An infrahuman analysis and extension of the conservation of anxiety principle. *Journal of Abnormal Psychology, 88,* 107–120.

Meyer, V. (1966). Modification of expectations in cases with obsessional rituals. *Behaviour Research and Therapy, 4,* 273–280.

Mackintosh, N. J. (1983). *Conditioning and associative learning.* New York: Oxford University Press.

Mineka, S., Cook, M., & Miller, S. (1984). Fear conditioned with escapable and inescapable shock: Effects of a feedback stimulus. *Journal of Experimental Psychology: Animal Behavior Processes, 10,* 307–323.

Rescorla, R. A. (1969). Establishment of a positive reinforcer through contrast with shock. *Journal of Comparative and Physiological Psychology, 67,* 260–263.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: Vol. 2. Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Solomon, R. L., & Wynne, L. C. (1954). Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review, 61,* 353–385.

Soltysik, S., Wolfe, G. E., Nicholas, T., Wilson, W. J., & Garcia-Sanchez, J. L. (1983). Blocking of inhibitory conditioning within a serial conditioned stimulus-conditioned inhibitor compound: Maintenance of acquired behavior without an unconditioned stimulus. *Learning and Motivation, 14,* 1–29.

Stampfl, T. G. (1987). Theoretical implications of the neurotic paradox as a problem in behavior theory: An experimental resolution. *The Behavior Analyst, 10,* 161–173.

Received November 10, 2000
Final acceptance December 12, 2000