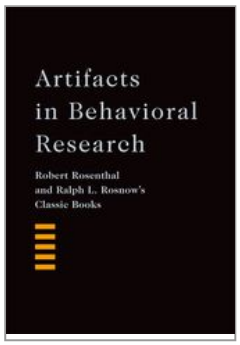


Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books
Robert Rosenthal and Ralph L. Rosnow

Print publication date: 2009
Print ISBN-13: 9780195385540
Published to Oxford Scholarship Online: September 2009
DOI: 10.1093/acprof:oso/9780195385540.001.0001

The Conditions and Consequences of Evaluation Apprehension

Milton J. Rosenberg

DOI:10.1093/acprof:oso/9780195385540.003.0007

Abstract and Keywords

On the assumption that many research subjects are apprehensive about how they will be evaluated, this chapter describes the circumstances in which this source of artifacts seems most and least apt to occur.

Keywords: evaluation apprehension, social desirability, attitude research, research subjects

Just as it keeps rats pressing levers, intermittent reinforcement keeps psychologists theorizing and neologizing. The best reinforcer I know is not the student's imitation of his professor's crotchets, nor is it a “successful replication” of one's experiment by another: instead it is to have some theoretical term that one has coined be often quoted and then to watch the quotation marks fade away as the term begins to enjoy some common usage. To put a phrase into the language (even if that language is spoken by only a few dozen others) confirms the sometimes faltering sense that one has really said something.

This seems to have begun to happen with the term “evaluation apprehension” which I first used in some unpublished documents in 1960–1961 and in an obscure article in 1963, and which I then explicated in a more visible one in 1965. Yet the diffusion of the term is not at all due to its being the key to some arcane and profound insight. Most experimental psychologists had long since come to the unhappy awareness that their subjects were prone to “faking it” and, particularly, to faking it “good.” But as a sort of contrast to Mark Twain's aphorism about people and the weather, the problem of self presentation in

experiments seemed to be something that virtually nobody was talking about[1] though a great deal could be done about it.

That the term “evaluation apprehension” has recently gained some currency must, then, be due to its helping to fill a need—the need, I should say, for experimental psychologists, social and otherwise, to come to terms with an obvious and fascinating source of trouble in their experimental procedures and rituals. In recent years my own sense of that need has led me beyond the initial conceptualization and into this possibly paradoxical commitment: to try to do systematic **(p.212)** experimentation on evaluation apprehension as a source of systematic bias in psychological experiments.

This chapter is intended as a rather loose, narrative account of the main directions taken and the major findings gleaned in that research program. All but the first of the studies to be described are previously unpublished, though they have been presented in various colloquia over the last two years. Some of these studies will be described in full detail in forthcoming articles; and it is my ambition to bring all this work, and related studies, into tight but expansive focus in an as yet unwritten book.

Evaluation Apprehension as Concept and Process

To begin, I had better not assume that the partial diffusion of the term “evaluation apprehension” has also spread abroad its full intended conceptual meaning. Thus what is called for, first, is a statement of definition. Then I shall need to outline my conception of how evaluation apprehension gets aroused and, after arousal, sometimes interacts with features of the experimental situation in ways that produce systematic biasing of experimental response data. Following these necessary preliminaries I shall turn, in the last portion of this introductory section, to some of the reasoning that lies behind the basic conceptualization and then we can begin to look at its research implications.

What, then, is the working conception of evaluation apprehension around which my recent research and this chapter are organized? The summary given in an earlier article (Rosenberg, 1965) is, I think, worth repeating here:

> It is proposed that the typical human subject approaches the typical psychological experiment with a preliminary expectation that the psychologist may undertake to evaluate his (the subject's) emotional adequacy, his mental health or lack of it. Members of the general public, including students in introductory psychology courses, have usually learned (despite our occasional efforts to persuade them otherwise) to attribute special abilities along these lines to those whose work is perceived as involving psychological interests and skills. Even when the subject is convinced that his adjustment is not being directly studied he is

likely to think that the experimenter is nevertheless bound to be sensitive to any behavior that bespeaks poor adjustment or immaturity.

In experiments the subject's initial suspicion that he may be exposing himself to evaluation will usually be confirmed or disconfirmed (as he perceives it) in the early stages of his encounter with the experimenter. Whenever it *is* confirmed, or to the extent that it is, the typical subject will be likely to experience *evaluation apprehension*; that is, an active, anxiety-toned concern that he win a positive evaluation from the experimenter, or at least that he provide no grounds for a negative one. Personality variables will have some bearing upon the extent to which this pattern of apprehension develops. But equally important are various aspects of the experimental design such as the experimenter's explanatory 'pitch,' the types of measures used, and the experimental manipulations themselves.

Such factors may operate with equal potency across all cells of an experiment; but we shall focus upon the more troublesome situation in which treatment differences between experimental groups make for differential arousal and confirmation of evaluation apprehension. The particular difficulty with this state of affairs is that subjects in groups experiencing comparatively high levels of evaluation apprehension will be more **(p.213)** prone than subjects in other groups to interpret the experimenter's instructions, explanations, and measures for what they may convey about the kinds of responses that will be considered healthy or unhealthy, mature or immature. In other words, they will develop *hypotheses* about how to win positive evaluation or to avoid negative evaluation. And usually the subjects in such an experimental group are enough alike in their perceptual reactions to the situation so that there will be considerable similarity in the hypotheses at which they separately arrive. This similarity may, in turn, operate to systematically influence experimental responding in ways that foster false confirmation of the experimenter's predictions.

What suggests this view of the secret side of the structured transaction between experimenter and subject? What, if anything, confirms the view?

One answer to the first of these questions concerns the modal theme that is usually encountered when one engages subjects in extended postexperimental discussion. Experienced experimenters who bother to talk to their subjects have all heard questions like these: "How did I do—were my responses (answers) normal?" "What were you really trying to find out, whether I'm some kind of neurotic?" "Did I react the same as most people do?" If one goes further in postexperimental inquiry, as I have regularly tried to do in recent years in my experimental work on attitude change (see Abelson *et al.*, 1968; Rosenberg *et al.*, 1960), and asks subjects to attempt a reconstruction of their private experience

of the experimental transaction, one often picks up another theme that I take to be quite significant. Subjects will report—sometimes with uncertainty and sometimes with great clarity—that they were burdened or preoccupied with the question “What is the real purpose of this experiment?” and that when some striking aspect of the experimental situation was revealed to them (whether through further instructions from the experimenter or, often, through first encounter with the instrument designed to elicit dependent variable measures) this generated a flash of “insight” about what the experimenter was “really trying to find out about me.” Though such “insights” are almost always incorrect they are of the sort that is capable of affecting the subject's further behavior in the experimental situation. The fact that such influence upon experimental responding has occurred is often the precise burden of the subject's remarks.

Thus, conversations with subjects (and also with graduate students and colleagues as they muse upon their memories from undergraduate years when they were the recruited subjects rather than the recruiters) have helped to shape the basic conception of the evaluation apprehension process. Yet another contributing influence has been the fact that experienced experimental social psychologists seem to share a certain basic style when engaged in professional “yesbutism.” What I mean is that when they suspect that someone else's data “are too neat—and the hypothesis can’t be *that* true” their first line of reinterpretation is usually to suggest that something about the experimental instructions or manipulations probably “aroused” the subjects in some unintended way or direction. Who has not heard reinterpretations similar to these illustrative ones? “The instructions probably made the subjects in the experimental group quite anxious about how they would be accepted and that, rather than the attributions of expertise as such, would be enough to make them conform to the views of the other group members.” Or: “By telling the subjects that prejudiced people are people who have repressed their hostility toward parents you are really making it necessary for them to show that the tolerance message influences **(p.214)** them; it isn’t ‘insight’ that accounts for the change, it's their need to get the psychologist's approval.”

The penchant for this sort of reinterpretation in terms of self-presentation dilemmas is widespread. Is this simply because it is a normative style in our profession; or has it become so because it reflects a persisting social psychological reality in the conduct of psychological research? Obviously, I would suggest that the latter is the case.

But if observations and speculations of the sort that I have indicated have helped to suggest the evaluation apprehension view, they do not, of course, in any way serve to confirm it. Confirmation can only be accomplished through further research. Thus, one of the basic aims of the experimental program that I and my various colleagues have been conducting has been to demonstrate that evaluation apprehension, once aroused, can significantly influence dependent

variable data. We have intended also to show that this influence often works not merely to increase “random error variance” but rather that it exerts *systematic* bias upon experimental responding; i.e., it “tilts” data distributions toward one or the other end of the response continuum and thus generates “significant” findings that happen also to be illusory ones.

We have had other purposes in mind as well—particularly to investigate the conditions under which evaluation apprehension is more or less likely to be aroused and, if aroused, more or less likely to induce systematic bias in dependent variable data.

I shall return to these matters later. Our first task is to review and discuss some “demonstration” studies. What they are intended to demonstrate is, simply, that when evaluation apprehension is aroused (and when it is coupled with the provision of cues that hint how the normal or “healthy” person would be likely to respond) this can induce systematic bias. Of course, it must be clearly understood that any demonstration that this can happen does not establish that it always or usually will happen. But there is no point in worrying about evaluation apprehension at all or in spending effort on trying to control and reduce it, unless we have first satisfied ourselves that it can actually be shown to exert biasing influence upon experimental responding.

Demonstration Through Altered Replication

There are at least two ways in which our basic point can be demonstrated. The one that will occupy us now is, in essence, a classic strategy. It is the one that is commonly employed when one suspects that the findings obtained in some reported “successful” experiment are in reality not due to the validity of the experimenter's hypothesis but to some unintended influence let loose by his poorly designed operations of manipulation or measurement. In this strategy one redesigns the suspected operations and repeats the experiment. If, despite the operational changes, the original findings are replicated one now has presumptive evidence that one's objections and doubts were ill taken; if meaningfully different, nonreplicative data are obtained, one has some claim (though it should not be over-indulged) to emit the prideful chortle: “I told you so” or “Thus do I refute Professor Berkeley.”

How does this bear upon our intention to confirm, by empirical demonstration, that unsuspected arousal of evaluation apprehension does sometimes generate false confirmations of hypotheses? Obviously, when we suspect that this has happened, **(p.215)** and where we have a speculative interpretation of how it happened, we may undertake an altered replication of the original study. The object would be to change those operations which we believe to have aroused evaluation apprehension and to have fostered the expectation that a certain way of responding would bring positive evaluation from the experimenter. If such an altered replication were to yield data that, as predicted, were quite different

from the findings of the original study this could be taken as evidence that our original concern over evaluation apprehension was neither excessive nor misplaced. In effect, such an outcome would be a demonstration, through the construct validation method, that evaluation apprehension can generate systematic bias in experimental data—though, of course, a single successful instance would hardly stand as an incontrovertibly definitive demonstration.

The strategy that I have just described was the first one employed in the demonstration phase of our inquiry into the evaluation apprehension phenomenon. The substantive area of concern was research in support of a basic hypothesis derived from cognitive dissonance theory. Some early experiments—most notably those by Festinger and Carlsmith (1959) and Cohen (in Brehm and Cohen, 1962) had supposedly confirmed the hypothesized relationship: when counterattitudinal advocacy (i.e., arguing in support of an attitude position opposite to one's own true conviction) is undertaken with little justification (e.g., for a small monetary reward) this will induce more attitude change in the advocate than when counterattitudinal advocacy is undertaken with strong justification (e.g., for a comparatively large monetary reward).

However, as many observers (among them Chapanis and Chapanis, 1964; Brown, 1962) have pointed out, dissonance studies of this type confront the subjects (particularly those in “low dissonance” experimental groups) with startling and ambiguous experiences and conditions. Agreeing with Chapanis and Chapanis that a likely consequence will be the arousal of “suspicion,” I thought it possible to be even more specific about the intervening, response-affecting, patterns of arousal that may occur with subjects in such studies. The particular case in point upon which we focussed was the well-known study by Cohen. In this experiment Yale undergraduates had been recruited to write essays in support of a position opposite to the one they actually held on a currently salient campus issue. The issue concerned “the actions of the New Haven police” in a recent campus riot. The undergraduates uniformly felt that the police had behaved badly. The essay they were requested to write was on the topic: “Why the actions of the New Haven police were justified.”

Having appeared at randomly chosen dormitory rooms the experimenter requested the potential subject to write such an essay and as an inducement offered a financial reward of either $.50, $1.00, $5.00 or $10.00. After the essay had been completed the experimenter asked the subject to fill out an attitude measure indicating how much he approved or disapproved of “the actions of the New Haven police.” As this measure was handed to him the subject was invited to take into account, if he so chose, the pro-police arguments he had just improvised in writing the counterattitudinal essay.

The prediction derived from dissonance theory was that the lesser magnitude of reward would generate a greater magnitude of dissonance and thus greater attitude change: that is, an inverse monotonic relationship was expected between the amount of money offered to elicit the counterattitudinal advocacy and the degree of attitude change toward the pro-police position. This prediction was apparently confirmed; **(p.216)** the $.50 reward group showed greatest attitude change in the pro-police direction, the $1.00 group next greatest change and the $5.00 and $10.00 reward groups did not differ from a control group which, without any prior counterattitudinal advocacy, had merely filled out an attitude scale concerning the question of whether “the actions of the New Haven police” were justified.

On the basis of attitude theory considerations that need not be reviewed here I thought that the opposite prediction made more sense: that the degree of attitude change would be a positive, rather than an inverse, function of the amount of monetary payment that was offered to elicit the counterattitudinal advocacy. Also it seemed likely to me that Cohen's results could be due to an unsuspected arousal of evaluation apprehension and a strong, but implicit, cueing which would have led most low dissonance (i.e., high reward) subjects to withhold evidence that they had influenced themselves in the pro-police direction.

Exactly what leads us toward this sort of interpretation of what really happened in this and similar early dissonance experiments on attitude change? The answer can best be conveyed by some extended quotations from the original article (Rosenberg, 1965) which posed the evaluation apprehension reinterpretation of the Cohen study and then went on to report the altered replication by which that reinterpretation was tested.

> It seems quite conceivable that in certain dissonance experiments the use of surprisingly large monetary rewards for eliciting counterattitudinal arguments may seem quite strange to the subject, may suggest that he is being treated disingenuously. This in turn is likely to confirm initial expectations that evaluation is somehow being undertaken. As a result the typical subject, once exposed to this manipulation, may be aroused to a comparatively high level of evaluation apprehension; and, guided by the figural fact that an excessive reward has been offered, he may be led to hypothesize that the experimental situation is one in which his autonomy, his honesty, his resoluteness in resisting a special kind of bribe, are being tested. Thus, given the patterning of their initial expectations and the routinized cultural meanings of some of the main features of the experimental situation, most low-dissonance subjects may come to reason somewhat as follows: “they probably want to see whether getting paid so

> much will affect my own attitude, whether it will influence me, whether I am the kind of person whose views can be changed by buying him off."
>
> The subject who has formulated such a subjective hypothesis about the real purpose of the experimental situation will be prone to resist giving evidence of attitude change: for to do so would, as he perceives it, convey something unattractive about himself, would lead to his being negatively evaluated by the experimenter. On the other hand, a similar hypothesis would be less likely to occur to the subject who is offered a smaller monetary reward and thus he would be less likely to resist giving evidence of attitude change.

On the basis of these speculative considerations I suggested, regarding Cohen's experiment, that

> in this study, as in others of similar design, the low-dissonance (high-reward) subjects would be more likely to suspect that the experimenter had some unrevealed purpose. The gross discrepancy between spending a few minutes writing an essay and the large sum offered, the fact that this large sum had not yet been delivered by the time the **(p.217)** subject was handed the attitude questionnaire, the fact that he was virtually invited to show that he had become more positive toward the New Haven police: all these could have served to engender suspicion and thus to arouse evaluation apprehension and negative affect toward the experimenter. Either or both of these motivating states could probably be most efficiently reduced by the subject refusing to show anything but fairly strong disapproval of the New Haven police; for the subject who had come to believe that his autonomy in the face of a monetary lure was being assessed, remaining 'anti-police' would demonstrate that he *had* autonomy; for the subject who perceived an indirect and disingenuous attempt to change his attitude and felt some reactive anger, holding fast to his original attitude could appear to be a relevant way of frustrating the experimenter. Furthermore, with each *step* of increase in reward we could expect an increase in the proportion of subjects who had been brought to a motivating level of evaluation apprehension or affect arousal.

But such a reinterpretation is merely another instance of applied "wise-guyism" unless one attempts to put it to a close and demanding further experimental test. To properly employ the altered replication strategy that I have already described, it was necessary to remove the posited evaluation apprehension dynamic, or at least to subdue it, and otherwise to hew as closely as possible to the design and operations of the original study.

How might the first of these desiderata best be implemented? The reinterpretation in terms of evaluation apprehension had an obvious

methodological implication. If the posited data biasing dynamic had actually occurred this had been made possible by the fact that the experimenter conducted both the dissonance arousal and subsequent attitude measurement. For evaluation apprehension and negative affect, if they had been aroused in the high reward subjects, would have been focused upon the experimenter; and it would have been either to avoid his negative evaluation or to frustrate him, or both, that the high reward subject would hold back (from the experimenter and possibly even from himself) any evidence that he had been influenced by the pro-police arguments that he had elaborated in the essay he had just completed.

Thus, quoting again from the original article (Rosenberg, 1965), these considerations led us toward the basic alteration employed in our replication:

> The most effective way then to eliminate the influence of the biasing factors would be to separate the dissonance arousal phase of the experiment from the attitude measurement phase. The experiment should be organized so that it appears to the subject to be two separate, unrelated studies, conducted by investigators who have little or no relationship with each other and who are pursuing different research interests. In such a situation the evaluation apprehension and negative affect that are focused upon the dissonance-arousing experimenter would probably be lessened and, more important, they would not govern the subject's responses to the attitude-measuring experimenter and to the information that he seeks from the subject.

We need not tarry here over the details of the staging of the two-experiment disguise. It will suffice to say that the disguise (judged by what the subjects said in quite probing postexperimental interviews) worked well, and that adaptations of it have since been used successfully both by others (e.g., Carlsmith, Collins, and Helmreich, 1966) and in my own continuing research program on attitude change (Rosenberg, 1968).

(p.218) Nor do we have to linger over precise descriptions of the instructions and measurement procedures used with the subjects. Except for changes required by our use of the two-experiment disguise all but two aspects of the procedure were identical with those used by Cohen in the original experiment. The two deviations from the original experiment were necessitated by the fact that it was conducted at Yale University and the altered replication at Ohio State University. Thus in the second study Yale undergraduates did not serve as subjects and the issue for counterattitudinal advocacy could not be the same one employed at Yale.

The issue that was used concerned the subjects’ attitudes toward a proposed ban upon any further participation by the O.S.U. football team in the Rose Bowl contest. Such a ban had been enacted, and later rescinded, by the faculty senate

during the previous year and extreme student opposition had been expressed through demonstrations and some riot-like group activity.

The experimental subjects wrote essays favoring the restoration of the Rose Bowl ban. The three experimental groups wrote the essays for promised rewards (delivered after completion of the essay) of $.50, $1.00 and $5.00 respectively. A control group merely took the dependent variable measure—a questionnaire on seven different campus issues, one of which was the Rose Bowl ban, while another dealt with the desirability of O.S.U. abandoning its policy of giving athletic scholarships.

On the Rose Bowl issue the Kruskal–Wallis one-way analysis of variance disclosed a significant relationship ($p < .001$) and inspection shows this to be of the positive, monotonic type: the larger the financial reward for counterattitudinal advocacy the greater the degree of attitude change (as estimated by comparison to the baseline attitude data provided by the control group). The $.50 and $1.00 groups showed greater favorability toward the Rose Bowl ban than did the control group ($p < .01$) and less favorability than the $5.00 group ($p < .02$).[2]

A similar overall finding ($p < .005$) was obtained on the athletic scholarship issue, though the differences between the groups were of lesser (but still significant) magnitude. This finding was also predicted, and is interpreted as evidence of some generalization of the main attitude change effect to a related, antiathletic issue.

Avoiding the lure of another theoretical area I have so far said nothing about the substantive issues in this experiment. And I shall resist the temptation to do so now—except to note that the positive relationship obtained between degree of reward for counterattitudinal advocacy and degree of resultant attitude change confirms the prediction drawn from my own affective-cognitive consistency theory and disconfirms the prediction derived from dissonance theory. But these issues of attitude theory need not be examined here. They are fully treated in some of my earlier publications (Rosenberg 1956; 1960a; 1960b; 1968) and in a published debate between myself and Aronson, the latter writing as an advocate of a sophisticated, modified version of dissonance theory (Aronson, 1966; Rosenberg, 1966).

Before I turn away completely from the whirlpool of attitude theory around which I have been skirting, I should like to make clear that the controversy concerning **(p.219)** counterattitudinal advocacy effects was not, by any means, fully resolved on the basis of this one study. Indeed, new issues have since been discovered in this by now middle-aged area of theoretical debate, experiment and counter-experiment. But the two-experiment disguise is now fairly standard in this particular research area. Also, the fact that under some conditions, at

least, the “incentive” rather than the “dissonance” relationship does obtain is now credited by the main participants in the persisting debate, though they continue to disagree (see the contributions of Janis, Carlsmith, Collins, Aronson, and Rosenberg in Abelson *et al.*, 1968) about the nature and provenance of those conditions.

Of greater pertinence at the moment are two points that have nothing to do with counterattitudinal advocacy as such, though they are grounded upon the Rose Bowl counterattitudinal advocacy study. The results of the altered replication can be taken as at least an indirect demonstration of the possibility that evaluation apprehension is capable of inducing systematic bias in experimental responding, and thus of generating undetected Type I or Type II errors (in the sense of invalid confirmations and disconfirmations of hypotheses). The second point is that such bias effects need not remain undetected, nor need they be left in the realm of the merely suspected. Variations of the altered replication strategy could probably be designed in most instances where an evaluation apprehension artifact is suspected to have induced systematic bias in the array of dependent variable data.

Inventiveness and care in the design of altered replications, and a readiness to resort to them frequently could probably do much to improve the reliability of the data that experimental psychologists collect to test hypotheses and in reaction to which they often develop new hypotheses.

Evaluation apprehension is by no means the only conceivable source of systematic biasing of data, nor is it an equally threatening possibility in all realms of psychological research. But whenever our experiments are heavy on surprise and whenever the experimenter's purposes are likely to seem mysterious to subjects (or whenever subjects are likely to sense disingenuousness in the experimenter's explanatory communication), we would do well to adopt the cautionary stance of obsessive concern over the evaluation apprehension problem. And having adopted this stance, we would do well to go beyond mere obsession or mere disputatiousness and get back to the laboratory where we can put our suspicions to test by conducting the relevantly redesigned altered replications.

Anyone who resorts to this strategy, however, had better be prepared to find himself at the receiving end of the ironic justice process. For the criticized and their partisans can reverse the tactic on the aspiring critic. An altered replication designed to remove a suspected evaluation apprehension contaminant from some previously reported experiment can, in itself, be interpreted as having been contaminated by evaluation apprehension or by some other biasing force (e.g., experimenter expectancy, demand characteristics, subject presensitization).

The mind reels, and one's strength does quaver a bit, at the conceivable prospect of an infinite regress in which a study is designed to take systematic bias out of a study that was designed to take systematic bias out of a study that was … but in all likelihood, even fully consensual devotion to the expunging of evaluation apprehension effects will stop far short of such total, unlimited doubt. At some point it should become clear that particular experimental paradigms and particular substantive areas of experimental inquiry have been pretty thoroughly "debugged." And meanwhile, **(p.220)** whatever temporary disruption, confusion, and outraged pride may result, the ultimate outcome can be trusted to be beneficial—not only in that it will probably elevate the trustworthiness of data in the contested area, but also because there is nothing more restorative of the scientific temper than an occasional encounter with the hard, intractable fact that one has made, and remains capable of making, mistakes.

Demonstration Through Manipulated Arousal and Cueing

The construct validation strategy, useful as it has been in theory testing generally and in our own research program, is really a version of the Platonic analogy of the cave. The shadows that are projected across the wall (i.e., our data) denote that something is passing between us and the sun—but we are still tantalizingly out of direct contact with its substance.

Thus, the foregoing study and others of similar design, though they seem to confirm the reality of the evaluation apprehension dynamic, do not bring us into direct contact with it. To look more closely at that process it appeared necessary to *arouse* it, rather than reduce it as was done in the Rose Bowl study.

My first effort in this direction was undertaken in an experiment in which I had valuable collaboration from Dr. Raymond Mulry who was, at that time, one of my graduate students. Our working plan was simple, perhaps even crude.

Through a printed "Background Information Sheet" we conveyed to two separate experimental groups the following points: They were about to participate in a study of social perception, in which they were to judge how much they liked or disliked various pictured persons. Past research by others, they were informed, had shown that liking-disliking reactions to strangers were correlated with personality, particularly with whether the rater was psychologically "mature" or "immature." To one experimental group it was disclosed further that the main burden of the past research (various invented journal articles were cited) was that psychologically mature and healthy people show greater liking for strangers than do immature people. To a second experimental group the printed communication conveyed the opposite: past research had shown that it was psychologically immature and comparatively unhealthy persons who showed greater liking for strangers.

Beyond this crucial point of difference the two forms of the manipulative communication again converged. All of the past research, it was asserted, had been done with subjects in face-to-face contact with real strangers. Would the same relationship between psychological maturity and liking hold for mere photographs of unknown people? This, the subjects were told, was a question that we planned to pursue in further research. But first it was necessary to "standardize" a set of photographs; to determine how much, on the average, they elicited liking or disliking reactions. Thus, in the present study, according to the concluding paragraph of the Background Information Sheet, we were not testing the personalities of our subjects; rather, we were simply establishing normative data against which we would later compare the liking-disliking ratings elicited from subjects whose personality qualities had already been assessed.

The simplicity and directness of this manipulation make clear its intended purpose. We were attempting to arouse evaluation apprehension by confirming, for our **(p.221)** subjects, the sort of expectancy that subjects often bring to experiments; namely, in the present instance, that as researchers we were ordinarily interested, among other things, in personality assessment. And furthermore we were cueing our subjects about what past research had shown (and thus about the likely content of our own expectations) concerning the ways in which "mature" and "immature" people tended to react to strangers. Why did we add that we had no idea whether the same relationship would hold with pictures as with reactions to directly encountered real persons? Partly to enhance the general credibility of our communication and partly to reduce what might otherwise be a too overwhelming influence upon the individual's judgments of the pictures. Also this made the present study somewhat more comparable to many others in which the common strategy (whatever has gone before) is to provide overt reassurance that the subjects' personalities are not being scrutinized.

Apart from the two experimental groups (both aroused to evaluation apprehension and cued either toward liking or disliking responses respectively) we also set up a control group. These subjects received a brief neutral communication which did nothing to arouse evaluation apprehension or to provide directional cueing. The data from this group served, then, as the baseline against which we could assess the significance of the deflections, toward the liking and disliking ends of the scale, of the experimental subjects' self-reported judgments.

I have already suggested that the Rose Bowl altered replication study could be interpreted as providing indirect evidence that evaluation apprehension can contaminate experimental data but not that it always or usually will. The same stricture is all the more applicable in limiting the meaning of the present study. Only a failure to find significant differences between the mean liking ratings of

the three groups could be taken as definitive; for this would mean that, even under optimum conditions, evaluation apprehension does not get aroused or, if aroused, does not affect experimental data.

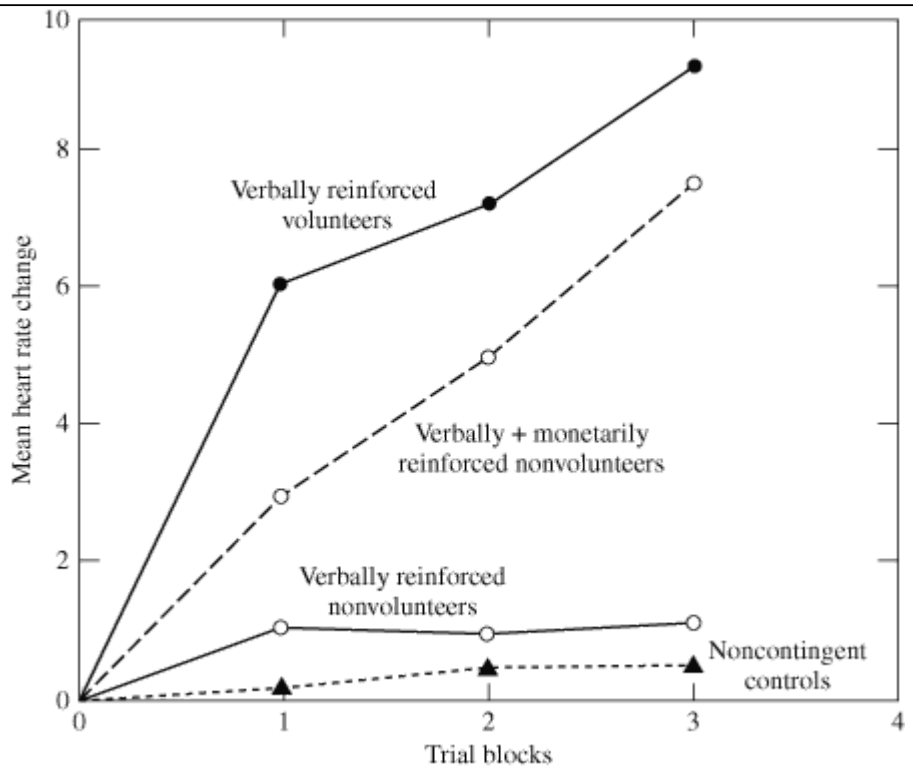
But if significant differences were obtained just what would they tell us? Merely that the data biasing dynamic that we suspect to be unintentionally induced in certain kinds of experiments can be intentionally induced by rather direct manipulation. In essence, then, we were giving ourselves a chance to increase the pertinence of the null hypothesis or provisionally to reject it.

If the data seemed to allow the latter (i.e., if they showed that, at least by intentional amplification, evaluation apprehension can be made to affect experimental responding) we would also be in a position to carry out inquiry a few steps further. We would then be able to ask what kinds of people, situational definitions, and experimental tasks tend to facilitate or diminish the operation of the process in which data are systematically biased under the influence of evaluation apprehension.

These foregoing considerations set the context in which we can now proceed to discuss the findings of the first evaluation apprehension manipulative study; and they are equally relevant to the various other studies that followed it and employed the same basic design paradigm. Obviously I would have no claim to write this chapter if the results of this first study, and of the others that followed upon it, had failed to render the null hypothesis improbable. Thus there will be little surprise in the disclosure that in the first of these studies a large and significant difference was obtained between the two experimental groups.

For the 12 pictures of male faces (each rated on a 21-point like-dislike scale ranging from +10 to −10) the algebraic sums of each subject's judgments were **(p.222)**

Table 7–1 Like-Dislike Mean Sums for Groups; and Probabilities of Differences Between Groups



computed. The means of these scores for separate groups of male and female subjects and the probabilities of the differences between various pairs of means are displayed in Table 7–1

For male subjects in the experimental group that was cued to think that mature people like strangers, the mean algebraic sum of the ratings was +23.20. For the male subjects cued to think that immature people show greater liking for strangers, the mean algebraic sum was −8.65. The significance of the difference between these groups (computed by the Mann–Whitney Rank Sum statistic, as are most of the other simple differences between groups that are reported in this chapter), was clearly established ($p < .0001$).

However, as reference to Table 7–1 makes clear, we also encountered an interesting complication. The disliking treatment did not, in fact, exert a significant influence upon the male subjects who received it. This is apparent from the fact that the picture ratings from the unmanipulated control group are, on the average, just as negative (the mean is −11.25) as those from the disliking group and, of course, there is no significant difference between these groups.

Does this signify that the disliking cueing that we employed was simply not credible? Or that, though credible, subjects could not bring themselves to behave in opposition to the normative standard (at least with typical middle-class Americans) that whatever our private disposition may be, strangers are to be approached with external affability?

Either of these interpretations would be plausible if it were not for various other available findings. The most striking is that with the separate groups of female subjects (judging pictures of males, it should be remembered) the disliking treatment does influence the picture ratings and they are as deviant from the control mean in the negative direction ($p < .01$) as the mean for the liking treatment is in the positive direction ($p < .03$).

Furthermore, even with the male subjects, there is evidence suggesting that a personality-linked variable has mediated the influence of the disliking treatment. For all subjects we had available their scores on the Marlowe–Crowne (1961) Social **(p.223)** Desirability (SD) Scale which had been administered sometime before the present experiment was conducted. When the male subjects are split into high and low halves on the Social Desirability Scale distribution we find that a trend in the predicted direction is visible between the High SD subjects in the control and disliking groups respectively. But with the Low SD subjects that trend is reversed and approaches significance ($p < .10$) in the counterhypothesis direction. If the latter group had shown a trend no stronger than that obtained from the former the overall finding would have supported the predicted relationship at an acceptable level of statistical significance. Thus it is the Low SD males who, needing less approval from others (and, we may assume, from the experimenter), are not willing to respond against the normative grain and win a judgment of normality by representing themselves as disliking certain strangers more than they otherwise would.

There is a glimmer of a paradox in these last data; for generally one would expect people with a high need for approval (the High SD scorers) also to show a more persisting proclivity for representing themselves as positively disposed toward random others. In fact, passing beyond the data from the disliking condition, we find that the social desirability factor did exert the expected influence within the experimental group that was cued to believe that liking strangers is a sign of maturity. The High SD male subjects in that condition give more extreme liking scores (the mean sum of their picture ratings is +34.77) than do the comparable Low SD subjects (whose mean score is +13.72). The difference between these groups is significant ($p < .03$).

Despite the few tantalizing ambiguities that I have discussed above, the overall import of this first manipulative study seemed quite clear; with intentional arousal of evaluation apprehension the subsequent directional cueing does “take”—that is, it influences the subjects’ experimental responding. Postexperimental inquiry indicated to our tentative satisfaction that these results were not due to any easy comprehension of our unrevealed purposes. The subjects usually insisted that the preliminary material that they read concerning “earlier studies” on reactions to strangers had not particularly influenced them. I do not take these reports as veridical; but neither do I think that they are due to a simple intention to deceive the experimenter. From

interviewing conducted after data collection in this study and others, I have formed the impression that subjects will usually obscure from themselves the extent to which they regulate their responding so as to win favorable judgments from the experimenter. And though I cannot anchor the following judgment on a base of hard data I would hazard the psychologically obvious interpretation that this sort of motivated inattention is due to the typical subject's need to conserve a positive image of himself even as he half-knowingly seeks to make a positive impression upon the experimenter.

Upon completion of this first demonstration study it would have been possible to plunge directly into studies concerned with variables that facilitate or suppress the evaluation apprehension-data biasing process. But the rating of pictures for their likability is a rather special sort of task and, as we have seen, certain complications did arise on the dislike cueing side of the experiment. To satisfy ourselves that the process under study was a fairly general one, it seemed necessary to adapt the basic experimental paradigm to some other and quite different sorts of experimental tasks.

Two further studies of this type were successfully carried out with male undergraduate subjects at Dartmouth College. I shall describe them somewhat more briefly than the preceding study, since they are useful here only in adding some empirical **(p.224)** weight to an assertion that I have already registered more than once: that is, that evaluation apprehension combined with some hints about how “normal people” react (and thus implying something about how the experimenter's approval can be obtained) does exert systematic biasing influence upon experimental responding.

In one of these additional studies I was joined by two coinvestigators, Philip Corsi (who developed the basic experimental design and operations) and Edward Holmes, both of whom were advanced undergraduate students in the Psychology Department of Dartmouth College. We used an extremely simple task: the subject taps upon a key with his right and left index fingers for six separate ten-second intervals, half of these with one finger and half with the other. The number of taps is automatically registered on a Veeder Root meter. Normally there is a considerable discrepancy between the performance of the two fingers, the index finger of the dominant hand producing more taps than that of the nondominant hand.

As the subject entered the experimental room he was asked by the experimenter: “Did you take the general abilities test and the personality inventory during freshman week?” The purpose of this query was to stir some initial prompting toward evaluation apprehension.

Following the administration of three brief abilities tests focused on verbal and symbolic skills (and intended to rouse the subjects’ interest in their

performances) the experimenter proceeded to give a memorized, verbal explanation of the finger-tapping task. For the control subjects (eleven right handed undergraduates) this consisted only of a simple description of the task. Working with only this information these subjects did, indeed, produce more taps with their right than with their left index fingers. The mean difference between the sum of right index taps minus the sum of left index taps was 22.45 for this group.

With an experimental group of the same number of right handed subjects the preliminary communication contained some additional information designed to heighten evaluation apprehension and to turn it in a particular response direction. Thus they were told that recent research with graduate students at Yale and at the University of Michigan had turned up the surprising finding that the number of taps with the nondominant index finger was virtually equal to the number with the dominant index finger. The clear implication was that people with higher intelligence (or perhaps of higher educational attainment) performed differently than did other, more ordinary, persons.

The result was striking. The mean difference between the sums of right and left index finger taps was only 10.73 and this was quite significantly ($p < .005$) different from the comparable score of 22.45 obtained with the control group. A clear hint about the relation between performance on the experimental task and the likely evaluation that the experimenter would draw from the subject's performance had produced a “transcendence effect.” The experimental subjects performed far more efficiently with the left index finger than subjects (both our own control subjects and those in many other studies) ordinarily do.

One further finding from this study is of particular interest. The control and experimental data just described were obtained under a “full feedback” condition; that is, the meter registering the cumulative number of taps was visible to each subject as he went through the separate right and left finger tapping trials. But on six other trials with these same subjects we used a “partial feedback” condition in which the meter was exposed only after the ten-second tapping interval had been **(p.225)** completed. (For half of the subjects the partial feedback trials, and for the other half the full feedback trials, were run first.) Under the partial feedback condition the same sort of difference is obtained between the experimental and control groups but it is of lesser magnitude ($p < .03$) and the difference between the means for the two groups is correspondingly smaller. It seems clear, then, that the continuous availability of information about the quality of his performance will assist the subject in guiding that performance in the direction that he expects will bring him a favorable evaluation.

A more ambitious study was undertaken at the same time. This one followed our original experimental paradigm (as used in the picture-rating study) in all its basic details, except for using a quite different sort of experimental task.

In two basic experimental conditions and in a separate control condition the subjects (20 in each group) worked with a booklet of addition problems, each page of which had the same number of total digits. The booklet was far too lengthy for completion and the subject was free to stop whenever he chose to do so. Each fourth page in the booklet presented some scales from the Nowlis (1965) Mood Adjective Check List which the subject was to fill out before he went on to the next page of addition problems.

Again the experiment began with the subject reading a Background Information Sheet. This document, modeled after the one used in the original picture rating study, explained that past research by other investigators had revealed a relationship between personality competence and the amount of efficiency shown, and pleasure experienced, in executing dull, routine tasks. One form of the information sheet went on to report that the clear finding from many studies (again fictional citations were given) was that the mature, psychologically healthy person experienced more pleasure and was more efficient at such tasks than was the immature person.

There followed a paragraph explaining the probable psychological basis of this finding (i.e., emotionally healthy people are less prone to distraction and enjoy challenges more than emotionally handicapped persons). The text then went on to explain that all of the previous studies had dealt with “motoric and manual” tasks and there was no proof that the same relationship would hold true with clerical or other “paper and pencil” tasks. This, it was explained, was the question that would be addressed in later research that the experimenters planned, but first it was necessary to develop a standard task and to determine what the typical levels of performance would be. Thus the subject was about to participate in a pilot study focused not upon his personality but rather upon determining average performance levels for various kinds of clerical tasks—and at present the task whose utility was to be explored was arithmetic addition.

The content of the Background Information Sheet used with the other experimental group was, of course, exactly parallel in form but opposite in content. It reported that past data with routine motoric performances had shown that psychologically healthy and mature persons were less efficient at carrying out such tasks and got less pleasure from them than did psychologically immature persons. Again a brief psychological explanation of the basis for this finding was offered and this was followed by exactly the same further comments that were used with the first version of the manipulative communication.

Obviously this manipulation, just as the one used in the picture rating study, was likely to exert a strong force toward arousal of evaluation apprehension and at the **(p.226)** same time would provide unambiguous cues that could be used to regulate experimental responding so as to maximize the chance that one would be judged “normal” by the psychological experimenter. However, despite the directness of the manipulation, postexperimental questionnaires and postexperimental interviewing with a sample of the subjects revealed little acknowledged penetration of the purpose of the experiment. Subjects did show accurate recall for the content of the Background Information Sheet but usually they insisted that they did not feel that their personalities were being scrutinized; instead, they reported that they had simply worked on the problems until they got bored or fatigued.

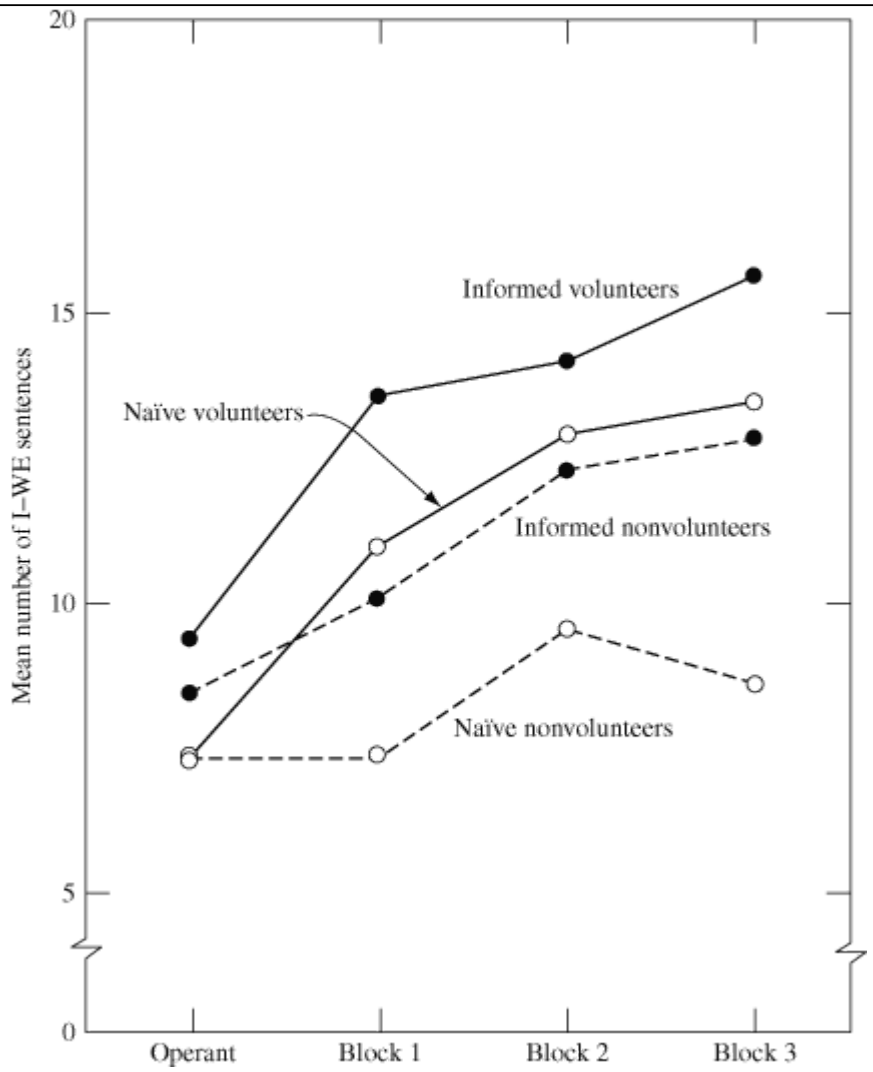
But, though evaluation apprehension or concern for performing in the “normal” way typically was not acknowledged to the interviewer (and, possibly, not fully acknowledged to the self) it did clearly influence the actual performances of the subjects on the arithmetic addition task.

That this is the case is clear from the findings presented in Table 7–2. The means reported there are from the two experimental groups. The table also displays means from a control group whose members worked on the addition problems without any previous arousal of evaluation apprehension, thus establishing the baseline performance levels against which the experimental groups can be judged.

It is clear that the two experimental groups differ from one another in the predicted direction. On the average, the subjects who were led to believe that mature people tend to be comparatively dysphoric and inept toward routine tasks completed ten less addition problems than did the opposite experimental group ($p < .03$). Similarly, they correctly solved eleven fewer problems than did the other experimental group ($p < .003$).

However, examination of Table 7–2 will quickly suggest that the treatment emphasizing that mature people do not perform well on dull tasks was a far stronger influence upon the performance than was the opposite treatment. While the former experimental group differs significantly from the control group on both the number

Table 7–2 Performance Means for Groups; and Probabilities of Differences Between Groups



(p.227) of problems completed and the number solved, the latter group does not. An obvious proposition virtually suggests itself and it is one that may well deserve an important place in the general theory of evaluation apprehension processes that is emerging as we pursue our experimental program. Simply stated it is this: cues suggesting a response pattern that is likely to bring approval from the experimenter will have stronger influence upon actual responding when that pattern is also less effortful in execution.

There is, of course, an alternative interpretation that is quite plausible in the present instance: the “displeasure and inefficiency” version of the Background Information Sheet may simply have been more credible, more in accord with the initial expectations of the subjects. After all it does seem likely, at a common sense level, that only “odd” people will enjoy routine, repetitious tasks. But this interpretation is weakened by the fact that, by a somewhat subtler analysis, we find that the subjects in the opposite experimental group were also influenced in their performances on the addition task. Apart from comparing the groups on the total number of problems completed and correctly completed we went on to compute for each subject an index based upon the ratio between these two separate scores. Dividing the number of problems correctly solved by the

number completed we obtain a meaningful estimate of the quality of the subject's performance in relation to the scope of that performance.

As Table 7–2 shows, on this index the experimental groups are again significantly different ($p < .03$) from one another in the predicted direction. However in this instance the group cued to the expectation that mature people perform poorly does not differ from the control group. Though they are completing fewer problems, the percent of these that are correctly completed is the same as with the control group. But the opposite experimental group does differ from the control group. On the average the subjects in this experimental group complete a few less problems and solve a few more. In consequence the difference between the control group and the “pleasure and efficiency” cued group attains significance ($p < .05$). Clearly these experimental subjects have been putting somewhat more effort into the task; they have been concentrating more closely on the truly tiresome task of adding columns of digits, and in consequence they have attained somewhat greater accuracy.

Equally interesting and meaningful in the light of the finding just reported, are some additional findings obtained with the Mood Adjective Check List.

Avoiding the task of describing the scoring or analytical procedures, I shall content myself here with simply reporting that on some of the subscales of this instrument we find the experimental groups differing either significantly or at borderline levels from one another and, in some instances, from the control group as well. The subjects in the group cued to think that normal people enjoy routine tasks characterize themselves as feeling less dysphoric while doing the addition problems than those cued to think that normal people do not enjoy them. And these characterizations tend to persist across the various intervals (every fourth page in the addition problem booklet) at which the subjects were required to report upon their mood states.

I have reviewed three manipulative studies each of which successfully demonstrated our basic point: that systematic bias in experimental responding can be produced through the arousal of evaluation apprehension and the cueing of particular response patterns as likely to foster positive evaluation.

(p.228) However, two defects of this group of studies are apparent and they should be noted here. The first is, simply, that they do not cover as broad a range of experimental tasks as could be desired. Suspecting that virtually any type of experimental performance could be systematically biased by the evaluation apprehension process, we might well have gone on to similar studies in such diverse areas as conditioning and other learning phenomena, psychophysical judgment, impression formation, concept formation, and many other areas. Particularly I should have liked to test the proposition that the degree of attitude change shown by subjects (in their responses to

questionnaires administered after a persuasive communication has been received) will be influenced by the prior suggestion that attitude change reflects a mature quality of “openmindedness” or an immature quality of “inconstancy.”

Further work along some of these lines is planned. But, happily, the task of demonstrating the broad relevance of evaluation apprehension as a data biasing process has now been taken up by some other investigators. By rather different experimental techniques than those that I have employed, Silverman (1968; Silverman and Regula, 1968) has been providing some evidence that could be easily fitted to the general picture developed here. And Sigall, Aronson, and Van Hoose (1968) have recently reported a study in which subjects are exposed to evaluation apprehension cueing and also to the “demand characteristic” of the experimenter's expectancy about their performances, the latter ostensibly based upon the scientific hypothesis he is testing. With subjects for whom both forces converge, suggesting that a certain mode of responding will prove the experimenter's hypothesis and also make the subject appear a competent and adequate personality, strong influence on experimental responding is obtained. With another group of subjects these forces are made to diverge, so that the subject must violate the experimenter's hypothesis if, as he sees it, he is to appear competent and psychologically adequate. The typical subject yields to the latter rather than the former force. Thus, even with a strong demand characteristic opposing it, the evaluation apprehension dynamic is found to exert a statistically significant influence upon the experimental responding of the subjects.

Interesting and heartening as such studies are, much more experimental exploration will be required before we can take as established the claim that the systematic biasing of data through the evaluation apprehension dynamic is a *general* phenomenon, one that can be made to occur over the vast range of response dimensions with which modern experimental psychology is concerned. My expectation is that such a program of “parametric” exploratory studies would in fact reveal considerable generality of this sort. At the same time it would probably disclose that certain types of experimental responding are more prone, and others more resistant, to this type of systematic biasing. Indeed, I think it likely that one would also find that, within a given behavioral realm, certain directions of responding are more easily affected by evaluation apprehension pressures than are others. This has already become apparent through our discoveries that liking of strangers or inefficient performance on routine tasks are more readily inducible response patterns than are their opposites.

A momentary lapse into unrestrained programmatic fantasy (an easy indulgence if one puts aside the fact that someone must actually undertake the vast labors that are contemplated) suggests the desirability of constructing, through empirical **(p.229)** techniques, a sort of evaluation apprehension atlas of

response dimensions. The hundreds of types of elicited behaviors which now serve as dependent variables in psychological research could be separately submitted to evaluation apprehension cueing of the sort employed in our demonstration experiments. The degree of influenceability of each particular response pattern (and of separate response directions) could then be assessed. Ideally, this would need to be done with systematic variation in types of subjects, types of evaluation apprehension arousal, and types of directional cueing. The result would probably have high payoff in terms of increasing our ability to do uncontaminated, bias-free research—or at least to come closer in approaching that utopian state of affairs.

I said earlier that I perceive two main defects in the group of demonstration studies described here. The first, as discussed above, can be handled only by doing more demonstration (and parametric exploration) studies over the broad range of common dependent variables employed in psychological research. The second defect is one that bears upon the *way* in which such further studies might be conducted. What I have in mind is the fact that in all of the foregoing studies the manipulation had two separate components: evaluation apprehension was aroused or heightened by our telling the subjects, in a fairly direct way, that the responses they were about to make would have some revelatory significance concerning their own personalities; then, in a separate and subsequent portion of the communication, some hints (usually rather strong ones) were given concerning the response differences that might be expected as between normal and abnormal or “mature” and “immature” persons.

Are both portions of the induction required? For that matter, can subtler inductions be used without the loss of the systematic bias effects? These questions point up a basic limit in the group of demonstration studies so far reviewed: namely, that they have not featured enough cross experiment systematic variation in ways of inducing evaluation apprehension. When such variation is attempted what are we likely to find? Both through speculative rumination and also in the light of some of the data from additional studies that I shall shortly discuss, I am willing to hazard some informed guesses. The first is that the evaluation apprehension biasing effect does not depend upon providing the subjects with an initial statement defining the experimenter as one who is interested in the study of personality or who is otherwise sensitive to the personality revealing implications of the data he is collecting. When this is done it probably does boost the data biasing process, but the same sort of process is likely to be set in motion merely by providing some cues suggesting that one mode of responding as compared to another is more “normal” or “competent” or “mature.” The latter strategy was the one employed by Sigall, Aronson, and Van Hoose (1968) and it was sufficient to induce significant systematic bias.

However, what of the situation in which no direct cueing toward the “normal” pattern of response is provided? Surely this is the typical state of affairs in experiments in which evaluation apprehension is an inadvertent rather than an intended influence upon subjects’ responding. Theoretical analysis has rather persuaded me (and some studies, reported later, on the mediation of the experimenter expectancy effect have turned persuasion toward conviction) of this basic point: arousing the subject to the general expectancy that his personality competence will be available for judgment by the psychological experimenter sets him examining salient aspects **(p.230)** of the situation for what they might reveal about “the way a normal person would respond.” In other words, when a general state of evaluation apprehension has been aroused by intention (as we attempt to do with the first portion of the Background Information Sheet communication) or unintentionally, direct cueing of the normality-revealing behavioral model is not required. Subtler hints will be picked up and private hypotheses will be formulated by the subjects—and, to the extent that the separate subjects attend to the same hints and draw the same interpretations, systematic biasing of response data will be likely to occur. An implied methodological corrective is lurking in these last comments. Though I shall return to it at a later point it deserves bold preliminary iteration here: techniques for reducing any stirrings of evaluation apprehension that subjects bring into an experimental situation, for disconfirming any initial concern that their psychological adequacy or inadequacy will be open to judgment, are bound to improve the trustworthiness of the data collected in that experiment.

In setting the stage for the foregoing discussion of the limits of our manipulation techniques I asked: can subtler inductions be used without loss of the systematic bias effects? By “subtler” I mean communications which do the work of our Background Information Sheet (i.e., arousing general evaluation apprehension, or cueing the subject in a particular response direction, or both) far less explicitly, with more “natural” indirectness. I am fairly sure that the answer is yes—that such subtler manipulation will induce systematic bias in experimental responding. To this purpose, a number of further demonstration studies have been planned, but not yet executed. If their results are successful they will give us a stronger empirical basis than we have yet established, for the claim that the evaluation apprehension dynamic does often operate where it is usually unsuspected: for example, in experiments undertaken to test substantive issues and hypotheses relevant to important matters of psychological theory. But, if this point is not yet fully established through our demonstration efforts I must, nevertheless, confess that the studies we have already completed (both those described above and those that follow in the next sections) have considerably strengthened my own original suspicion, namely: Evaluation apprehension does contaminate a fair portion of the experimental work now being conducted over the broad range from social psychology to psychophysics.

To be sure, as I make this declaration I am mindful of various limiting considerations: some of my readers will certainly think it a considerable leap beyond the data—and they are right; but scientific inquiry, like other more muscular pursuits, is advanced by the judicious use of audacity. Also I am mindful that this sort of *j'accuse*, as it concerns any experiment in which one suspects that evaluation apprehension has distorted the data, cannot be sustained by a hundred, let alone three, demonstration experiments; instead the logic of inquiry forces us back to the necessity for undertaking carefully designed altered replication studies.

However, the more we can learn about evaluation apprehension through intentionally arousing it, the better equipped we will be to search it out and bring it to heel through the altered replication strategy. Thus, in further research I and my colleagues have gone on creating evaluation apprehension and expanding our inquiry to encompass subsidiary variables which may work to heighten or reduce its influence upon experimental responding. I shall now turn to a review and discussion of some of these further studies.

(p.231) Variables Influencing the Evaluation Apprehension Process

Though all of our original demonstration studies showed clear main effects, there was a fair degree of intersubject variance within, as well as between, conditions. This suggested that uncontrolled factors relating to the subjects' personalities, their sensitivities to aspects of the situation, and their patterns of past experience as subjects might be affecting how much evaluation apprehension they felt and how they were acting to reduce it.

Clearly a host of variables might be found to influence the evaluation apprehension data biasing process—and the direction of such influence might be either to facilitate or subdue the overall operation of the process. I found it useful to conceive such "booster" and "suppressor" variables as falling into five major categories. They could be: personality attributes (or overall personality patterns) of the subject; aspects of the subject's recent, preexperimental experience; aspects or attributes of the experimenter; or of the experimental setting; or of the experimental task. We need not think of this taxonomy as the most logical of all possible ones, nor need we assert that it would incorporate all relevant variables. Its main value was, simply, that it was enough to get us started.

But we are just *barely* started on this line of inquiry. While many relevant variables are easily conceivable, only four major ones have been investigated in specific experiments. The results which I shall shortly present have been quite informative both in confirming our initial hypotheses and also, in two of these studies, by disclosing certain more complex interactions which have, in turn, suggested some new lines of theoretical speculation.

The four variables upon which this work has so far focused are: the need for approval as an attribute of personality, the salience of the “clinical” orientation as an attribute of the experimenter or of the experimental setting, the experimenter's “gate-keeper” power over the subject, and the ambiguity of the experimental stimulus materials.

I have already reported that the need for approval (as indexed by scores on the Social Desirability Scale) seemed to play a response affecting role in the first of our demonstration studies. The same appeared to be true in the study reported above in which “efficient” and “inefficient” performance on routine addition problems were separately cued as reflecting personality competence. In this instance we found that under the cueing treatment suggesting that bored and inefficient performance on routine tasks is a correlate of “maturity,” subjects with high (upper 50 per cent) Social Desirability scores did not differ from the control group. However those with low Social Desirability scores completed and correctly solved significantly fewer problems. This finding though unexpected is a meaningful one, in the sense that hindsight suggests some hypotheses that deserve further experimental testing: perhaps those with strong approval need have in the past been more concerned with (and, thus, more rewarded for) seeking social acceptance through overt display of competence; or perhaps, because of their greater general concern with techniques of ingratiation, they find the cueing communication less credible than do the subjects who are comparatively low in the need for social approval.

The personality attribute of need for approval, as indexed by SD scores, was studied within the context of our original demonstration studies. Our more directly **(p.232)** focused work on booster and suppressor effects began with two studies dealing with another variable: the definition of the experimenter or the experimental situation as high or comparatively low in clinical interest. In effect this can be viewed as a simple “additive” variable: that is, another direct force making for evaluation apprehension in the subject and thus, in our typical experimental situation, adding to the weight of forces that have already served to create that motivating state.

In the first of our studies on the clinical interest variable, the same sort of Background Information Sheet that was used in our earlier studies was read by all experimental subjects. Once again its first portion was designed to rouse general evaluation apprehension while its second portion conveyed the clear hint for one group that mature people are comparatively high, on liking for strangers, and for the other group, on disliking for strangers. Control subjects received no such preliminary communication. So far this study is essentially a replication of our original demonstration experiment. The additional variable was introduced through a few memorized sentences which the experimenter addressed to the subject.

The experimenters (senior students in an advanced experimental psychology course at Dartmouth College) each ran six subjects: two who had received the liking cueing, two the disliking cueing, and two control subjects. To one subject from each of these three categories the experimenter represented himself as having a “clinical” orientation. He did this by saying, just before presenting the instructions for the picture rating task, that he was quite fascinated by the experiment as, indeed, he was by psychology generally. “I guess,” he continued, “that's because I’m always thinking about what makes people tick. That's why I’m hoping to go into psychiatry after I finish medical school.”

In the opposite “nonclinical” role that he played with his three other subjects each experimenter said that he did not particularly see the importance of the present experiment. He continued: “For that matter I’m not sure what I’m doing in this course but they said, at the School of Engineering, that I had to take it.”

The purpose of this manipulation was, then, simply to convey that the experimenter either had or lacked interest in understanding and judging the personalities of others. At the same time, for the “clinical” role, it was clear that the experimenter was not yet professionally trained or skilled in this direction. As he alternated between these two roles the experimenter ran his subjects without any knowledge of whether they were in the control group or in the groups that had been respectively cued to the suggestion that liking or disliking for strangers was characteristic of psychologically mature persons.

One hundred and fifty subjects gave their liking-disliking ratings for 15 photographs of male faces and the data from 130 of these were analyzed. (The data from the 20 other subjects were discarded because postexperimental questionnaire data showed that they had not understood or retained the content of the like-dislike portion of the communication.)

The data clearly indicate that the definition of the experimenter as either having or lacking a clinical orientation does, as predicted, have some influence upon the amount of systematically biased responding by the subjects. Under both the clinical and nonclinical experimenter conditions the control subjects (who received neither evaluation apprehension arousal nor directional cueing) lean toward an overall liking response pattern; and there is no difference between the mean algebraic sums of the ratings for the control subjects run by clinical and nonclinical experimenters. For the **(p.233)** former the mean of the algebraic sums is +25.20 and for the latter +23.00. In the clinical experimenter condition the subjects who received the “disliking is mature” cueing have a mean sum of +.13, while under the nonclinical experimenter condition the mean sum is +5.80. Apparently somewhat greater deflections away from the control group basal levels are occurring under the clinical condition. However, in both

instances the differences from the relevant control groups are quite significant ($p < .00005$ and $p < .0003$, respectively).

A more clear-cut booster effect is obtained with the subjects who received the "liking is mature" cueing. The subgroup run by nonclinical experimenters has a mean sum of +20.80; and this is not significantly different from the mean for the nonclinical, control group. However, the liking subgroup run by clinical experimenters shows a mean of +39.12. This differs significantly both from the means of the clinical control group ($p < .01$) and the nonclinical, liking-cued group ($p < .003$).

The following conclusions seem reasonable: In this subject population there was some tendency, as indicated by the control group data, to give moderately positive judgments of the pictured persons. Thus, for the liking-cued groups, the information that liking of strangers is a sign of maturity was congruent with their initial response disposition. But the identification of the experimenter as having a special interest in "what makes people tick" and in "psychiatry" operated to raise the stakes for the subjects run by the clinical experimenter. To guarantee the winning of a positive evaluation from him the typical subject in this group strives to give extreme, and thus unambiguous, proof that he possesses the defined hallmark of the mature, psychologically healthy person.

These are, then, unexceptional data. They seem to confirm the obvious and predicted relationship between the aroused strength of evaluation apprehension (the clinical definition having served to increase it, and the nonclinical definition to decrease it) and the degree to which the subject relies upon a response style he believes to be indicative of normal or attractive personality.

But, persisting in the mood of parametric exploration rather than theoretical expansion, the following simple question might be asked: Must the experimenter directly define himself as having a special interest or ability in evaluating personalities? Or can the same sort of evaluation atmosphere be induced by other means? One additional experiment seemed to confirm the latter possibility.

In this study we employed the same Background Information Sheet as in the previous one. By this means we again provided both for arousal of evaluation apprehension and directional cueing of responses in the "liking" and "disliking" directions respectively. And again we attempted to strengthen the evaluation apprehension dynamic by introducing an additional clinical implication into the experimental situation. Thus, before they read the Background Information Sheet the subjects in one main treatment read a printed announcement concerning an impending study. This told them that "Dr. P. J. Schroeder," a clinical psychologist from another institution, had asked our cooperation in recruiting subjects for a large study on "student personality and adjustment in college life." This study was being conducted on various different campuses.

Participation in it would involve the subject's being interviewed by Dr. Schroeder and allowing him to administer various “projective tests of personality.” Dr. Schroeder, it was made clear, would treat the findings as completely confidential and, specifically, he would not disclose them to the experimenter. The subjects were asked to sign for appointments “for this other, unrelated project” if they were so inclined. Virtually all the subjects did sign.

(p.234) When we compare the subjects in this treatment to others who were not exposed to it we find the former showing stronger directional bias effects than the latter. Under the “Schroeder is coming” condition the difference in ratings between subjects cued in the liking and disliking directions respectively is clearly significant ($p < .02$). Under the standard condition comparable to the prior experiment, but lacking any extra clinical implication, the comparable finding is $p < .10$. (Smaller samples were used in this study than in the previous one; and with variances of about the same magnitude the overall probabilities are, as would be expected, somewhat larger.)

As I have already suggested, these are studies of limited import and they offer no major surprises. Essentially, their value lies in lending support to this basic point: any aspect of the experimenter (or of the situation or setting in which he is encountered) that adds some further implication of interest in psychological evaluation will tend to increase the influence of the evaluation apprehension dynamic upon the subject's experimental responding. This statement assumes, of course, that some other provocations toward evaluation apprehension are also acting upon the subjects as, for example, the information that we conveyed through the Background Information Sheet. However, it would seem quite likely that our *additional* factors (i.e., the undergraduate experimenter's confessed clinical interest or the subject's elicited commitment to participate in a later personality evaluative study) could operate as *sufficient* factors in and of themselves. Further research would be required to confirm this rather obvious speculation.

But obvious relationships (even when they raise questions about the underlying and somewhat obscure sequences of events that mediate them) are less compelling than findings that raise new and unexpected issues. Therefore, rather than linger over the findings reviewed above, I shall turn now to some further preliminary studies concerning other variables. In both of these studies the major hypotheses were confirmed, but certain unexpected relationships were also encountered; and they are of a type that promises to deepen our inquiry into the operation of the evaluation apprehension dynamic.

In one of these studies we attempted to examine the consequences of making the experimenter a “gatekeeper” for the subjects. By this we meant, simply, that the experimenter was to be perceived by the subjects as likely to allow some, but not all of them into some rewarding activity area. In addition to setting up

gatekeeper and nongatekeeper conditions we also treated the manipulation of evaluation apprehension in a new way. In previous studies our Background Information Sheet had been designed to arouse (or confirm and amplify) evaluation apprehension; and control subjects who did not receive the Background Information Sheet provided the necessary baseline data. However, in the present study we used two forms of the Background Information Sheet, one intended to confirm and strengthen original promptings toward evaluation apprehension and the other to disconfirm and weaken them.

To convey fully the nature of this experiment I must linger briefly over the actual operations employed to establish the gatekeeper and nongatekeeper and high and low evaluation apprehension variables.

The gatekeeper manipulation consisted of the experimenter (myself) reading aloud a letter that, as he explained it, he had recently received from the executive director of the “National Policy Foundation” on whose advisory board, as the letter made clear, the experimenter himself had served. The letter invited the experimenter **(p.235)** to nominate some forty undergraduate students to participate in a “regional collegiate council” that would be discussing, as part of a funded national study, college student experiences and views on general issues concerning campus life and educational practices. These students, the letter made clear, would be well paid and through their participation they would “help to educate the experts and, hopefully, the academic and governmental hierarchs who will be receiving the progress reports and the final report of the total study.”

Having read this letter the experimenter said he had decided not to choose from among his own students but rather from the many subjects who were participating in “this week's experimental sessions,” and that he would be “contacting some of you about this within the next week or so.” In the nongatekeeper portion of the study we elected to say nothing along these lines, rather than to provide some contrasting or filler material.

The communication designed to arouse and suppress general evaluation apprehension explained that the study the subject was about to participate in was being conducted by the Personality Assessment Project (high evaluation apprehension) or the Mathematical Psychology Project (low evaluation apprehension). The low evaluation apprehension communication stressed that we were “interested in integrating the various dimensions or factors involved in social perceptual processes into a more simplified model by a method which statisticians term ‘stochastic-inferential mathematical modeling’.” With these, and similar comments, we intended to create in the subject the reassurance that we were studying some rather obscure and technical issues and had no interest in his unique personality. Thus, a further paragraph informed the subject that: “the photo ratings of all participants will be considered as a group and an overall

average score and statistical estimate of distribution variance will be obtained for this set of photographs."

The high evaluation apprehension communication informed the subjects that "these photographs comprise a recently developed *Test of Psychological Maturity*. The pattern of responses given to these photographs by psychologically mature undergraduate students has been shown to be significantly different from the responses of those manifesting psychological maladjustment." The communication went on to assert that our research was pointed toward further development of the picture rating personality test as an "instrument for picking out from the large college population those students who would be judged inadequate in their emotional adjustment or generally prone to psychological disorders."

Postexperimental questionnaire data established that the low evaluation apprehension communication did significantly "reassure" those who read it while the high evaluation apprehension communication did increase anxiety and set the person toward concern over what impression he would make upon the experimenter.

The contrast with our earlier uses of the Background Information Sheet should be clear. Not only were we attempting to remove evaluation apprehension in some subjects while strengthening it in others, but we were also providing no cueing that directly reported that liking or disliking for strangers had been found to be characteristic of psychologically mature persons. Instead, a more limited, or one might say, less obvious and intrusive form of directional cueing was employed. Each experimental subject, after he had been exposed to the gatekeeper or nongatekeeper and high or low evaluation apprehension manipulations, read a two-paragraph communication which simply reported that previous research with the pictures he **(p.236)** was about to rate had shown that most people judged them positively (liking) or negatively (disliking). While one third of the total subject population of 148 males received this form of the liking cueing, and another third the disliking cueing, the remaining third received no directional cueing and thus served as a control group.

In the actual administration of this experiment we were able to achieve a high level of efficiency by use of the language laboratory at the University of Chicago. Subjects were run in groups of eight to twelve. Each subject occupied a separate work booth. Seated in the booth he first heard the experimenter deliver the gatekeeper "pitch" or, for the nongatekeeper subjects, a brief and quite neutral introductory statement. The subject then read the high or low evaluation apprehension manipulation which, under instruction, he had removed from an envelope placed on the table within his booth. He then went on, unless a cueing control subject, to read the directional cueing communication. Following this he gave, on a rating sheet, his liking-disliking judgments for each of the 15 pictured

faces as they were projected on a screen easily visible to all subjects. After this rating sheet had been completed the sequence of pictures was presented again while the subject rated each of the pictured persons for “how successful” they had been. A third exposure of the pictures was then given while the subjects rated the pictured persons for “how intelligent” they appeared.[3] All pictures were exposed for ten seconds each, with a following ten second interval during which the subject wrote his rating on a scale from −10 to +10. Two postexperimental questionnaires, administered both before and after a thorough debriefing, provided strong evidence that the manipulations had been successful and that very little suspicion had been aroused as to our real purpose.

I have so far described the procedures of this study without any direct reference to the hypotheses that guided it. However, they are probably already apparent. The gatekeeper manipulation was intended to increase the desirability of winning a positive evaluation from the experimenter; for this would now have the additional payoff value of increasing the probability of being chosen for membership in the interesting and remunerative student discussion group that was being set up by the “National Policy Foundation.” Thus we predicted that response dependence upon the directional cueing would be greater for subjects in the gatekeeper condition than for those in the nongatekeeper condition.

Similarly we expected that subjects receiving the high evaluation apprehension manipulation would show stronger response bias effects than those receiving the communication that was designed to reduce evaluation apprehension. And, of course, we were interested in the possibility of a meaningful interaction between the two major variables, and also their respective and combined interactions with the like-dislike cueing variable.

This rather complex study, with 12 separate cells in a 2 × 2 × 3 design, and with considerable data drawn from postexperimental questionnaires and inquiry, yielded a great deal of information; and full presentation and analysis can only be attempted in a lengthy, separate article. Thus I shall dwell here only upon some of the major findings and their probable meaning.

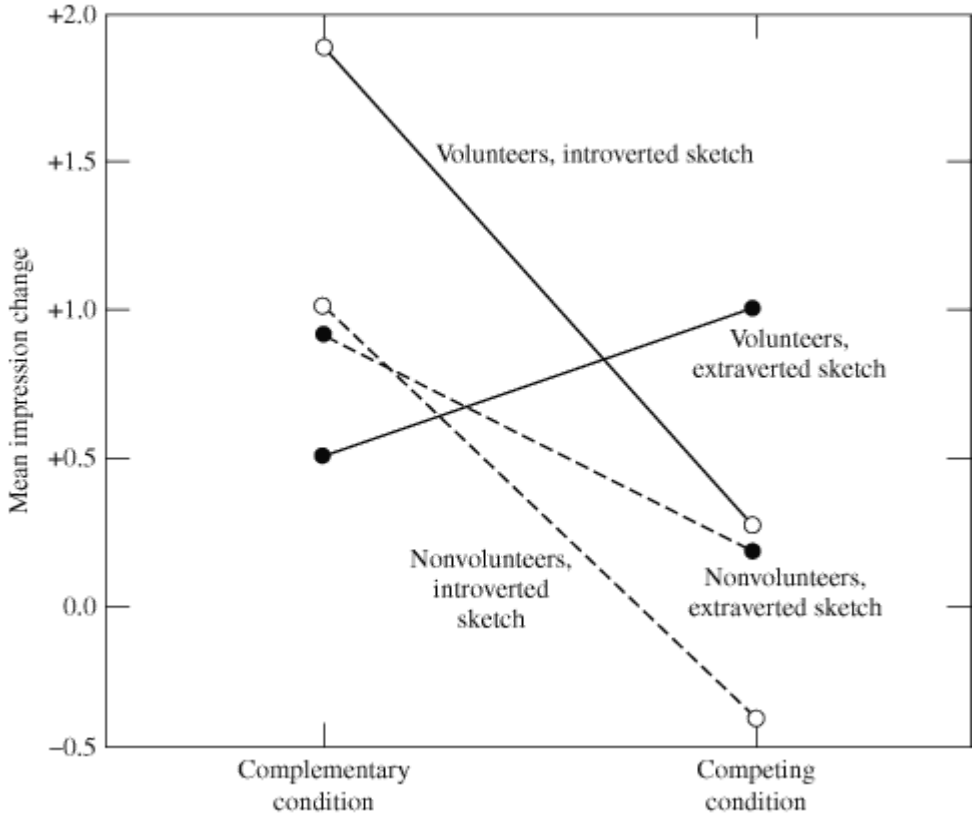
(p.237) Table 7–3 presents the mean algebraic sums of the liking ratings for the six cells that received the gatekeeper treatment and, separately, for the six cells in the nongatekeeper treatment. The probabilities of the differences between relevant pairs of cells are also presented. Reference to these tables will help to illuminate the findings from the separate analyses of variance that were carried out for both the gatekeeper and nongatekeeper conditions.

In both analyses we obtained clear evidence of a cueing effect. The algebraic sums of the subjects’ ratings on the like-dislike dimension strongly reflect the cueing that was received: those who got positive cueing gave more positive ratings than those who got no cueing and these, in turn, gave more positive

ratings than those who got negative cueing. In the nongatekeeper half of the experiment the *p* value for this effect is less than .0001. Also the effect does appear to generalize to the ratings of "success" ($p < .05$) and "intelligence" ($p < .006$).

Considering only the four groups that received cueing in either the positive or negative direction (i.e., eliminating the two no cueing, control groups) our other

Table 7–3 Like-Dislike Mean Sums for Groups; and Probabilities of Differences Between Groups



(p.238) major prediction was confirmed. In the nongatekeeper portion of the study a significant interaction is obtained between cueing and evaluation apprehension level as regards the liking ratings ($p < .03$). This is due to the fact that when subjects have been roused to a state of evaluation apprehension their picture ratings are more extremely influenced by either the positive or negative cueing than when they are at a low or suppressed level of evaluation apprehension. Thus, the mean of the algebraic sums for the high evaluation apprehension subjects who received positive cueing is some 53 points more positive than the mean for the high evaluation apprehension subjects who received negative cueing. For the low evaluation apprehension group the comparable discrepancy, while in the same direction, is only 25 points.

Similar effects of lesser magnitude and statistical significance are obtained when we compare the two evaluation apprehension groups on their ratings of the pictures for success and intelligence. The probabilities for the overall

evaluation apprehension by cueing interactions on these two dependent variables are less than .12 and .19, respectively.

In passing, it is worth noting that within the high evaluation apprehension condition the differences between the scores from the positively and negatively cued subjects are significant at probabilities of .008 or less for each of the three dependent variables; while the parallel analysis with the low evaluation apprehension subjects yields a significant probability only on the liking ratings.

I have dwelt upon these results because they suggest a point of particular interest both as concerns an emerging theory of the self-presentation process and also as they bear upon an important methodological issue. The kind of directional cueing intentionally provided in this study is often unintentionally present in other research situations, both of experimental and survey form (e.g., the respondent in the typical public affairs study often has a fairly clear idea, whether accurate or not, of “how most people would probably answer” on some of the more salient issues). More “valid” data (i.e., more accurate self-representations) are likely to be obtained when we attempt to reduce evaluation apprehension through some preliminary communication which disconfirms the subject's or respondent's concern that his psychological maturity (or, for that matter, his “public spiritedness” or “patriotism”) may be open to assessment and evaluative judgment.

Yet the fact is that even with an apparently successful reduction of evaluation apprehension (judging by the postexperimental questionnaire data from the low evaluation apprehension subjects) the directional cueing still exerts some influence. Probably this indicates some residuum of persisting evaluation apprehension and, if so interpreted, it points up the necessity for developing even more effective techniques for giving subjects or respondents the sort of reassurance which allows them to be their typical selves (i.e., uninfluenced by situational and inadvertent cueing factors) when reporting on their own judgmental or attitudinal processes.

So far the discussion has been restricted to the findings from the nongatekeeper portion of the experiment. With the data from the gatekeeper portion of this study we encounter a number of interesting patterns, particularly when they are viewed in relation to the comparable nongatekeeper experimental groups. Whereas the positively and negatively cued groups in the nongatekeeper, low evaluation apprehension condition differed significantly only on their liking ratings of the pictures, but not on the success or intelligence ratings, the low evaluation apprehension groups who received the gatekeeper manipulation show significant cueing effects on the liking, success, and intelligence ratings (respectively, $p < .00003$, $p < .0002$, $p < .0005$).

(p.239) This is further reflected in the difference between the liking means for the positively and negatively cued, low evaluation apprehension groups. Under the nongatekeeper condition this difference is 25.67, while the comparable difference under the gatekeeper condition is 44.83. For the success ratings the differences between the means for the two cueing groups are 3.60 for the low evaluation apprehension nongatekeeper condition and 37.34 for the low evaluation apprehension gatekeeper condition. With the ratings of intelligence the respective difference scores are 10.80 and 38.73.

It is clear that when we make the subject dependent upon the experimenter's judgment of him we restore something like evaluation apprehension. The subject, knowing that the experimenter is a psychologist and probably desiring that he “let him through the gate” to a rewarding experience, regulates his responding by reference to the cues that tell him how “most others” respond.

So far the results from the gatekeeper condition confirm our original hypotheses. However, where we examine the data from the high evaluation apprehension gatekeeper subjects, one major surprise is encountered: Unlike the results with the low evaluation apprehension subjects, the introduction of the gatekeeper condition (which was intended as an extra force compelling the subject toward reliance upon the directional cueing) seems in fact to *reduce* such reliance for the positively, but not negatively, cued group. In the high evaluation apprehension nongatekeeper and gatekeeper conditions the mean algebraic sums for the liking ratings in the absence of any directional cueing are 5.42 and 5.09, respectively. But whereas the liking sums for the positively cued subjects in the former group have a mean of 33.50 (and thus the difference between the control and positively cued subjects is 28.08), in the latter group the positively cued subjects yield a mean sum of only 13.82 (making the difference between the control and positively cued subjects only 8.73). Similar findings are obtained with the dependent variables of success ratings and intelligence ratings.

A possible interpretation is that the combination of the high evaluation apprehension and gatekeeper treatments strains the subjects’ credulity or, perhaps, puts them under a degree of tension which inhibits or otherwise disrupts their readiness to be influenced by the directional cueing. But the absence of the same pattern with the negatively cued groups limits the applicability of this interpretation. Subtler possibilities have occurred to us, but their explication had best await the results of further data analyses that are yet to be executed. These last findings comprise one of the valuable surprises of which I spoke earlier; and I must confess considerable interest in further experimental investigation in this particular realm as well as considerable frustration over the tantalizing ambiguity that presently beclouds the issue.

Among many further subsidiary findings obtained in this experiment I shall mention only one other. A postexperimental index of the “anxiety” aroused by the high evaluation apprehension communication is strongly correlated with the degree to which the subjects in the experimental groups were influenced by the directional cueing that they received. This serves to reinforce our general theoretical view while also suggesting the importance of apprehension-proneness as a mediating, personality-linked variable.

While I have not here attempted a full description of the procedures of this complex study or of all the available analyses, enough has been presented to make clear the basis for the following conclusions: Evaluation apprehension has again **(p.240)** been shown to be a factor, or process, that mediates systematic biasing of the sort that is due to cueing (in this study, somewhat more indirect cueing than in our previous work) of the preferred pattern of experimental responding. A second variable, namely the perception of the experimenter as a “gatekeeper” (i.e., as one who controls access to further reward or ego-enhancement) has been shown to facilitate yielding to directional cueing, particularly when evaluation apprehension has been brought to a low, or inoperative, level. But the combination of high evaluation apprehension and the gatekeeper variables has not, as we thought it would, worked to maximize the degree of influence upon experimental responding that is exerted by directional cueing. Whether this is due to some artifactual considerations (or to some unintended and subtler pattern of evaluation apprehension that has, in turn, generated a more obscure response strategy) or whether it is our first encounter with a truly general effect remains to be determined through further research.

In general this study does appear to add force to the claim that evaluation apprehension can contaminate the data gathering process, and it directs us toward a more complex consideration of other variables that interact with evaluation apprehension.

The last study that I shall treat in this section was, in all but two respects, a close duplicate of the one just described. Thus, its design and procedures can be outlined in short compass. Subjects were again exposed to the high and low evaluation apprehension treatments and then to either positive or negative directional cueing or to no cueing at all. Again the experiment was conducted in the language laboratory setting with each subject working in a separate booth and all viewing and rating the projected pictures at the same time.

The two major differences between this study and the previous one were: All subjects were female undergraduates; in place of the gatekeeper manipulation we attempted a systematic, two-stage variation in which the pictures to be rated were presented under conditions of high and low ambiguity. Operationally, this meant that under the nonambiguity condition each successive picture was exposed in sharp focus for ten full seconds and the subject was to give her rating

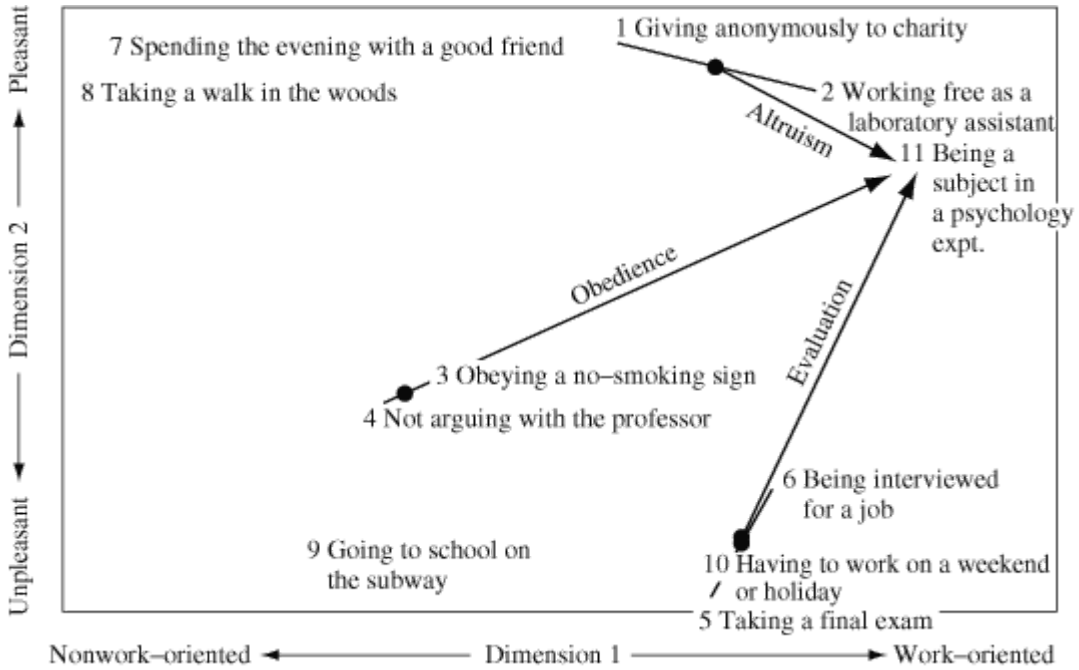
only after the exposure was completed. In the ambiguity condition a stable level of poor focus (low resolution) was employed and the picture was exposed for only three seconds.

The basic hypothesis which led us toward this study on the evaluation apprehension × cueing × ambiguity interaction was that biased experimental responding due to evaluation apprehension (in interaction with directional cueing) will be a direct partial function of the degree of ambiguity in the stimulus materials to which such responding is coordinated. Basic to this prediction was the notion that the stimulus attributes of the particular pictures do, in interaction with the subject's own judgmental standards, exert some influence upon his ratings. This is likely to be true even when a larger part of the variance in the ratings is controlled by the arousal and directional channeling of evaluation apprehension. To make the pictures more ambiguous is to make the stimulus attributes less readily available. This, in sum, should foster a further intensification of the subject's reliance upon such cueing as he may have received and thus the bias effects should be intensified.

Table 7–4 presents the mean algebraic sums of the liking ratings for the six cells in the ambiguity treatment and, separately, for the six cells in the nonambiguity treatment. The significant differences reported in the table help to make clear the findings from the analyses of variance that we carried out for both the ambiguity and nonambiguity conditions.

(p.241)

Table 7–4 Like-Dislike Mean Sums for Groups; and Probabilities of Differences Between Groups



Analysis of variance of the ratings from the half of the study in which the subjects rated the unambiguous photographs reveals comparatively strong cueing effects. On the liking ratings the cueing effect is highly significant ($p < .0001$); and for the success and intelligence ratings they are of borderline significance ($p < .15$, $p < .07$ respectively).

Analysis of variance of the ratings from the half of the study run under the condition of stimulus ambiguity also reveals a significant main effect for the liking ratings ($p < .002$), but no effect for the success ratings ($p < .68$), and a borderline effect for the intelligence ratings ($p < .14$).

However, while the like-dislike directional cueing exerts the predicted influence, we find that two other expectations are not directly confirmed: Within the separate ambiguity conditions we do not find that the high evaluation apprehension subjects are significantly more influenced by both types of directional cueing than are the low evaluation apprehension subjects; nor do we find significantly more biasing of responses in the cued directions in the ambiguous as compared to the nonambiguous treatments. Instead, what stands out is a complex interaction between direction of **(p.242)** cueing, low and high evaluation apprehension and the ambiguity-nonambiguity variable. This significant interaction can best be described in these terms: Under the ambiguity condition the positive directional cueing has greater influence upon the liking ratings than does the negative directional cueing; under the nonambiguity condition the negative directional cueing has greater influence than the positive directional cueing; and while this pattern is visible with both high evaluation apprehension and low evaluation apprehension subjects, it is somewhat stronger with the former under the nonambiguity condition and with the latter under the ambiguity condition.

Further and more complex analysis of these data, and of related data gathered with an extensive postexperimental questionnaire, carries us partway toward unraveling the meaning of the triple interaction reported above. But all such interpretation remains uncertain without recourse to further, replicative study. At this point the speculative path that seems most accessible is one which highlights the interaction between our independent variables and the special meaning of the experimental task. This speculative path begins with the assumption that most (American middle class) persons take it to be socially desirable to show openness and liking toward others. Those subjects who have received cueing suggesting that most persons in past research have rated the pictures negatively face a conflict between their own expectations or half-shaped hypothesis and the directional cueing that has been addressed to them. With stimulus ambiguity high they may, in the resultant state of uncertainty, fall back upon their own, original expectations; and thus the positive cueing works more effectively upon them than does the negative cueing. However, with high clarity and detail in the photographs, typical subjects may be able to find evidence in

facial and expressive characteristics onto which they can more readily impose the negative judgments that, according to the negative directional cueing, are typically made by “most people” who view these particular photographs.

That the yielding to the negative cueing under the nonambiguous condition is greater for high evaluation apprehension than low evaluation apprehension subjects (the difference between control and dislike group means being 38.91 for the former and 24.72 for the latter) suggests the further pertinence of the interpretation offered here: for the high evaluation apprehension subjects, believing they are undergoing indirect personality assessment, have a greater stake in regulating their responses in the cued direction. In effect, our interpretation, reduced to its simplest form, suggests this further hypothesis: to yield to directional cueing that endorses an unpracticed response style, the person needs “something to work with,” that is, some supporting aspects in the experimental situation or in the proffered stimulus material which will enable him to view his yielding to the directional cueing as having some basis in “reality” rather than solely in his need to win a positive evaluation.

Clearly this line of speculation, if strengthened by later research, moves our inquiry into self-presentation processes toward a subtler and more difficult kind of theorizing; one which will have to give fuller representation than heretofore to the limits and lures that the total experimental context provides for the subject who is attempting to regulate his experimental responding in a way that serves both his need for approval from others and, at the same time, from himself.

As I said in opening this section, “we are just barely started on this line of inquiry.” Having now reviewed our completed studies on variables that strengthen or reduce the data-biasing influence of evaluation apprehension I am all the more **(p.243)** sensitive to the fact that this work has a decidedly preliminary air about it. Much more inquiry is required and as it proceeds we must get beyond our present and too simple classificatory taxonomy of variables and into the construction of a process or systems model of the flow of the evaluation apprehension dynamic. Further work along these lines, both experimental and theoretical, is contemplated. But for now we can, I think, conclude that at least this much has been established: Between the initial arousal of evaluation apprehension and the ultimate tilting of experimental responses in the direction that, as the subject sees it, will maximize positive evaluation, there is scope for influence through many intervening and subsidiary variables. The few we have so far investigated appear to me to derive their influence in either or both of two ways: they may directly affect the subject's perceptions of *how* his responses will be judged; or they may affect his estimate of the *importance* of winning a positive evaluation from the particular experimenter in his particular experimental setting.

Evaluation Apprehension and the Experimenter Expectancy Effect

Three research strategies have been featured in the work I have already reported: altered replication, demonstration experiments and experiments on intervening or additive variables. Yet one other related research approach has figured in our recent work on the evaluation apprehension process. Simply described, this involves manipulating evaluation apprehension (by arousing or confirming it for some groups and suppressing or disconfirming it for others) and then examining the consequences for some other phenomenon or relationship of psychological interest. In general this strategy would appear to be relevant whenever one suspects that evaluation apprehension operates as a mediating or facilitating condition for an already established relationship between other variables.

Directly illustrative of my meaning is the possibility that evaluation apprehension may well be involved in the experimenter expectancy effect: that is, a state of concern over whether the experimenter will judge him as “normal” or “abnormal” may affect the way in which the subject perceives the experimenter's meanings, preferences, and aspirations within the experimental situation. To be specific: if, as the work of Rosenthal (1966) and Friedman (1967) suggests, the experimenter's expectancy is subtly communicated by aspects of his expressive style, the subject who is possessed of a concern over evaluation may well be more closely and accurately attuned to such indirect communication; or he may be more motivated to act upon the basis of what has been indirectly communicated.

To investigate such a possibility, then, one would attempt to replicate a standard experimenter expectancy study with at least two groups of subjects—one aroused to a high level of evaluation apprehension and one in which all tendencies toward this pattern of concern have been effectively diminished.

In a sense this research strategy can be viewed not as a forth and new one, but as a variant of the altered replication approach described in the first section of this chapter; but in this variant, instead of eliminating evaluation apprehension we attempt also to arouse it. However one wishes to classify it, this strategy has proved effective in the one realm in which it has already been employed. As the title of this section and the illustration offered above have already suggested, that realm has been the further study of the mediation of the experimenter expectancy effect.

(p.244) Before I turn to an account of our studies in this area I should like to comment briefly upon the relationship between my own preoccupation with the evaluation apprehension process and the work of other investigators of the “social psychology of the psychological experiment.” From the record of research (much of which is summarized in other chapters in its volume) on demand characteristics, subject presensitization, volunteer effects, and

experimenter expectancy effects, it seems abundantly clear that there are a number of sources of systematic bias in experimental data. For a long time these went unsuspected and, it can be assumed, contributed considerable nonrandom error to the data through which theoretical propositions were tested or inspired.

In the main I am persuaded by the work of others that the various processes that have been conceived as making for systematic bias do, in fact, have considerable operative force. And, obviously, I think and have tried to show that the same is true of the evaluation apprehension process.

We have then developed an empirically verified catalogue of data-biasing variables and processes. So far so good. But it seems apparent to me that we have now reached a stage at which we need not be content with a mere catalogue. Some larger, more integrative theory of the experimental-transactional process is required. The development of such a theory will afford intellectual satisfaction in itself; but, equally important, it will probably also contribute to a richer understanding of the role of self-representational dynamics in nonexperimental, interaction situations; and, of course, it will promise considerable further advance in improving the methods of research design and execution in all those disciplines (psychology is only one) whose data are gathered through interaction between the investigator and other, investigated persons.

I shall not presume to suggest the possible shape of a full and general integrative theory of the experimental process, though in the concluding section I shall risk a few preliminary speculations upon some aspects of such a theory. However, at this point I want only to register this obvious point: the development of this sort of theory will be advanced by—indeed, it may require—the prior investigation of the interaction and overlap between the biasing processes that are now separately delineated in our catalogue. A few investigations of this type have already been attempted; the study by Sigall, Aronson, and Van Hoose (1968) discussed above, is one. The three studies I shall now describe represent another such contribution. They are all focused upon the interaction between evaluation apprehension and experimenter expectancy. More particularly they are attempts to test the proposition already advanced: that is, that the experimenter expectancy effect is mediated or facilitated by evaluation apprehension. At the same time, the last of these studies also bears upon another important aspect of the experimenter expectancy effect; namely, the paralinguistic content of the experimenter's communications to the subject.

In our first effort in this realm my coinvestigator was Marshall Minor and a portion of this study served as his doctoral dissertation (Minor, 1967). We had two basic purposes: to replicate Rosenthal's finding that the expectancy held by an experimenter can introduce “experimenter bias” into the research situation so that the expectancy is confirmed by the response data elicited from subjects;

and, as I have already indicated, to show that the experimenter bias effect is mediated by evaluation apprehension. Particularly we hypothesized that the experimenter bias effect will be intensified when subjects believe that their experimental responses **(p.245)** may be utilized to evaluate their psychological adequacy, and that the effect will be diminished when they define the situation as one in which their psychological adequacy is not likely to be evaluated. The design of this study called for 16 naive male experimenters (eight given the +5 expectancy and eight the −5 expectancy) to separately run four subjects (two male and two female) through the Rosenthal picture rating task. (In this standardized task the subject rates each of a series of pictured persons on a scale from −10 to +10 for "whether the person pictured has been experiencing success or failure.") Half of all subjects would have already been roused to a high level of evaluation apprehension and the other half to a low or "suppressed" level of evaluation apprehension. Balancing in the assignment of subjects to experimenters was to be arranged so as to enable a statistical control for sequence effects, sex differences, and other possible unintended influences upon the response data.

Difficulties in recruiting the full complement of subjects during the University of Chicago summer quarter of 1966, and the failure of one experimenter to keep his scheduled appointment, reduced the actual situation to one in which 15 experimenters ran 23 male and 16 female subjects. However, it was possible to maintain partial balance in subject assignment and to effect statistical analytic controls for the "holes" in the matrix of experimenter-subject pairs that were actually completed. The latter type of control was made possible by use of the University of Chicago MESA 95 computer program (we are indebted to Professor Darrel Bock for introducing us to this program and for helping us to fit it to our needs) in which, as each mean square for an effect is determined, a transformation is made on the model by eliminating previously estimated effects from the succeeding mean squares. This made it possible for us to get an estimate of the expectancy × evaluation apprehension interaction in which estimate was independent of sex and sequence variables. (A full account of the handling of this and other analytic problems is found in Minor, 1967 and the special analysis of variance program employed is described in Bock, 1965.)

Some words about procedures and operations are required before we review the basic findings of this study.

The naive experimenters, all male graduate students in the school of education, reported individually for the scheduled experiment in whose execution they had promised to assist. After receiving a standardized background introduction to the experimental task each experimenter read a document giving further instructions. These expectancy-establishing instructions comprised a modified

version of one of the methods used by Rosenthal (1966). The crucial content was in the last paragraph:

> For your information, we have found from past research that certain types of people tend to rate these pictures in very particular ways. On the basis of personality test data that we collected earlier this quarter from the subjects you will be running, you should expect them to give an overall average rating of +5 (−5). This is because our earlier testing has shown that all of the subjects whom you will be running are in the category that we are calling “success perceivers” (“failure perceivers”). Thus if you will follow instructions and use proper experimental procedure, they will rate the pictures more extremely positive (extremely negative)—thus obtaining an average of about +5 (−5)—than would be the case if your subjects had simply been selected at random.

Each subject, before being introduced to the experimenter who was to administer the experimental task, was asked to “please read this sheet which will give you a **(p.246)** brief description of the purpose of today's experiment.” Half of the subjects read an explanation intended to arouse a high and stable magnitude of evaluation apprehension. The other half read an explanation designed to reduce and, hopefully, eliminate any tendency toward evaluation apprehension that the subjects might have brought with them into the experimental situation. The full logic that lay behind the construction of these two communications (which are similar to the ones used in our earlier studies on the gatekeeper and ambiguity variables) will not be traced here. However, much of it should be apparent from the content of the high and low evaluation apprehension communications which are fully reproduced below.

High Evaluation Apprehension Communication

“Today, you will be participating in a psychological experiment; and shortly you will be assigned to an experimenter who will explain the task to you. Although we are not able to answer any questions until after the experiment is over, we do want to give you a brief description of the purpose of the experiment. This should make participating more interesting and meaningful for you. Also, a growing number of psychological researchers are beginning to realize that they have an ethical responsibility to make the purpose of their experiments known to the individuals who are helping them out by participating in their research”.

“We are interested in studying social perception (i.e., how people perceive other people). More specifically, we want to find the factors which increase or decrease the accuracy of an individual's perception of other people. Certainly, with the lack of understanding in the world today, we do need to find out as much as we can about the reasons for inaccurate social judgment”.

“Prior research by ourselves and others indicates that, typically, poor social perception is associated with psychopathology. That is, people who are not able to accurately perceive how other people are feeling, or what they are experiencing, usually are found to be psychologically maladjusted. Much of our initial research in this area indicates that on the basis of performance on the social perception task, we can pick out from a college population those students who would be judged clinically to be maladjusted”.

“Several other researchers have presented data which support the preceding findings. Morgan and Provino (J. of Abnormal and Social Psychology, 1963) for example, report that in a college setting, the Social Perception Test could make rather subtle discriminations between varying degrees of emotional maladjustment and normalcy”.

“The purpose of today's experiment, therefore, is to replicate the previous results, and thus to test further the generality of the finding that people who cannot accurately judge what other people are experiencing tend to be psychologically maladjusted.”

Low Evaluation Apprehension Communication

“Today, you will be helping us to collect some preliminary data which we will use in setting up a subsequent research project. Shortly, you will be assigned to an experimenter who will explain the task to you. Time does not permit us to answer any questions, but we are able to give you a brief description of the purpose of the study. This should make participating more interesting and meaningful for you”.

(p.247) “We are interested in studying social perception (i.e., how people perceive other people). More specifically, we want to find the factors (e.g., fatigue, practice, etc.) which increase or decrease the accuracy of an individual's perception of other people”.

“Before we can investigate these different factors, however, we have to know how people perceive the feelings and experiences of others when these experimental factors are not present”.

“That is, we need a control, or standardization, group to use as a baseline against which we can judge the effects that our experimental factors have on social perception. This is the reason for your participation today”.

“We intend to average the performance of all of the students participating today, so that we will have a measure of how subjects perform on the task when such experimental variables as fatigue and prior practice are not present. This information will allow us to judge the effects which our experimental variables have when they are used with a subsequent group of students”.

“In other words, today's group will help us to find out how subjects typically perform on the task. Later, we can use the data we receive here to judge the performances of subsequent experimental groups of subjects.”

As in the typical Rosenthal experiment, interaction between experimenter and subject was held to a minimum level in which the experimenter read the picture rating instructions to the subject and collected his ratings for each of the ten pictures. Upon completion of this phase the subject, no longer in contact with the experimenter, filled out an extensive postexperimental questionnaire and was thoroughly interviewed. The same was done with each experimenter after he had completed running all of his assigned subjects. In a last phase experimenters and subjects were brought together for a full “debriefing” and for extended discussion, considerable care being taken to alleviate any lingering concern that might be felt by subjects who had been assigned to the high evaluation condition.

From the full analysis of variance three significant findings were obtained: Between experimenters, the expectancy variable (+5 versus −5 experimenter expectancy) controls a significant portion of the variance in their subjects’ ratings of how successful the pictured persons have been ($p < .05$). In the “within experimenters” analysis the sex of the subjects operates significantly ($p < .03$) reflecting a general tendency for females in either the + 5 or −5 expectancy groups to rate the pictured persons as less successful than the respective male subjects in the same expectancy treatments. Most relevant to our major interest is the finding of a rather strong interaction between expectancy and evaluation apprehension ($p < .02$).

The basis for this significant interaction is clearly revealed by a comparison of the mean photo ratings obtained from the +5 and −5 expectancy groups under both the high and low evaluation apprehension conditions respectively. (The male-female proportions are roughly equivalent in each of these four groups.) With evaluation apprehension reduced or suppressed (i.e., under the low evaluation apprehension treatment) the mean picture ratings for the +5 and −5 expectancy groups are −.78 and −.59 respectively. The difference between these means is not significant. Under the high evaluation apprehension condition the +5 and −5 expectancy group means are +.16 and −1.06 respectively. This difference is significant at a probability lower than .002.

Figure 7–1 provides a graphic representation of our basic finding that the experimenter expectancy effect is obtained, as predicted, when evaluation **(p. 248)**

apprehension has been aroused and is not obtained when evaluation apprehension has been reduced or eliminated.

Many additional aspects of the data analysis serve to develop further detail on the picture sketched here and to further strengthen our overall conclusions. These matters will be more fully reported in a separate publication. However, one particular subsidiary finding is worth noting here. An index reflecting the degree to which each experimenter was successful in inducing bias under the high evaluation apprehension condition was computed. This index was separately correlated with the scores from various questionnaires that were administered to the experimenters after they had run all their subjects. The two strongest correlations obtained were those with the Marlowe–Crowne Social Desirability scale ($r = .40$, $p < .06$) and the Sarason Test Anxiety Scale ($r = .57$, $p < .01$). This suggests the possibility that something like evaluation apprehension is involved not only in mediating the responsiveness of the subject to the experimenter's bias-inducing cues but perhaps also in setting the experimenter to emit such cues. At any rate, these findings suggest an empirical hypothesis worthy of further and more direct study, namely: that assigned expectancies will have a greater influence upon subject performance when the experimenters to whom these expectancies have been assigned have a high need for approval and a tendency to be apprehensive over the evaluation of their own competence.

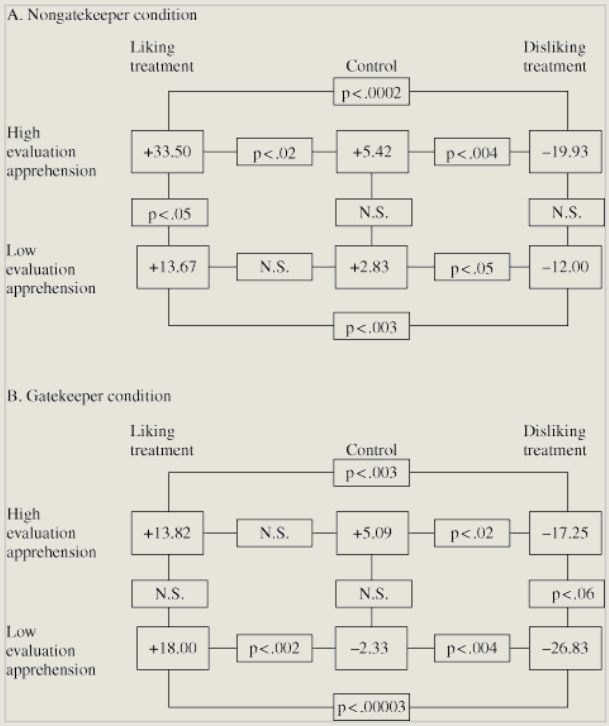


Figure 7–1 Response to Experimenter Expectancy as a Function of Evaluation Apprehension Level

(p.249) Upon completion of the analysis of this study we decided to attempt an altered and expanded replication. The major intended changes were these: to completely fill the matrix of required experimenter-subject combinations and thus handle the problem of sequence effects without recourse to the sort of statistical corrections that were required in the previous study; to run an "evaluation apprehension control" condition in which we would attempt to neither increase nor diminish the subjects' original, nonmanipulated, evaluation apprehension level; to run a "zero expectancy" as well as a +5 and −5 expectancy condition. A further purpose was to try out a way of staging the

study which combined some features of mass administration (e.g., subjects reading the initial evaluation apprehension communications while waiting in a large reception room) with the individual running of subjects on the picture rating task. Our hope in this last regard was to increase the efficiency of our own experimental procedures and those employed earlier by the Rosenthal group.

In this study, then, each of 33 male experimenters (11 each having been given the +5, 0, or −5 expectancies respectively) ran 3 male subjects on the Rosenthal picture rating task (one each having first received the high, low, or control evaluation apprehension communication respectively). The first two of these communications were slightly modified versions of the ones used in the earlier study and the last was a simpler one that merely advised the subject that he would shortly be assigned to an experimenter and asking him to wait until called upon.

The main results of this study can be quickly told. We failed to replicate the basic experimenter expectancy effect.

Analysis of postexperimental questionnaire and interview data from both subjects and experimenters suggests that this was due to our having failed to provide a credible experimental staging. In attempting to maximize efficiency in the routing of subjects and experimenters we seem to have aroused considerable suspicion about our own unrevealed purposes and about the actual contents of the communications intended to manipulate evaluation apprehension.

Without going further into the details of our post-hoc analysis of the suspicion-arousing aspects of the experimental procedures used, it may be said that a number of valuable cautionary points became clear to us and that we have profited from these in our further attempts at experimental investigation of experimenter bias, evaluation apprehension and kindred processes. In fact, on the basis of our first attempt at running a partially group-administered experiment in this realm, we were able to develop a different approach which was used in our next study in this sequence. This approach is just as efficient or more so, and yet seems to keep suspicion and other intrusive artifacts at a very low level.

Before turning to a description of the major study just referred to it will be necessary to briefly describe a study that was stimulated by our earlier work but was not undertaken as part of our research program.

Starkey Duncan, a clinical psychologist who had been working on paralinguistic aspects of communication within the psychotherapeutic situation, became interested in our work on experimenter bias effects. Through our joint consultations he came to the conclusion that the mediation of biasing cues in the Rosenthal paradigmatic situation might largely depend upon variations in the

nonlinguistic aspects of the experimenter's spoken communications to the subject. Particularly, he conjectured that the way in which the experimenter varied the intensity, intonation, pitch, and rhythm aspects of his reading of the instructions for the picture rating task might **(p.250)** convey to the subject an extra-linguistic (or, more properly, a “paralinguistic”) indication of the experimenter's expectancy regarding the responses the subject was about to make.

Duncan and Rosenthal proceeded to design a preliminary study to test this hypothesis. From films provided by Rosenthal, Duncan transcribed sound tapes of vocal readings of the instructions; three from each of two comparatively high biasing experimenters and four from a third high biasing experimenter. Together with the films from which the tapes were made, Rosenthal also provided the picture rating data obtained from the respective subjects who had received these separate vocal readings of the instructions. The taped readings were blindly coded on a number of different paralinguistic dimensions. The coding procedure used was based upon Duncan's earlier work. This procedure is extremely detailed and, with trained coders, yields high inter-judge reliability scores.

While the coding method will not be further described here, the results of this preliminary study can be simply summarized. Based only upon the coding of the instruction-reading tapes, Duncan was able to demonstrate that a large amount of the variance in the mean picture ratings given by the subjects could be accounted for by reference to the “Differential Emphasis Score” for each of the separate instruction readings that the respective subjects had received.

The Differential Emphasis score is a single index which reflects the degree to which the experimenter, in his vocal reading, has emphasized (through variations in volume, pitch, rhythm, etc.) either “success” or “failure” and either the positive or negative ends of the rating scale. The correlation between differential vocal emphasis and the subjects’ subsequent picture ratings was +.72 ($p < .01$); and all subjects who had heard greater emphasis on the rating alternatives associated with success subsequently rated the photos as being of more successful people than the subjects who heard readings that placed greater emphasis on the failure alternatives ($p < .001$).

An additional finding of considerable interest was that the correlation between experimenters’ assigned expectancies and the Differential Emphasis Scores was only .24. This suggests that though the pattern of emphasis used by the experimenters is influenced by the assigned expectancy it often varies from that expectancy in the direction of giving either greater or lesser than average emphasis to it. It suggests further that even where the relation between assigned expectancy and the subjects’ picture ratings is low, the experimenter

may actually be influencing the subject (through his deviant pattern of vocal emphasis) a good deal more than has previously been suspected.

From this preliminary study it seemed clear that with paralinguistic analysis considerable further progress could be achieved in pursuit of the difficult question of just how experimenter expectancy effects are mediated. Since the Duncan–Rosenthal study had used a variant of the method of postdiction it seemed especially desirable to attempt a more ambitious and more fully controlled study. We would reverse the procedure, moving from postdiction to prediction; this would be accomplished by exposing subjects to vocal readings selected for their paralinguistic direction (i.e. “success” or “failure”) and the degree of differential paralinguistic emphasis. Thus by experimental manipulation we could gain a closer and more stringent test of the hypothesis that in the typical experimenter bias study (and also in studies that may be inadvertently contaminated by bias effects) the subject's responses are influenced through paralinguistic aspects of the experimenter's communication to him.

(p.251) At the same time we planned to extend our earlier inquiry into the way in which the experimenter expectancy effect is mediated by the subject's state of evaluation apprehension. Thus the next study (in which I was joined by Duncan and Jonathan Finkelstein) was an experimental, manipulative investigation of the separate and interacting influences upon subjects’ response patterns of both paralinguistic emphasis and evaluation apprehension.

In the first phase we obtained from a number of colleagues and graduate students taped readings of the basic instructions for the Rosenthal picture rating task. Our request was that the first reading be given in an “objective and balanced” manner and that subsequent readings be “slightly shaded” in either a positive (i.e., “success” stressing) or negative (i.e., “failure” stressing) direction. None of these “experimenters” heard the readings of any other and each went about his “balancing” and “shading” in strictly his own manner.

After all the resulting speech samples were transcribed and scored for paralinguistic Differential Emphasis we were able to select a set of nine readings (three from each of three readers) to be used in the study. The three instruction readings taken from one of these experimenters were scored as balanced (i.e., no differential emphasis), moderate positive (i.e., intermediate bias toward an emphasis on perceiving the pictured persons as successful), and strong positive, respectively. From a second reader we had balanced, moderate negative, and strong negative readings; and from a third we had balanced, moderate positive, and moderate negative readings.

In the basic design of this study each of the nine instruction tapes was combined with each of three evaluation apprehension conditions, thus yielding a 27 cell

design. The evaluation apprehension conditions employed were High, Control, and Low. As in our earlier studies the evaluation apprehension manipulations were effected through a “Background Information Sheet” read by the subject. The evaluation apprehension bolstering and suppressing communications were similar to those used in earlier studies. The control evaluation apprehension group received no Background Information Sheet and was given no advance “explanation of the experiment.”

The subjects were 216 female undergraduates (eight per cell) who had volunteered in response to telephone calls requesting their participation in a study of person-perception. No payment or other rewards were offered. All experimental sessions were run in the University of Chicago language laboratory. In this facility the separate listening booths with multi-channel receivers could be easily adapted to a basic requirement of our design: namely, that within each administration group (*N* varied from 8 to 12 for the successive groups) each of the three thirds would respectively hear one of the three different readings of the instructions recorded by a single experimenter.

At the beginning of the experimental session, after all subjects were seated in their randomly assigned booths, they first heard a taped message thanking them for coming and, for the high and low evaluation apprehension groups, directing them to read the Background Information Sheet which was in a packet in front of the subject. After a five-minute pause for this purpose (control subjects were run in separate groups and had no such pause) each subject heard one of the taped readings of the Rosenthal instructions. Immediately following this the photographs to be rated for degree of “success” or “failure” were projected onto a screen in front of the **(p.252)** booths, each for ten seconds. The subjects recorded their own ratings on a standardized rating sheet which they were also required to sign. After the rating sheets had been collected a postexperimental questionnaire was distributed and following its completion and collection all subjects went through a thorough debriefing and were pledged to keep the purpose and design of the study confidential.

Before data analysis was undertaken 35 subjects were eliminated on the basis of important manipulation validation items from the post-experimental questionnaire. Thirteen were eliminated because they indicated that they had been aware of the purpose of the experiment; and 22 were eliminated either because they were in the low evaluation apprehension conditions and rated the Background Information Sheet as “anxiety arousing” or in the high evaluation apprehension condition and rated the Background Information Sheet “reassuring.”

Analysis of the data was based on the mean picture rating for each subject. Because comparisons between experimenters were not made in transcribing or scoring the readings of the instructions, and thus were not reflected in the

Differential Emphasis scores, it was necessary to adjust the subject means to take into account any differences among the experimenters. For each experimenter, therefore, the mean of all subjects in his control condition (i.e., the mean of the picture rating means from the subjects who heard his "balanced" reading of the instructions and had not received evaluation apprehension manipulation) was subtracted from the separate means of all his other subjects (i.e., those who heard his "biased" readings). These adjusted scores were used in our basic analysis.

Preliminary analysis indicated no significant difference in bias induction between the subjects who heard the moderate and strong positive readings from one of the experimenters or between the subjects who heard the moderate and strong negative readings from another of the experimenters. However, on inspection, clear differences were visible between those who heard positive and negative readings respectively, while those who heard "balanced" readings occupied an intermediate position. It was apparent, then, that somewhat subtler shadings of volume, pitch, and rhythm were just as effective as more pronounced ones in conveying a differential emphasis which influenced subject response patterns. In our further analysis we combined the data from subjects who had heard the moderate and strong positive readings of the instructions and, separately, from those who heard the moderate and strong negative readings.

When we tested the difference in scores between all subjects who had received the positive differential emphasis and those who had received the negative differential emphasis, the predicted main effect was confirmed ($p < .02$).

In a further and more detailed analysis the mean scores from the six separate cells were arranged in the ascending order that could be predicted from the assumption that the effects of differential emphasis would be facilitated to the degree that evaluation apprehension was experienced. That predicted order was: high EA, negative differential emphasis; control EA, negative emphasis; low EA, negative emphasis; low EA, positive emphasis; control EA, positive emphasis; high EA, positive emphasis. An analysis of variance was executed to determine whether the predicted order did, in fact, obtain. The resultant linear trend was found to be significant ($p < .02$).

Figure 7–2 reveals the basis for this summary statistic and reports further probabilities obtained through application of the Mann–Whitney Rank Sum Test. Thus **(p.253)**

subjects who had first read the Background Information Sheet designed to remove or reduce evaluation apprehension were apparently uninfluenced by the differential emphases conveyed in the various instruction tapes that they heard; no significant difference is found between the scores of the low evaluation apprehension groups that respectively heard positively and negatively biased readings of the instructions. On the other hand, when subjects have read a communication designed to confirm and heighten evaluation apprehension their picture ratings are very strongly influenced by the paralinguistic shading of the instruction tapes in either positive or negative directions ($p < .006$). Similarly we find that when evaluation apprehension is not manipulated (i.e., when, in the evaluation apprehension control condition, it is allowed to operate at a level that we may assume to be set by the interaction between the experimental task and the subject's personality) we obtain a smaller, but still significant, difference between subjects exposed to positively and negatively shaded readings of the instructions. The scope of this difference ($p < .02$) is roughly the same as that reported in typical successful experiments by the Rosenthal group.

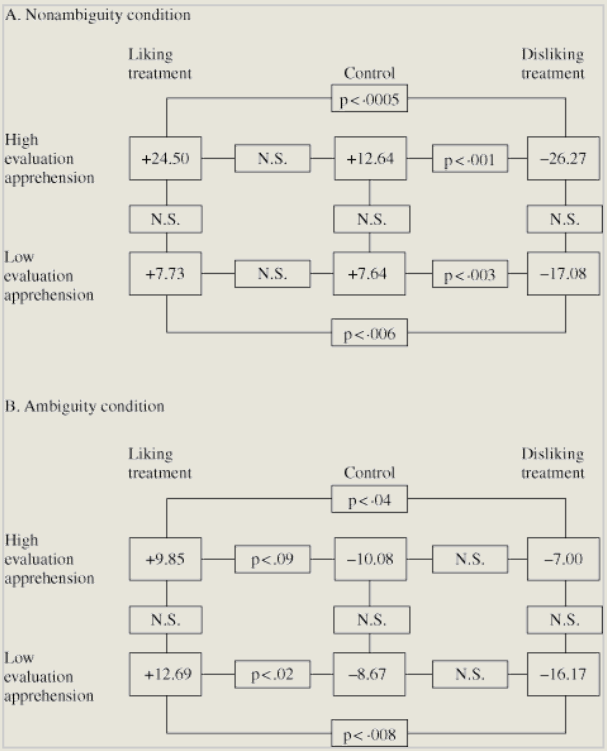


Figure 7–2 Response to Paralinguistic Differential Emphasis as a Function of Evaluation Apprehension Level

Further and more detailed analysis of these data remains to be carried out—particularly an analysis program that will draw upon material gathered through an open-ended postexperimental questionnaire. But on the basis of the main findings reported above we feel that we can now more clearly discern the nature and dynamics **(p.254)** of the experimenter bias effect. Loosely stated, the process appears to be one in which subtle paralinguistic shadings in the experimenter's communications do convey his expectancies or preferences as regards the response choices that the subject must make.[4] Whether the subject will be attuned to these paralinguistic cues or, if attuned, whether he will allow them to influence his experimental responding, may depend upon a number of things; but we are now in a position to conclude that one of these considerations, and probably an extremely important one, is whether the subject

has come to perceive the experiment as one in which the experimenter is likely to form judgments about the subject's psychological adequacy or attractiveness.

Some Recommendations and Open Issues

In bringing to conclusion a chapter that has already probably taxed the reader by its length and detail I shall resist the temptation to elaborate further upon my basic argument and its supporting evidence. By way of general summary it shall suffice to say that I have tried to explicate a conceptualization of the evaluation apprehension process and that all of the present studies appear to show that this process does induce systematic bias in experimental responding. Some of the research studies that have been reviewed have served an additional purpose: they have delineated and examined certain variables that appear to facilitate or restrict the operation of the evaluation apprehension biasing process. On the basis of the present studies I think it reasonable to put the seal of provisional validation (and the judgment that they are worthy of further experimental study) upon the following propositions:

The biasing influence of evaluation apprehension upon response data will be reduced if those data are collected by an experimenter other than the one whose evaluative judgment was the original focus of the subject's concern.

When a response pattern cued as likely to bring positive evaluation is also counternormative, subjects high on the need for approval will be more likely to produce that response pattern than subjects low on the need for approval.

The availability of continuous feedback about the quality of a subject's performance will facilitate his shaping that performance in the direction he thinks likely to earn him a favorable evaluation from the experimenter.

The less effortful the response direction that has been cued as likely to bring positive evaluation, the more will the subject go in that direction.

When the experimenter is perceived by the subject as having “power” over him (in the sense of controlling his access to some goal region or activity) this will foster the biasing of his responses in cued directions; and this will be particularly likely in the absence of other conditions that directly arouse evaluation apprehension.

(p.255) When the subject expects that a particular type of judgmental response will earn him positive evaluation from the experimenter, and when that type of response is also counternormative or unpracticed, his adoption of it will be facilitated by clarity in the stimuli to be judged.

Still other propositions supported or suggested by the research reviewed in this chapter could be summarized. But the past work is prologue to present and future concerns. Thus, my main purpose in this concluding section will be to

address some interesting implications and open issues that seem to be suggested by the studies that have been reported here.

The first of these is the question of whether one can draw from the present research and analysis any clear prescription concerning the conduct of psychological and related forms of research. A number of fairly obvious recommendations do come easily to mind. One of these I have already suggested: the altered replication approach does seem to afford a way of testing reinterpretations of experiments whenever it is suspected that the original data were influenced by inadvertent arousal of evaluation apprehension. This strategy can and should be more widely employed. Disputatious reinterpretation of the other man's research is easier than research itself; but the latter legitimates the former and assesses its relevance. Thus, whenever possible, these activities should be joined.

Another prescription follows rather obviously from our studies, described in the previous section, which demonstrated that evaluation apprehension mediates the experimenter-expectancy effect. These studies do seem to show that evaluation apprehension and its data distorting effects can be reduced (or, at least, minimized) if one defines the experimental situation in a certain way for the subject. Whatever the particular details of such preliminary communications, they should lead the subject to perceive at least two things about the experimenter and his experiment: that his interest is focused not so much upon individuals in their uniqueness as upon aggregates of persons in their normative or nomothetic aspect; that some purpose far more technical (and perhaps more “dry”) than personality study is being prosecuted by the experimenter.[5]

Credible messages to this effect, or “accidental” revelations of the same order, can probably be rather easily developed and almost as easily pretested. Undoubtedly, content and style will need to be varied with types of subjects and types of experimental situations; but if interest in handling this problem became widespread we would soon develop a technology of evaluation apprehension control that would, I think, contribute significantly to improving the quality and trustworthiness of psychological research.

However such a change in standard procedure would raise an important problem concerning an aspect of experimental method that has, in recent years, become quite ritualized. Should the postexperimental “debriefing” include an explanation of the evaluation apprehension problem and of the way it was brought under control? Recently, some commentators have argued that debriefing should not be conducted unless it is required to reduce anxieties or ego-injuries directly due to the experiment, **(p.256)** and unless it is also clear that the debriefing itself will not embarrass the subject or diminish his self-esteem by demonstrating his gullibility.

Against these considerations I would give great weight to the notion that experimenters have an ethical obligation to be as frank as possible with their subjects, even though full, disingenuous revelation must be deferred until all data are collected. Nor do I think that such revelation need be degrading. Whether the subject comes out of the debriefing feeling tricked, and exposed as an “easy mark,” or whether he comes out with a sense of having participated in a useful endeavor in which he played an important part and was honorably treated would, I think, depend largely upon the secret motives and visible style of the experimenter. Surely, as Kenneth Ring (1968) has suggested, the “fun and games” approach to experimental social psychology degrades subjects, trivializes research and, I would add, quite probably activates the evaluation apprehension dynamic so as to induce unsuspected but sizeable systematic bias in resultant data.

Candid and thorough debriefing, unmarred by any proclivity toward gloating, can do much for the experimenter's self-image and probably it also serves the enrichment of the subject's experience and knowledge. However it does generate a further problem—as much for experimenters who may employ evaluation apprehension control procedures as for experimenters pursuing other approaches. I refer, of course, to the risk that, despite the elicitation from the subject of a pledge to say nothing about the experiment to other potential subjects, the vessel of secrecy may spring leaks. This, in turn, may spoil the host culture of the naive subject pool without the experimenter knowing that anything of the sort has happened. It is my impression that the pledge to postexperimental secrecy is usually internalized when a bond of mutual trust has been woven; and I am not aware of anything that works better to insure that bond than full and candid postexperimental debriefing.

Furthermore, the postexperimental discussion that can be opened up by mutual debriefing tends to free the subject to reveal much of his own recent, subjective experience in the experimental situation. The information thereby gleaned can be of considerable help in determining whether evaluation apprehension or other contaminating processes may have been operating during the experimental transaction. Such discussion also provides the experimenter with a fairly comfortable occasion for asking whether “you had heard anything about this experiment from a previous subject,” just as it provides a facilitating context in which subjects are likely to respond to that query with candor.

I am aware, of course, that I am dealing here in lore and impressions. Clearly, more systematic research is required on the effects of the debriefing strategy upon the subject's self-esteem, upon his maintenance of the secrecy pledge and, for that matter, upon the value of his introspections about his experiences in the experiment. But until such a body of research has been undertaken and reported I do think it reasonable to hew to the general standard favoring postexperimental revelation of all deceptions and of all major experimental

purposes; and this should include discussion of the evaluation apprehension problem and of the techniques that have been used to bring that problem under control.

I shall turn now to a second major matter that requires some discussion. Among various other issues heretofore unattended is a question that any thoughtful reader must have already conceived as he has worked through these pages: Is the desire to win the experimenter's approval, and his judgment that one is psychologically **(p.257)** adequate, the only motive of interpersonal relevance activated in the subject during the experimental transaction?

Assuredly the case cannot be that simple. Even if we reduce our range of conjectural scan to patterns of motivational arousal that directly affect the subject's way of relating to the experimenter (and, thus, how he responds in the latter's experimental situation) a number of other possibilities come easily to mind. Though they are probably far less common than the process upon which this chapter has focused they do require discussion. At least one of these additional data-biasing processes is quite familiar to all psychological experimenters of sufficient experience and sensitivity. There are some subjects who most of the time (and many subjects who some of the time) are likely to want to confound the experimenter, to disconfirm what they perceive as his expectations, to violate what they construe as his apparent scientific hypothesis.

I am mindful that this observation partially contradicts the view elaborated by Martin Orne (1962). For him the experimenter's hypothesis is a “demand characteristic” to which subjects, by the very nature of their role, are prone to yield. This may often happen though, as I shall argue later in this section, when it does it is probably mediated less by a general role-based standard of cooperativeness than by the evaluation apprehension dynamic.

Under what sorts of circumstances, or with what kinds of persons, does the opposite tend to occur; that is, what accounts for the not uncommon instance in which the subject's purpose seems to be to “screw up the works”?

With an ease and haste that may bespeak defensiveness, psychologists are often prone to interpret such behavior as due to general hostility, to character-based “anality,” or to lingering reverberations of the oedipal revolt against authority figures. Such may indeed be the case with occasional subjects. But another dynamic process seems to me to be far more common. Evaluation apprehension, when strongly experienced, may sometimes generate a sort of reactive anger toward the experimenter; or it may be so intolerable as to require immediate “distancing” from the experimental situation and its evaluative implications. Either of these purposes, and yet other comparably defensive ones, can be served by turning the tables on the experimenter and giving indirect expression to a negative evaluation of him. Given the constraints of the usual experimental

situation, the most effective way of doing this may often be to disrupt the experimenter's enterprise by emitting just those responses which will, as the subject sees it, confound or disappoint him. Also if this can be done with a “light” style, with some visibly amused irresponsibility, a further defensive stratagem is brought into operation. The subject may then be able to believe that he has destroyed the evaluative significance of the experimental transaction; for, if he is clearly not taking the situation seriously, his behavior cannot be meaningfully interpreted as saying much about his true psychological nature or competence.

From the viewpoint of the experimenter, the problem posed by this sort of process is not so much that it may occur as that it may not be easily or reliably discerned. While skillful postexperimental inquiry may be of some use in reducing this problem, there is, I think, another important alternative. There may well be some personality patterns and some foci of regnant conflict that tend to heighten the likelihood that subjects will take recourse to the “confound the experimenter” strategy. The question begs for early investigation and psychologists interested in **(p.258)** the social psychology of the experiment will need to turn their investigative skills in this direction.

Equally compelling and probably even more readily open to systematic investigation, is the question of what attributes of the experimenter and of his instructions and preliminary explanations work toward the same effect. It is my untested impression that experimenters who are perceived by subjects as rather severe and unrevealing while, at the same time, intrusively “nosy,” are the ones most likely to arouse special data biasing patterns of resistance in some of their subjects. And obviously, it could be hypothesized also that the same is true of experiments that are perceived (or misperceived) as probing too deeply into anxiety-laden or low self-esteem areas of the private self.

At least one other rather obvious bias-inducing pattern requires discussion in the present context even though, as far as I know, it has not been submitted to any systematic study whatever. I have in mind the occasional sounding of the “cry for help” by a genuinely troubled or unhappy subject who thinks he ought to be, but presently is not, a patient.

Undoubtedly, this is far less common than the aspiration to appear “normal” and win a positive evaluation, but just how uncommon it is I do not know. From my own experience and that of colleagues with whom I have discussed the matter I would hazard this judgment: with some small number of undergraduate subjects (and, perhaps, most often with freshmen at times of situational stress) contact with a “psychologist” does activate the regressive longing for some show of support and sympathy from a wise, compassionate parent surrogate.

When this does occur a number of problems arise. The most important, in the light of our methodological concern, is that out of this background there may issue a pattern of experimental responding opposite to that with which this chapter has been most concerned, but just as troublesome for its data biasing consequences. Where involvement in the experimental situation fosters the tendency to emit the “cry for help” the subject, utilizing the same directional cues as are available to other subjects and either with or without fully conscious intent, may shape his experimental responding so as to make himself appear “abnormal” or troubled or anxious. In consequence, his pattern of experimental responding will lack valid bearing upon the hypotheses that are being put to experimental test.

The obvious corrective, again, is to submit the problem to systematic scrutiny through further research. A paradigm experimental situation such as the picture rating task used in some of our studies can be employed, and to it there can be attached fairly clear implications of “normal” and “abnormal” response patterns. Response deflection in the “abnormal” direction could be taken as an index of the motivation toward negative or “needful” self-representation. And variations in such an index could be examined against coordinate variation in personality indices, in systematic manipulations of the experimenter's style, the experimental script, and prior inductions of psychological stress. Out of some such research program there would probably emerge a set of useful cautionary strictures that would help to further reduce the problem of systematic bias in psychological and kindred types of research data.

My purpose in these last few pages has been to note that, in addition to the subject's striving toward positive self-representation as a way of reducing evaluation apprehension, there are some other, related trends which may also induce systematic **(p.259)** bias in response data. Returning to our main focus upon the former process, I should now like to address an issue that has haunted the discussion at a number of points but has not yet been fully confronted: What is the relationship between the evaluation apprehension dynamic and such other sources of systematic bias as the experimenter expectancy and demand characteristic processes?

The answer that I think most acceptable, though only in a provisional way, is already implicit in my earlier discussion of the two experiments in which we found that the activation of evaluation apprehension facilitated, and its reduction obliterated, Rosenthal's experimenter expectancy effect.

In their separate research and theorizing Rosenthal and, to a lesser degree, Orne have both emphasized the experimenter side of the experimenter-subject interaction: that is, they have delineated and demonstrated that experimenters do indirectly reveal what sorts of responses they would welcome from their subjects and they have also shown that this does, somehow, affect the responses

of those subjects. However they have had far less to say about the subject's side of the transaction; about the patterns of concern, apprehension, and ego-defensiveness which move him toward acting out, or at least coordinating to, the experimenter's implicit demands.

It is, of course, true that Rosenthal has addressed this issue in some of his fascinating side-excursions (Rosenthal, 1966) into the personality attributes of comparatively biasable and unbiasable subjects. But what has been required as well is a narrower or more process-oriented focus upon the actual psychological events that carry the subject through the experiment and up to the point at which he “delivers” the elicited gift of his responses.[6]

The evaluation apprehension process as defined in this chapter and as exemplified in our various studies appears now to be an important part of the subject side of the total experimental transaction. In the emerging general theory of the “social psychology of the experiment” it does not replace the account of experimenter expectancy effects developed by Rosenthal. Rather it extends it and perhaps also deepens that account by adding further clarity about the conditions under which experimenter bias is likely to be induced. As regards the demand characteristic process posited by Orne, the present approach does inevitably raise some difficulties and disposes me toward one note of disagreement. This concerns the motivational-perceptual pattern which facilitates the subject's yielding to the “experimenter's scientific hypothesis.” Where the experimenter's true hypothesis is clear to the subject (and I would think that usually it is not) yielding to it would most likely be mediated by the expectation that this will somehow bring approval or other immediate social rewards from the experimenter. To be sure Orne might be interpreted as saying that positive self-evaluation is being sought by the subject, particularly in that he may take pleasure in viewing himself as an accommodating and helpful person. But the present studies, **(p.260)** coupled with the very pertinent one by Sigall, Aronson, and Van Hoose (1968), suggest that evaluation apprehension focused upon the experimenter is a more potent and more basic pattern of subject sensitivity. Thus, I would hazard the hypothesis that the subject's readiness to help the experimenter make his scientific point, if experienced at all, is an instrumental stage in his search for reassuring evidence that the experimenter judges him as an acceptable or even attractively “normal” person.

My basic argument, then, is that our focus on evaluation apprehension adds to the picture developed by Rosenthal and other major contributors. By carrying us beyond the kind of biasing processes which can be traced to variability in the experimenter's behavior it directs us toward those which may be due to the figural highlights and ambiguities of the *experiment* itself.

While they do not logically require it, the experimenter-oriented theories sometimes tend to view the subject as a comparatively passive recipient of implicit “messages” or “cues” from the experimenter. This would suggest that where such cues are absent or imperceptible, systematic bias would be unlikely to occur. In distinction, a subject-oriented theory of the experimental transaction views the subject as seeking something from the experimental experience. In the present theoretical view that “something” is the experimenter's judgmental validation of the subject's psychological adequacy and on this basis, the ultimate maintenance or enhancement of the subject's self-esteem.

However, whether this or some other private purpose animates the typical subject is of less importance for the moment than the altered perspective that is opened to us when we lay basic stress upon the subject as seeker. From this emphasis there follows the necessary recognition that even when there is no direct cueing conveyed through the experimenter's behavior, the subject may be prone to construct some personal interpretation of the “true meaning” of the experiment. More often than not, he will speculatively examine the instructions he has received, the overall rationale that has been provided, the procedures and measuring devices to which he has been exposed; and out of the questions these raise for him and the hints they convey to him he will, if at all possible, draw some meaning, some guiding hypothesis about what is really being investigated and how he can best display himself to the investigator.

In this view, then, the experimental situation and, for that matter nonexperimental research situations as well, can activate the subject to search for their meaning. Whether the meaning found is often focused upon the evaluation theme, as I have argued, or upon yet other themes, there ensues a consequence as intellectually fascinating as it is methodologically troublesome. The subject's final “definition of the situation” will affect his responding and thus will be reflected in the dependent variable data.

To turn again to the problem of improving research procedures, the foregoing argument clearly suggests a further caution. The danger of inadvertent systematic bias in response data cannot be fully reduced by effective elimination of the experimenter expectancy and demand characteristic problems. We must remain sensitive to the possibility that the subject, no matter how acquiescent or calm he appears, may be actively processing his impressions toward the development of some interpretive hypothesis, one that will lead him to adopt a response strategy that may distort the resulting data.

An analogue for this whole process is provided by the larger number of our present studies, excepting those focused upon the experimenter expectancy **(p. 261)** phenomenon. In the former group of studies the systematic biasing of the subject's response patterns was not demonstrably due to any intraexperimenter or interexperimenter variations in behavioral style. Rather, the differences in

subjects' performances could be directly traced to the fact that the preparatory materials they read contained hints that they could then rather easily shape into hypotheses about the purpose, or the indirect revelatory significance, of the experiment.

In substantive research focused upon other psychological issues and conducted by experimenters who do not *intend* their experimental procedures to induce systematic bias, the suspicions aroused and the hints conveyed by the instructions, manipulations, and measures may be of more obscure origin and less certain import. Yet "seeking" subjects are prone to pick up whatever cues may be available in the structural and procedural detail of the experiment itself.

The more figural and prominent are the cues of this type, the more likely that separate subjects will come to the same or similar interpretive hypotheses about how to assure positive evaluation for themselves, or, for that matter, about how to reach still other social goals that they may be seeking. In consequence, it will be more likely that a systematic bias in one or another response direction will result. In contrast, the more obscure and the more numerous such provocations toward suspicion and interpretation, the more likely that subjects will reach comparatively unique interpretive hypotheses; and this will tend to foster "random" rather than systematic bias.

Either way, the consequence is an increase in the possibility that intrinsically valid hypotheses will be "disconfirmed" and intrinsically invalid ones "confirmed." Thus it becomes imperative that we submit to far closer scrutiny the processes by which subjects engage in active information seeking, ambiguity reduction and the development of interpretive hypotheses.

Whether subjects engage in such activities with full "consciousness" (i.e., with purposive self-direction and ratiocinative clarity) or, as I think more likely, with intersubject and intrasubject variability in motivation, effort, and attentiveness, is an interesting issue but not a crucial one. At the present stage what is most important is that we translate our research interest in such processes into the more specific questions that will make possible their controlled investigation. In my view the most useful focus of the required further research effort would be to ask just what variables determine when and how subjects go about formulating hypotheses; and what other variables influence the content and certainty of those hypotheses and the ways in which they are transformed into actual, data-yielding responses.

Equally important, of course, is the search for conditions which reduce the likelihood that such activities will take place at all. The reduction or elimination of evaluation apprehension (or the structuring of an experiment so that evaluation apprehension never arises) appears to be one such important

condition. But there are probably others and their discovery would be a great boon to the whole experimental enterprise.

Until all these matters have been more fully clarified through further research it is necessary that experimenters strive to abandon the image of the “average” subject as a passive and patient human component within a total experimental system; a component that, by processing inputs into outputs, somehow automatically reveals immutable psychological laws.

Having said this much I must hasten to add that I do believe that such laws exist in nature, and that the experimental method has been and will remain essential to the task of apprehending and confirming them.

(p.262) Those psychologists who have responded to recent research on the social psychology of the experiment with despair over the prospects of the experimental method itself are, I think, guilty of unjustifiable reactive depression and are casting out the baby with the bath. When they call for renewed recourse to “field studies,” to “natural observation” with “non-reactive measures,” and to phenomenological inquiry they are doing the behavioral disciplines a useful service. Those ways of gathering data (though equally open to systematic bias effects) can do a great deal to enrich inquiry into the regularities that govern man's psychological development and his functioning in relation to the persons and institutions that define his existence.

However, when such critics suggest that the experimental God is dead, they appear to have missed the point implicit in all research on the social psychology of the experiment. That point is that the experimental method can readily be used to perfect, or at least to significantly improve, itself. Any experimental demonstration of some source of systematic bias and of the process by which it operates immediately suggests procedures for the control and elimination of that source of bias. Another heartening consideration is, simply, that on the basis of present knowledge a great deal is already known about how to reduce the dangers of contamination and systematic bias. Such knowledge can also inform the critical evaluation of the worth of particular experiments as these are reported. The wheat, then, can even now often be separated from the chaff—and the yield is not a grossly unfavorable one.

A truly exciting and optimistic prospect has been opened by a decade of work on the social psychology of the experiment and I hope that it has been further advanced by the present inquiry into the evaluation apprehension process. We are approaching the point at which we may achieve a practical (if not philosophically perfected) solution to the classic epistemological problem of detaching the knower from the known; of allowing the order inherent in behavioral and social processes to tell us its own true story without any distortion due to promptings from the listener or failings of his listening device.

The velocity of further advance toward the improvement of both experimental and nonexperimental investigative procedures is likely to increase as research on the social psychology of social inquiry is vigorously prosecuted. And if, on occasion, one is troubled by the ostensible paradox that the processes inducing systematic bias may operate in our very investigations of systematic bias, there are at least two types of reassurance available. The lesser one is that every investigation in this realm profits the succeeding one; error should fall away as we continue to “zero in” toward the goal of bias-free research. The greater reassurance is that paradox itself is a goad toward intellectual and scientific adventurousness; the more closed off and ostensibly circular the problem, the more deserving it is of assault and solution.

References

Bibliography references:

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., and Tannenbaum, P. H. (Eds.) *Theories of Cognitive Consistency: A Sourcebook*. Chicago: Rand McNally, 1968.

Aronson, E. The psychology of insufficient justification: an analysis of some conflicting data. *In* S. Feldman (Ed.) *Cognitive Consistency*. New York: Academic Press, 1966.

Bock, R. D. A computer program for univariate and multivariate analysis of variance. *Proceedings of the I.B.M. Computer Symposium on Statistics*. White Plains, New York: I.B.M. Data Processing Division, 1965, 69–111.

(p.263) Brehm, J. W., and Cohen, A. R. *Explorations in Cognitive Dissonance*. New York: Wiley, 1962.

Brown, R. Models of attitude change. *In* R. Brown, E. Galanter, E. Hess, and G. Mandler. *New Directions in Psychology*. New York: Holt, Rinehart and Winston, 1962, 1–85.

Carlsmith, J. M., Collins, B. C., and Helmreich, R. L. Studies in forced compliance: I. The effect of pressure for compliance on attitude change produced by face-to-face role-playing and anonymous essay writing. *Journal of Personality and Social Psychology*, 1966, **4**, 1–13.

Chapanis, N., and Chapanis, A. Cognitive dissonance: five years later. *Psychology Bulletin*, 1964, **61**, 1–22.

Crowne, D. P., and Marlowe, D. A new scale of social desirability independent of psychopadiology. *Journal of Consulting Psychology*, 1960, **24**, 349–354.

Crowne, D. P., and Marlowe, D. *The Approval Motive*. New York: Wiley, 1964.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden, 1957.

Festinger, L., and Carlsmith, J. M. Cognitive consequence of forced compliance. *Journal of Abnormal and Social Psychology*, 1959, **58**, 203–210.

Friedman, N. *The Social Nature of Psychological Research: The Psychological Experiment as a Social Interaction*. New York: Basic Books, 1967.

Minor, M. W. *Experimenter Expectancy Effect as a Function of Evaluation Apprehension*. Unpublished doctoral dissertation, University of Chicago, 1967.

Nowlis, V. Research with the Mood Adjective Check List. *In* S. S. Tomkins and C. E. Izard (Eds.), *Affect, Cognition, and Personality*. New York: Springer, 1965.

Orne, M. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implication. *American Psychologist*, 1962, **17**, 776–783.

Riecken, H. W. A program for research on experiments in social psychology. *In* N. F. Washburne (Ed.), *Decisions, Values, and Groups*. Vol. 2. New York: Pergamon Press, 1962.

Ring, K. Experimental social psychology: some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, 1967, **2**, 113–123.

Rosenberg, M. J. Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, 1956, **53**, 367–372.

Rosenberg, M. J. An analysis of affective-cognitive consistency. *In* Rosenberg, M. J., Hovland, C. I. *et al., Attitude Organization and Change*. New Haven: Yale University Press, 1960. (a).

Rosenberg, M. J. Cognitive reorganization in response to the hypnotic reversal of attitudinal affect. *Journal of Personality*, 1960, **28**, 39–63. (b).

Rosenberg, M. J. When dissonance fails: on eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology*, 1965, **1**, 18–42.

Rosenberg, M. J. Some limits of dissonance: toward a differentiated view of counter-attitudinal performance. *In* S. Feldman (Ed.) *Cognitive Consistency*. New York: Academic Press, 1966.

Rosenberg, M. J. Hedonism, inauthenticity, and other goads toward expansion of a consistency theory. *In* R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb,

M. J. Rosenberg, and P. H. Tannenbaum (Ed.) *Theories of Cognitive Consistency: A Sourcebook*. Chicago: Rand McNally, 1968.

Rosenberg, M. J., Hovland, C. I., McGuire, W. J., Abelson, R. P., and Brehm, J. W. *Attitude Organization and Change*. New Haven: Yale University Press, 1960.

Rosenthal, R. *Experimenter Effects in Behavioral Research*. New York: Appleton-Century Crofts, 1966.

Sigall, H., Aronson, E., and Van Hoose, T. *The Cooperative Subject: Myth or Reality*? Dept. of Psychology, University of Texas, 1968 (mimeographed).

Silverman, I. Role-related behavior of subjects in laboratory studies of attitude change. *Journal of Personality and Social Psychology*, 1968, **8**, 343–348.

Silverman, I., and Regula, C. R. Evaluation apprehension, demand characteristics, and the effects of distraction on persuasibility. *Journal of Social Psychology*, 1968, **75**, 273–281.

Notes:

(1) One clear voice that helped break the silence was that of Henry Riecken. In a valuable article published in 1962 he proffered a general view of the psychological experiment as a sort of ritualized exchange between subject and experimenter. An important aspect of the exchange dynamic, as he saw it, was the subject's desire to "put his best foot forward." However, in Riecken's view, this was basically a source of "unintended variance" in data and the possibility that it could exert systematic influence making for false confirmation or disconfirmation of hypotheses was not directly examined.

Also focused upon the self-presentation process were the inquiries by Edwards (1957) and Crowne and Marlowe (1964) concerning the "social desirability" variable. In distinction to the work described in this chapter, their basic interest has been with the contaminating influence of positive self-presentation upon psychological testing and its results rather than upon psychological experiments.

(2) The probabilities reported here as confirming the differences between the groups in this study are all based upon the one-tailed test. Throughout this chapter the same convention has been employed whenever the direction of a difference was predicted—though, as will be seen, most of the findings would easily retain their statistical significance even if the more stringent, but less appropriate, two-tailed standard were applied. Within the tables summarizing the statistical findings a designation of "N.S." (i.e., not significant) represents a probability value larger than .10, usually considerably larger.

(3) It should be clear that the subjects had not received any directional cueing concerning the personality revealing relevance of judgments that others have

been successful or are intelligent. However, judging another as possessing these qualities would represent a positive evaluation of him. Thus we expected some generalization from the subjects' judgments on the like-dislike dimension onto these two other judgmental scales. Also, evidence of such generalization (or of such indirect cueing effects) could be taken as an additional measure of the degree to which the directional cueing was utilized by the subject.

(4) Of course, we cannot logically rule out the possibility that still other modes of mediation such as "kinesic" cueing may play a role in conveying the experimenter's expectancy to the subject. What is clear is that no such additional channel of indirect communication was open in the present study since the only experienced differences between the three separate instruction readings contributed by any single "experimenter" lay in their differential paralinguistic emphases. Also relevant in this connection is this further fact: In the present study, the magnitude of the experimenter expectancy effect under both the control and high evaluation apprehension conditions was as great, or greater than, that obtained in most experiments concerned with this type of bias. This strongly suggests that differential paralinguistic emphasis is the main, if not the only, process through which the direction of the experimenter's expectancy is transmitted to the subject.

(5) At the same time it would probably be necessary to guard against making the experiment seem so empty of purpose or relevance as to destroy the subject's motivation to remain psychologically involved in it. Clearly, some art (and some validation of its products) will be required in the further development of techniques for limiting and reducing evaluation apprehension.

(6) While it has been insufficiently developed, this sort of concern has not been totally ignored during the short period in which the social psychology of the experiment has commanded intellectual interest. Some usefully provocative beginnings in this direction were elaborated by Riecken in his seminal article (1962); and Orne, despite the experimenter-oriented nature of the demand characteristic concept, has also been somewhat sensitive to these matters.

However, while the focus upon subject processes has not been totally absent in earlier speculative writing, it has lagged in development. Perhaps this is due to its having been obscured by the deserved figural prominence of the work on experimenter expectancy and demand phenomena. The proper corrective lies not in abandoning the latter interest but in restoring and expanding our concern with the former.