JOURNAL OF RESEARCH IN PERSONALITY 9, 253–269 (1975)

Disinhibition of Aggression through Diffusion of Responsibility and Dehumanization of Victims

ALBERT BANDURA, BILL UNDERWOOD, AND MICHAEL E. FROMSON

Stanford University

The present study tested derivations from social learning theory on the disinhibition of aggression through processes that weaken self-deterring consequences to injurious conduct. Subjects were provided with opportunities to behave punitively under diffused or personalized responsibility toward groups that were characterized in either humanized, neutral, or dehumanized terms. Both dehumanization and lessened personal responsibility enhanced aggressiveness, with dehumanization serving as the more potent disinhibitor. Escalation of aggression under conditions of dehumanization was especially marked when punitiveness was dysfunctional in effecting desired changes. The uniformly low level of aggression directed toward humanized groups, regardless of variations in responsibility and instrumentality of the conduct, attested to the power of humanization to counteract punitiveness. Results of supplementary measures are consistent with the postulated relationship between self-disinhibiting processes and punitiveness. Dehumanization fostered self-absolving justifications that were in turn associated with increased punitiveness. Findings on the internal concomitants of behavior performed under different levels of responsibility suggest that reducing personal responsibility heightens aggressiveness more through social than personal sources of disinhibition.

Psychological explanations of aggression traditionally have been concerned with individual injurious acts that are aversively motivated. In most of these accounts, aggression is attributed not only to a limited set of instigators but to a narrow range of regulative influences as well. In recent years social scientists have enlarged their view of what constitutes aggression and have begun to reexamine its sources from a broader perspective.

A complete theory of aggression must explain how aggressive patterns are developed, what provokes people to behave aggressively, and what regulates their aggressive actions. Figure 1 summarizes the determinants of these separable processes from the perspective of social learning theory (Bandura, 1973). This theory is designed to encompass different

This research was supported by Public Health Research Grant M-5162 from the National Institute of Mental Health. The authors are indebted to Caleb Blodgett, Candice Bonner, and Gary Zinik, who assisted us in the research. Requests for reprints should be sent to Albert Bandura, Department of Psychology, Stanford University, Stanford, California 94305. Dr. Underwood is now at the University of Texas at Austin.

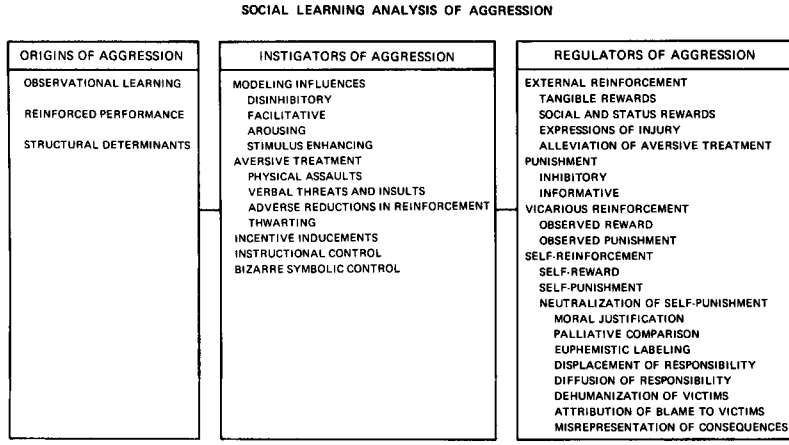

FIG. 1. Schematic outline of the origins, instigators, and regulators of aggression in social learning theory.

facets of aggression, whether individual or collective, personally or institutionally sanctioned.

Among the processes controlling aggressive behavior, self-reinforcement plays an especially influential role. After people acquire standards of conduct through modeling and selective reinforcement, they partly regulate their own actions by self-created consequences. They do things that give them self-satisfaction and a sense of self-worth but refrain from conduct that produces self-devaluative consequences. Internalization of standards, however, does not create an invariable control mechanism within the person. Because activation of self-reinforcement is under discriminative control, variations in moral conduct often occur with the same internalized moral codes.

Self-generated consequences can be disengaged from censurable acts by self-exonerating practices (Bandura, 1973; Kelman, 1973; Sanford & Comstock, 1971). One method is to make reprehensible behavior personally and socially acceptable by construing it in terms of high moral principle. Euphemistic labeling provides a convenient linguistic device for masking reprehensible activities or according them a respectable status. Self-deplored acts can also be made benign by contrasting them with more flagrant inhumanities. Moral justifications and palliative comparisons can serve as effective disinhibitors of aggression because they not only eliminate self-generated deterrents but also engage self-reward in the service of inhumane conduct. What was morally unacceptable becomes, through cognitive restructuring, a source of self-pride.

Self-reinforcing consequences are likely to be activated most strongly when the causal connection between moral behavior and its outcomes is

well defined. A common dissociative practice in everyday life is to obscure or distort the relationship between actions and the effects they cause. People will behave in aggressive ways they normally repudiate if a legitimate authority sanctions their conduct and acknowledges responsibility for its consequences (Milgram, 1974). By displacing responsibility elsewhere, people need not hold themselves accountable for what they do and are thus spared self-prohibiting reactions.

Exemption from self-censure can be facilitated by diffusing responsibility for culpable behavior. Through division of labor, division of decision making, and collective action, people can behave injuriously without feeling personally responsible. Anonymity can weaken restraints socially by reducing the chances of being blamed and punished by others for misdeeds. To the extent that dispersing responsibility also obscures individual contributions to collectively produced consequences, it can reduce personal restraints by attenuating self-disapproval.

Attributing blame to the victim is still another exonerative expedient. Victims are faulted for bringing suffering on themselves, or extraordinary circumstances are invoked as vindications for punitive conduct. One need not engage in self-reproof for committing acts dictated by circumstances.

The strength of self-evaluative reactions partly depends upon the characteristics of the people toward whom actions are directed. Inflicting harm upon individuals who are regarded as subhuman or debased is less apt to arouse self-reproof than if they are seen as human beings with dignifying qualities. The reason for this is that people who are reduced to base creatures are likely to be viewed as insensitive to maltreatment and influenceable only through the more primitive methods. Dehumanizing the victim is therefore a further means of reducing self-punishment for cruel actions. Additional ways of weakening self-deterring reactions operate by misrepresenting the consequences of actions. As long as detrimental effects are ignored or minimized there is little reason for self-censure.

Disinhibiting phenomena are amply documented by naturalistic accounts of collective violence, but they have only recently received systematic study under controlled conditions (Zimbardo, 1969; Diener, Note 1). Evidence is needed on how sanctioning arrangements interact to weaken restraints over aggression and on the mechanisms by which they achieve their effects.

EXPERIMENT 1

The present experiment investigated in a factorial design the influence of diffusion of responsibility and victim dehumanization on aggression

when control of punitive actions relied mainly on self-deterrents. Groups of subjects were provided with opportunities to behave punitively toward others under conditions of personalized or diffused responsibility. Within each of these conditions, the targets of aggression were characterized in either humanized, neutral, or dehumanized terms. It was hypothesized that both dehumanization and diffused responsibility would reduce self-deterring responses and enhance aggressiveness. Dehumanization was expected to be a more powerful disinhibitor of aggression under diffused than under individualized conditions of responsibility.

Self-absolving practices alone are unlikely to transform otherwise considerate people into callous aggressors instantly. Rather, the change is usually achieved through a process of gradual desensitization in which distress and self-reproof are extinguished with repeated performance of aggressive acts. It was predicted that aggression would be disinhibited more rapidly when responsibility is diffused and victims are dehumanized than when people are held personally responsible for their actions and victims are humanized. There was no a priori basis, however, for predicting the relative disinhibitory power of the responsibility and dehumanization conditions.

Method

Subjects

A total of 72 paid male volunteers from junior colleges participated in the study. The experiment employed a 3×2 factorial design based upon three variations in victim labeling and two levels of responsibility. Punishment trials constituted the repeated measure of aggression. Twelve subjects were randomly assigned to each experimental condition.

Apparatus

The experimental room was partitioned into three identical cubicles so that subjects could neither see nor communicate with each other. Each cubicle contained an aggression device for delivering shocks in 10 levels of intensity. A red light on the apparatus signaled inadequate solutions ostensibly recommended by a group of decision makers. Preconceptions concerning the personal characteristics of the decision makers were conveyed through headphones connected to a tape recorder.

Procedure

Subjects arrived in groups of three in all conditions and were assigned as a triad to one of the treatments. As the experimenter escorted them to the experimental room, his assistant announced, within hearing distance, that he would begin the instructions to the group of decision makers who had presumably arrived earlier and were assembled in a nearby room. The experimenter explained to the subjects that the study was concerned with the effects of punishment on the quality of collective decision making. The task was presented as an analog of organizational procedures in which members of a task group jointly decide the best course of action from several alternatives and project supervisors take corrective steps if advisors propose solutions that prove inadequate. It was further explained that in

order to increase the generality of the findings, participants from different social backgrounds were being included in the project but that each group was composed of people with similar attributes. Information on the homogeneous composition was included so that characterizations of the group would apply to individual members, as in conventional stereotyping.

The subjects were informed that they would serve as supervisors for a three-member team of decision makers assigned the task of formulating collective solutions for a series of bargaining situations. On each of 25 bargaining trials they would choose one of several options that, the experimenter explained, were known to vary in effectiveness. Whenever the decision makers proposed effective bargaining solutions, an amber light would flash on the supervisors' display panel, which required no action. However, on occasions when they recommended inadequate solutions, as signaled by the illumination of the red light, the supervisors were to punish them by administering a shock in any intensity they saw fit. They were informed that the ten shock intensities ranged from mild at level 1 to painful at level 10. Thus, participants who did not wish to behave punitively could easily meet the task requirements simply by selecting the weakest shock levels. Since, in fact, there were no decision makers, no pain was inflicted by subjects' punitive actions.

After the task was described, but before the treatment conditions were created, subjects were informed that they were free to leave without loss of pay if they did not wish to participate in the study. Eight declined. In these instances the other members of the triad were also dismissed because observed refusal could affect their subsequent behavior.

Responsibility Conditions

Subjects in the *individualized responsibility* condition were told that each of them was assigned supervisory responsibility for a member of the decision-making team and that they personally determined the level of shock that their supervisee received on each punishment trial.

In the *diffused responsibility condition,* subjects were informed that the shock levels they selected were automatically averaged by the device so that the level of shock received by their supervisees represented the average of their collective decision on each punishment trial.

Labeling of the Victims

Upon completing the responsibility induction, the experimenter announced that he was leaving for the bargaining room and would issue further instructions through the intercom when the decision makers were ready to begin. The subjects donned their headphones, and the experimenter departed.

A short time later subjects overheard a brief interchange between the experimenter and his assistant through an apparent inadvertence that served as the vehicle for characterizing the decision makers. The interchanges were tape recordings that varied in content in accordance with the treatment condition. Each recording opened with a click of the microphone switch followed by the experimenter's reporting that the bargaining sessions would soon begin. At this point, he was suddenly interrupted by the assistant, asking where the scoring forms were stored. The distracting interruption provided the pretext for leaving the microphone on. As they searched in the back of the room, the experimenter was heard asking his assistant if the decision makers had completed their questionnaires. After replying, he remarked, in a brief aside, that the personal qualities exhibited by the group confirmed the views of the person through whom the participants were recruited. To minimize any implied social sanctioning of aggressive actions the remarks were made matter-of-factly as a reiteration of the groups' reported characteristics rather than as a personal evaluation.

For subjects in the *humanized condition,* the decision makers were characterized as a perceptive, understanding, and otherwise humanized group. By contrast, in the *dehumanized condition,* the decision makers were described as an animalistic, rotten bunch. In the *neutral condition,* no evaluative references were made as to the characteristics of the group.

The prerecorded exchanges ended with the experimenter's expressing muffled alarm on discovering that the microphone had been inadvertently left on. A click sounded as though it were abruptly shut off. After a short pause the microphone was again activated, whereupon the experimenter announced the start of the bargaining series.

Dependent Measure

Inadequate solutions were signaled on 10 of the 25 trials and distributed in the following predetermined pattern: trials 3, 5, 6, 8, 10, 14, 16, 21, 22, 23. The intensity of shock administered on these occasions served as the measure of aggressiveness.

Postexperimental Questionnaire

When the formal procedures were concluded, subjects filled out a questionnaire on which they rated their supervisees' responsiveness to disciplinary measures and wrote their reactions to their supervisory role. The latter item provided data on the self-disinhibiting processes accompanying the treatment conditions. In addition, subjects' evaluations of the supervisees were assessed by the semantic differential technique. The form used consisted of 6-bipolar rating scales using pairs of contrasting adjectives along dimensions of rigidity, intelligence, indolence, competence, impulsiveness, and sensitivity. After all the procedures were completed, the participants were encouraged to comment freely on their experiences and were provided with a full explanation of the experiment.

Results

The mean of the pooled ratings on the semantic differential scales provides a check on the success of the dehumanization treatments. The labeling procedures were highly effective in creating differential evaluations of the group members ($F = 15.71$, $p < .001$). All three conditions differed significantly from one another beyond the .01 level, with the dehumanizing, neutral, and humanizing characterizations inducing degrading, neutral, and favorable evaluations, respectively.

Level of Aggression

Figure 2 presents the mean intensity of shocks administered by subjects as a function of treatment conditions. The main effects on punitiveness of victim labeling ($F = 54.96$, $p < .001$) and responsibility ($F = 18.21$, $p < .001$) are both highly significant. The interactions of these two factors with each other and with shock trials are also significant sources of variance. There was no significant variation across triads within conditions.

The way in which labeling interacted with responsibility ($F = 3.56$, $p < .05$) to produce differences in aggressiveness is depicted graphically in Fig. 2. Comparisons between means of the different conditions reveal that subjects behaved more punitively under diffused than individualized

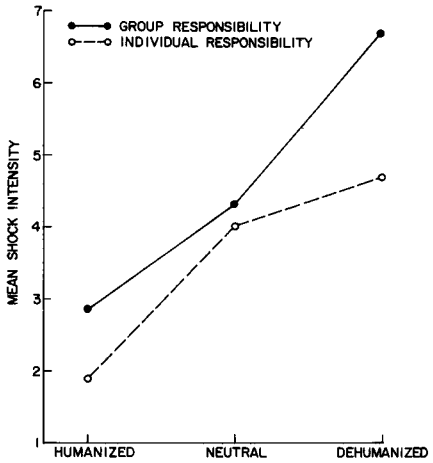

FIG. 2. Mean intensity of shocks administered by subjects as a function of diffusion of responsibility and dehumanization of the recipients of punishment.

responsibility when the group members were humanized ($t = 2.64$, $p < .01$) or dehumanized ($t = 3.33$, $p < .005$). Variations in responsibility did not affect aggressiveness, however, when the members were presented in neutral terms.

Under both responsibility conditions, aggressiveness mounted with dehumanization of the performers. Subjects who had a diminished sense of personal responsibility shocked dehumanized performers more severely than neutral ones ($t = 3.93$, $p < .001$), who, in turn, were treated more harshly than those invested with humanized qualities ($t = 3.62$, $p < .001$). A similar pattern of differences was obtained under personalized responsibility, although the rise in aggressiveness was less marked between the neutral and dehumanized treatments ($t = 1.93$, $p < .05$).

The significant interaction between labeling and trials ($F = 3.58$, $p < .001$) may be seen in Fig. 3. Intergroup differences were tested for significance at each trial. Characterization of the performers had a significant though weak effect on aggression at the outset. Compared to subjects in the humanized condition, those in the neutral ($t = 2.35$, $p < .025$) and dehumanized ($t = 3.01$, $p < .005$) treatments were slightly more punitive. On the second trial, subjects continued to treat the humanized performers in a relatively gentle manner but abruptly escalated their level of punitiveness with neutral ($t = 5.31$, $p < .001$) and dehumanized ($t = 6.41$, $p < .001$) members. The latter groups, however, did not differ significantly in this respect. On all subsequent occasions, the three groups were significantly differentiated from each other in levels of aggressiveness. Dehumanized performers were treated most punitively, humanized ones were spared painful shocks, while the neutral ones were administered an intermediate level of punishment.

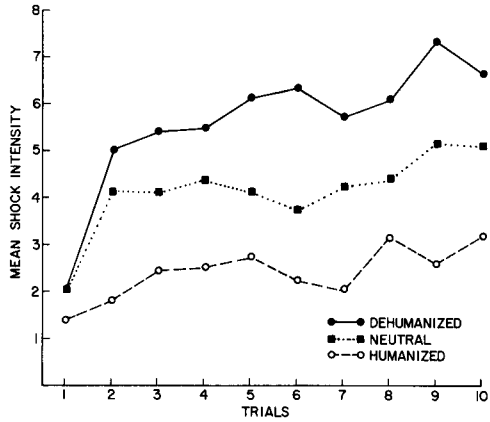

FIG. 3. Mean intensity of shocks administered on repeated occasions by subjects to performers who are represented as humanized, neutral, or devoid of humanness.

The interaction between responsibility and trials ($F = 1.94$, $p < .05$) is shown in Figure 4. Initially subjects behaved nonpunitively regardless of how responsible they were for the shocks inflicted on the performers. Once having aggressed, subjects operating under diffused responsibility promptly intensified their level of punitiveness. Beginning with the second trial and continuing through the eighth one, they administered significantly more severe shocks under diffused than under individualized responsibility. On the ninth trial, which signaled two consecutive inadequate performances, subjects who were personally responsible promptly heightened their punitiveness and did not differ significantly in this

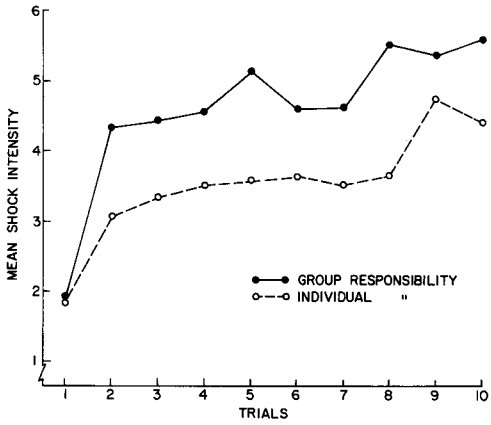

FIG. 4. Mean intensity of shocks administered on repeated occasions by subjects behaving under conditions of individualized or diffused responsibility for the effects of their actions.

respect from their counterparts in the diffused responsibility condition. When increased shock failed to improve the supervisees' performance on the subsequent trial, the individually responsible subjects reduced their punitiveness to a level significantly below that of subjects who were only partially responsible.

Self-Disinhibiting Processes

Forty-four percent of the subjects did not register any responses to the item concerning their reactions to having served in a punitive role, and the various groups did not differ in this respect. Nor did the nonresponders differ in level of punitiveness from those who recorded personal reactions. The responses of the latter subjects were scored for self-disinhibiting justifications or for expressions of disapproval of punitive measures.

In the justification category were included such responses as (a) ascribing culpability to the performers (e.g., "In many cases poor performance is indicative of laziness and a willingness to test the supervisor" and "People are basically evil and have to be put in their place"); (b) extolling the benefits or necessity of punishment (e.g., "It gets more efficiency out of the group" and "Although punishment is looked down upon, that's not going to influence me because I've seen it work"); (c) attributing their punitive behavior to situational or role requirements (e.g., "As an acting supervisor it was my job to punish poor performance" and "If doing my job as a supervisor means I must be a son of a bitch, so be it"); (d) displacing responsibility (e.g., "I administered shocks because I was told to"); (e) minimizing the painful consequences of their actions (e.g., "It would not hurt them too bad"); (f) disavowing conscious involvement in the activities (e.g., "I was reacting mechanically to the lights"); (g) emphasizing the prevalence of punishment (e.g., "Everyone is punished for something everyday").

Disapproval of punitive practices was indexed by such responses as (a) certifying the detrimental effects of punishment (e.g., "Physical punishment is not good. It just causes additional stress, which makes everything worse"); (b) affirming the relative superiority of nonpunitive methods (e.g., "People should not be punished for doing bad; they should be rewarded for doing good"); (c) registering concern over punishing people without sufficient acquaintance with them or the causes of their errors in judgment (e.g., "I felt uncomfortable because I was administering punishment without knowing whom I was punishing and also not knowing enough about their mistakes"); and (d) objecting to the use of excessive punishment on moral grounds (e.g., 'Morally and ethically, I couldn't give a painful punishment").

The responses were coded independently by two raters as representing self-disinhibition or disapproval or as inapplicable to these two

TABLE 1
PERCENTAGE OF SUBJECTS EXPRESSING SELF-DISINHIBITING JUSTIFICATIONS OR DISAPPROVAL OF PUNITIVE MEASURES AS A FUNCTION OF RESPONSIBILITY AND VICTIM LABELING

Treatment conditions	Self-disinhibiting justifications	Disapproval of punitiveness
Responsibility		
Diffused	18	50
Individualized	33	50
Victim labeling		
Humanized	0	81
Neutral	40	40
Dehumanized	43	21
Combined disinhibitors		
Individualized humanized	0	100
Individualized neutral	50	33
Individualized dehumanized	60	0
Diffused humanized	0	67
Diffused neutral	25	50
Diffused dehumanized	33	33

forms of self-response. In the latter instances subjects' reports contained no evaluative responses. Since subjects rarely expressed both justifying and disapproving evaluations of punitive behavior, they were rated as either self-disinhibiting or disapproving if they gave one or more responses within the same category. The raters agreed on 80% of their categorizations.

Table 1 shows the percentage of subjects who expressed self-disinhibiting or disapproving reactions under the different treatment conditions. Subjects who assumed personal responsibility for the punishments they administered were more inclined to generate self-absolving justifications than those operating under group responsibility, but the difference was not of statistically significant magnitude. Dehumanization, however, produced significant differences both in self-disinhibiting justifications ($\chi^2 = 8.92$, $p < .02$) and in disapproval of punitive sanctions ($\chi^2 = 11.24$, $p < .01$). When the performers were humanized, subjects strongly disapproved of physical punishment and rarely excused its use. By contrast, when performers were divested of humanness, subjects seldom condemned punitive techniques but often voiced self-absolving justifications. The neutral condition, in turn, produced an intermediate diversity of responses. Essentially the same pattern of results is replicated under both individualized and collective responsibility.

The previous analysis provides some evidence linking the sanctioning conditions to internal self-disinhibitory processes. In order to determine

whether self-disinhibition, in turn, is linked to punitive behavior, the shock intensities administered by self-exonerating subjects ($M = 5.24$) were compared with those of subjects who disapproved of punitive actions ($M = 3.41$). The findings show the self-exonerators to be significantly more punitive ($t = 2.54$, $p < .01$).

EXPERIMENT 2

In the preceding experiment aggression was abruptly heightened after subjects had administered punishment under circumstances conducive to self-disinhibition. One possible explanation for the precipitous rise in aggression is in terms of the apparent success of punitive actions. It will be recalled that in the early trials, punishment was followed by unerring solutions. Having apparently affected an improvement in performance through punishment, subjects who are freed of restraints might be quick to intensify their punitiveness when the next opportunity to behave aggressively arose.

It would be predicted from the process of self-disinhibition, however, that under dehumanizing conditions escalation of aggression would be even more precipitous when punitiveness is dysfunctional in improving performance. When punishment fails to achieve desired results, the negative feedback is likely to affect aggressors differentially, depending upon how they view their victims. Apparent lack of progress by degraded victims is apt to be interpreted as further evidence of their culpability and thereby justifies intensified punitiveness toward them. By contrast, a similar lack of improvement by valued individuals is more likely to be attributed to the disruptive consequences of punishment and thus supports self-restraints over aggression. The second experiment was primarily designed to test the hypothesized interactive effects of dehumanization and differential efficacy of punishment on the escalation of aggression.

Method

Subjects

The subjects were 72 paid volunteers from junior colleges. They were randomly assigned in groups of three to one of six experimental conditions.

Procedure

The procedures for measuring punitiveness and for labeling the performers were the same as those used in the preceding experiment. Subjects administered shocks in intensities of their own choosing for inadequate solutions allegedly made by supervisees characterized in either humanized, neutral, or dehumanized terms.

Aggressiveness was measured under conditions of individualized responsibility in all groups. Personalized rather than diffused responsibility was selected because the former

condition elicits lower levels of punitiveness and thus allows greater latitude for the escalation of aggression.

Within each of the three labeling treatments, half the subjects were assigned to the *functional aggression* condition. In this feedback series every punished trial was followed by a correct solution, which signifies that aggression was highly effective in improving the supervisees' performances. For the remaining subjects, who were assigned to the *dysfunctional aggression condition,* punishment trials were grouped sequentially so that aggression was repeatedly followed by failure. The predetermined pattern of punishment trials was 3, 5, 8, 10, 12, 15, 18, 20, 23, 25 for the functional series and 3, 4, 5, 9, 10, 16, 17, 18, 24, 25 for the dysfunctional sequence. Both groups thus administered punishment on 10 of the 25 trials, but the patterns of success and failure were arranged to convey differential efficacy for aggression.

Results

Level of Aggression

The significant effect on punitiveness of labeling ($F = 16.32$, $p < .001$) and the interaction between labeling and trials ($F = 2.70$, $p < .005$) replicates the findings of the initial experiment. Subjects were less aggressive when their actions were consistently successful than when their punishments usually failed to produce improvements in subsequent performance ($F = 6.71$, $p < .025$). The significant interaction between efficacy and trials ($F = 3.33$, $p < .001$) reveals that the differential escalation of aggression did not occur until after the second administration of shocks.

Efficacy feedback had quite different effects on punitiveness over trials, however, depending upon how the performers were portrayed. The significant triple interaction ($F = 2.43$, $p < .005$), which is depicted in Fig. 5, bears on the hypothesized relationship tested in this study.

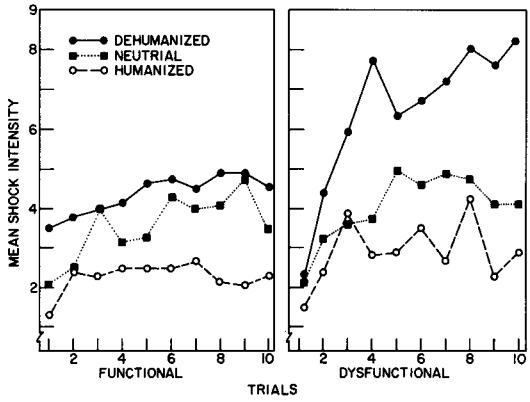

FIG. 5. Mean intensity of shocks administered to performers on repeated occasions as a function of dehumanization and efficacy of the punitive actions.

When punishment was consistently effective in rectifying errors, subjects gradually raised the intensity of shocks over trials with dehumanized and neutral performers but adhered to a consistently low level of punishment with humanized performers. Whereas at first neutral and humanized performers were treated comparably, as the trials progressed subjects punished neutral and dehumanized performers more severely and did not differ significantly from each other in this respect.

When punishment usually failed to eliminate errors, aggression was precipitously escalated to extreme levels with dehumanized performers; it increased gradually and then leveled off at a moderate level with neutral performers; and, after an initial rise, it deescalated and fluctuated around a low level of punishment with humanized performers.

Self-Disinhibiting Processes

Although results of the previous study on self-reactions to aggression are consistent with predictions, they require interpretative caution because of the reduced number of respondents. In the second experiment 97% of the subjects reported reactions on the questionnaire to their supervisory activities. These personal responses were coded independently by two raters, with 82% agreement in terms of the response categories described earlier.

The percentage of subjects in the various treatments who expressed self-disinhibiting or disapproving reactions is presented in Table 2. The

TABLE 2
PERCENTAGE OF SUBJECTS EXPRESSING SELF-DISINHIBITING JUSTIFICATIONS OR DISAPPROVAL OF PUNITIVE MEASURES AS A FUNCTION OF EFFICACY AND VICTIM LABELING

Treatment conditions	Self-disinhibiting justifications	Disapproval of punitiveness
Efficacy		
Functional	29	51
Dysfunctional	23	57
Victim labeling		
Humanized	4	74
Neutral	13	63
Dehumanized	61	23
Combined conditions		
Functional humanized	9	73
Functional neutral	25	58
Functional dehumanized	50	25
Dysfunctional humanized	0	75
Dysfunctional neutral	0	67
Dysfunctional dehumanized	73	27

internal concomitants of aggression did not differ as a function of instrumentality. Consistent with previous findings, self-disinhibiting justifications ($\chi^2 = 22.57$, $p < .001$) and repudiation of punitiveness ($\chi^2 = 11.95$, $p < .005$) varied under different labeling conditions. Subjects consistently disapproved of physical punishment with humanized performers, whereas they self-excused such conduct with dehumanized members. When they lacked knowledge about the performers, subjects' reactions were more variable though mainly self-disapproving of punitiveness. The differences in self-disinhibiting reactions were especially marked when aggression produced dysfunctional effects. The link between self-disinhibition and level of punitiveness was likewise replicated. Self-exonerators shocked performers more severely ($M = 5.24$) than did the disapprovers ($M = 2.95$), a difference that was highly significant ($t = 4.28$, $p < .001$).

DISCUSSION

Results of the present experiment confirm the hypothesized effects of responsibility, dehumanization, and the interaction of these factors on aggressiveness. Subjects behaved more punitively when responsibility was obscured by a collective instrumentality than when they were personally responsible for the amount of pain inflicted on others. Dehumanization had even greater disinhibitory power than did masking of responsibility links. Dehumanized performers were treated more than twice as punitively as those invested with human qualities and considerably more severely than the neutral group. As anticipated, dehumanization is especially conducive to aggression when people have a reduced sense of responsibility for the consequences of their actions.

The influential role played by victim labeling in the escalation process is most graphically revealed in changes in punitiveness over trials with differential feedback. Under conditions of functional feedback, subjects gradually increased their punitiveness toward dehumanized and neutral performers even in the face of evidence that weak shocks effectively improved performance and thus provided no justification for escalating aggression. By contrast, with humanized performers, subjects consistently adhered to mild punishment. The latter finding suggests some qualifying limits to the view that escalation of aggression is a general phenomenon that occurs independently of feedback or other factors (Buss, 1966; Goldstein, Davis, & Herman, 1975). Although many determinants have been varied in the studies demonstrating aggression escalation over trials, the recipients of aggression are always impersonalized.

Under dysfunctional feedback, subjects suddenly escalated punitiveness toward dehumanized performers to near maximum intensities. Although shocks typically failed to eliminate errors, increased punish-

ment was occasionally followed by a correct performance. It is conceivable that subjects were reinforced on those few occasions for escalated aggression. An explanation in terms of differential reinforcement, however, does not account for why subjects who also eventually succeeded through increased punishment of neutral and humanized performers were not reinforced into greater punitiveness. After the initial series of rising punishment produced a correct solution, on the subsequent (4th) error trial subjects escalated their punitiveness with dehumanized performers, did not change their level of punishment with neutral ones, and deescalated their punitiveness with humanized performers. This pattern of results, as well as that obtained under functional feedback, appears to be more consistent with differential disinhibition than with differential reinforcement processes.

The discussion thus far has focused on the disinhibiting power of labeling that divests people of human qualities. Of equal theoretical and social significance is the power of humanization in counteracting punitiveness. Subjects treated humanized performers in a mild fashion regardless of whether opportunities to aggress occured under personalized or diffused responsibility, with functional or dysfunctional feedback. These findings suggest that designations of others in terms that humanize them can serve as an effective corrective against aggression.

Lack of knowledge about the performers enhanced punitiveness, compared to conditions in which they were humanized. These findings confirm the common belief that impersonality is conducive to interpersonal aggression. Variations in responsibility, however, failed to produce differential amounts of aggression toward the neutral targets. It appears from the pattern of results that nonpersonalization substantially weakened self-restraints, thus reducing the capacity of mechanisms for obscuring responsibility to effect further changes in internalized control.

Conditions that dissociate or obfuscate the connection between actions and their effects do not automatically weaken self-restraints. Sequential analyses of aggression disclose that disinhibitory influences interact with aggressive experiences in determining levels of punitiveness. Neither shared responsibility nor dehumanization initially had much effect on punitive behavior. Aggression was abruptly heightened only after subjects had administered punishment under circumstances conducive to self-disinhibition. The uniformly low aggressiveness at the outset and the differential escalation of punitiveness under different feedback conditions indicate that the dehumanizing procedures produced their effects by divesting the victims of their humanness rather than through social sanctioning of punitive actions. In everyday life, of course, dehumanizing practices are almost invariably accompanied by active encouragement and reinforcement of maltreatment of those who are victimized.

Social learning theory provides a framework for analyzing different sources of disinhibition of aggression. Within this conceptualization, restraints over injurious conduct are largely governed by external, vicarious, and self-generated consequences. External and observed punishing consequences function as social inhibitors of aggressive responding, whereas self-evaluative reactions are the internal inhibitors. The operation of internal consequences is therefore best tested under conditions in which external consequences for injurious conduct are removed. It is in such situations that anticipatory self-reactions must serve as major deterrents of action.

The supplementary data on internal self-disinhibitory accompaniments of punitiveness attest to the numerous processes that may operate in weakening self-deterring consequences. When circumstances of personal responsibility and humanization made it difficult to avoid self-censure for injurious conduct, subjects disavowed the use of punitive measures and used predominantly weak shocks. By contrast, under sanctioning conditions, subjects resorted to a variety of self-disinhibiting devices and displayed a surprisingly high level of punitiveness. In studies of obedient aggression people are commanded to behave punitively. Here, participants escalated their punitiveness on their own.

Unlike dehumanization, diffusion of responsibility appears to increase aggression through social rather than personal disinhibition. The anonymity provided by collective action can effectively reduce restraints arising from fear of social censure. But since functionaries within a group are fully aware of how they behaved, a collective instrumentality may not necessarily lessen the sense of personal culpability. In the current study aggressiveness was measured under minimal threat of external censure. This factor might well account for why diffused responsibility emerged as a weaker disinhibitor than dehumanization. Evidence that sanctioning practices have differential internal concomitants attests to the value of distinguishing between social and personal sources of disinhibition.

The present study does not establish unequivocally the causal sequence between self-disinhibiting maneuvers and punitiveness. A detailed deterministic analysis would require an experimental design in which measures of self-disinhibition are taken prior to performance of aggression. The sanctioning process, however, is more likely a reciprocal than a simple unidirectional one. Disinhibiting social conditions facilitate expression of mildly reprehensible conduct, which, in turn, activates self-absolving responses that weaken restraints over more culpable behavior. If this is indeed the process, it would be best studied by sequential analysis of the interdependent changes in self-disinhibition and aggression.

REFERENCES

Bandura, A. *Aggression: A social learning analysis.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Buss, A. H. Instrumentality of aggression, feedback, and frustration as determinants of physical aggression. *Journal of Personality and Social Psychology,* 1966, **3,** 153–162.

Goldstein, J. H., Davis, R. W., & Herman, D. Escalation of aggression: Experimental studies. *Journal of Personality and Social Psychology,* 1975, **31,** 162–170.

Kelman, H. C. Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues,* 1973, **29,** 25–61.

Milgram, S. *Obedience to authority: An experimental view.* New York: Harper, 1974.

Sanford, N., & Comstock, C. *Sanctions for evil.* San Francisco: Jossey-Bass, 1971.

Zimbardo, P. G. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation, 1969.* Lincoln, Nebr.: University of Nebraska, 1969.

REFERENCE NOTE

1. Diener, E. *Deindividuation: Causes and characteristics.* Unpublished manuscript, University of Washington, 1974.