Influence of Subject Response Style Effects on Retrospective Measures

George S. Howard, Jim Millham, Stephen Slaten, and Louise O'Donnell
University of Houston

Recent attempts to reduce internal invalidity in studies employing pretest/posttest self-report indices of improvement have included the refinement of methodologies employing retrospective reports of pre-treatment states. The present study investigated the operation of social desirability and impression management response bias on such retrospective measures. The results do not support the hypothesis of greater bias on retrospective measurement and, in fact, are in a direction that might suggest an interpretation of reduced bias on such measures. The results also continue to support superior validity of retrospective over traditional pretest/posttest indices of improvement following treatment.

A science progresses by constantly revising, updating, and improving its research methodologies. Campbell and Stanley (1963), in a landmark work, analyzed the strengths and weaknesses of various experimental and quasi-experimental designs. One of their conclusions was that "true" experimental designs controlled for all potential sources of internal invalidity. A recent series of investigations (Howard & Dailey, 1979; Howard, Dailey, & Gulanick, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck, & Bray, 1979) have demonstrated an instrumentation-related source of internal invalidity in some true experimental designs, referred to as response-shift bias. The problem may arise whenever self-report instruments are employed to evaluate a treatment or training intervention.

Social scientists often evaluate interventions that are designed to alter not only a subject's behavior in a target domain (e.g., assertiveness, interviewing skills, dogmatism) but also his/her understanding or awareness of the target concept itself and his/her level of functioning with respect to that concept. Consequently, to the extent that a program meets its goals, subjects' understanding of the concept on which they are to self-report will be different at posttest than at pretest. The shift in understanding/awareness is referred to as a response-shift, and its presence renders pretest with posttest comparisons inappropriate.[1]

[1] A hypothetical example of a response-shift presented by Howard and Dailey (1979) was: A workshop participant might feel at pretest that he/she is an "average" leader. The intervention changes his/her understanding of the many skills involved in being a leader. Consequently, after the workshop, he/she understands that his/her level of functioning was really below average at pretest. Suppose this participant improved his/her leadership skills as a result of the intervention and moved from below average to average with respect to his/her *new* understanding of leadership. The ratings at pretest and posttest would both, then, be "average." If one does not consider that these ratings are based upon different understandings of the dimension of leadership, one might erroneously conclude that the subject had not profited from the workshop.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 1, Winter 1981, pp. 89–100

Campbell and Stanley (1963) controlled for instrumentation effects by recommending the use of objective raters. However, whenever self-report measures are employed, the subjects themselves serve as raters. Previous studies on response-shift bias indicate that while treatment subjects experience response-shifts, no-treatment control subjects, as expected, do not. Consequently, any traditional comparisons between treatment and control subjects, such as comparison of posttest-pretest change scores or posttest-only comparisons are inappropriate (Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979).

Howard and his colleagues recommend the use of retrospective pretests instead of traditional self-report pretests (Pre) as a means of controlling for response-shift bias effects. Retrospective pretests are obtained at the time of posttesting by asking each subject to respond to each item on the self-report measure twice. First, they are to report how they perceive themselves after the intervention (Post). Immediately after answering each item in this manner, they are to answer the same item again, this time in reference to how they now perceive themselves to have been just before the treatment was conducted (Retrospective Pre, or Then). Subjects are instructed to make the Then response in relation to the corresponding Post response in order to insure that both responses are made from the same perspective. Each set of ratings is scored separately to yield a Post score and a Then score.

When considering the use of retrospective measures, two issues become salient. First, does the use of retrospective measures lead to differing conclusions regarding the effectiveness of an intervention from traditional self-report pretests? There have been 10 studies to date where Pre/Post and Then/Post measures of change were compared. In four instances Then/Post analyses found significant treatment effects, whereas Pre/Post analyses did not find differences. In another four instances significant treatment effects were found using both approaches, and in the two remaining studies treatment effects were not observed with either approach. Therefore, in a substantial number of instances, use of retrospective measures does result in differing outcome conclusions from the traditional self-report Pre/Post approach.

A second issue to be considered is, which method provides the more valid results? In five separate analyses of the impact of intervention procedures ranging across assertiveness training, interview skills training, helping skills training, and interpersonal effectiveness training (Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979), the results from the Then/Post measurement approach were more similar to findings obtained from objective behavioral ratings of subjects' role-playing than were the results obtained from traditional Pre/Post self-report methods. Further, in a study investigating actual changes in amount of information acquired in a college course, Then/Post self-reports of content learned reflected more accurately the students' actual mastery than did the Pre/Post self-report approach (Howard, Schmeck, & Bray, 1979). In no study comparing Then/Post and Pre/Post self-report methods was the Pre/Post measure superior to, or even equivalent to, the Then/Post approach in reflecting behavioral indices of change.

It would appear from evidence available currently that the Then/Post measurement approach represents a significant and potentially more accurate alternative to Pre/Post self-report measures of change. It is important, therefore, to examine more extensively the parameters of Then/Post responding, particularly with respect to those sources of error that have seriously limited the usefulness of self-report measures as indices of change in intervention outcome studies.

The most widespread criticism of self-reports of change following treatment has been the operation of social desirability responding and related compliance with implicit task demands to report "improvement" following treatment. It has been argued that posttreatment retrospective self-reports might represent more accurate statements of pretreatment states than reports obtained prior to treatment because of greater familiarity with the behavioral dimensions being studied and a better opportunity for sensitized and reflective self-evaluation. However, it is possible that such enhanced familiarity with the behaviors and the goals of intervention might accentuate confounding due to social desirability responding and compliance with implicit task demands. That is, the superior accuracy of Then scores might be due neither to a greater understanding of the dimension of interest nor to subjects' increased awareness of their level of functioning on that dimension. Instead, the improved accuracy might be due to changes in subjects' susceptibility to various response-style influences. If this were the case, one obvious conclusion would be that Post/Then comparisons between treatment subjects (who are influenced by response-style effects) and no-treatment control subjects (who are not influenced as highly by response-style effects) would be inappropriate.

The present study investigated the operation of social desirability confounding in three ways. First, the relationship of individual differences in general social desirability responding to Pre and to Then self-reports of pretreatment states was determined. If posttreatment retrospective evaluations are more confounded than pretreatment measures by self-deceptive and impression management tendencies, a more powerful relationship would be expected between a social desirability measure reflecting these tendencies and the retrospective scores than that obtained between the pretreatment scores and the social desirability measure.

Second, a direct test was undertaken of the operation of impression management reflecting compliance with the implicit task demands to demonstrate improvement to the evaluator. A bogus pipeline technique was employed to assess the operation of such impression management responding. The bogus pipeline technique controls for a considerable portion of the variance attributed normally to social desirability responding in experimental situations. In the pipeline, subjects are led to believe that a "physiological monitoring device" is capable of assessing the truth or falsehood of their responses; it enables the experimenter to obtain responses uncontaminated by many of the biases that obscure paper-and-pencil measures (Jones & Sigall, 1971). The impression management component of social desirability responding has been demonstrated to be operating when significant differences in self-report evaluations are obtained between bogus pipeline and non-bogus pipeline testing conditions (Millham & Kellogg, in press). Therefore, differences in response-shift (Pre/Then) following treatment under bogus pipeline and under non-bogus-pipeline measurement would reflect operation of impression management in the retrospective evaluations.

Finally, all previous investigations of response-shift bias employed self-report measures that related to the intervention being evaluated. Larger Pre/Then differences for treatment subjects were consistently interpreted as due to subjects' changes along specific treatment dimensions rather than as a generalized compliance with task demands to demonstrate improvement following intervention. Therefore, it would be expected that no response-shift for treatment subjects on a measure unrelated to the treatment would be found. If such a response-shift was noted, then the operation of generalized compliance with task demands would have to be suspected. Consequently, a self-report measure of subjects' learning styles was included in this study to determine if Pre/Then differences for treatment subjects were obtained on a measure unrelated to the content of the treatment intervention.

Bogus Pipeline Pilot Study

Method

Prior to beginning the study, a pilot study that investigated the adequacy of the bogus pipeline manipulation was conducted. Forty students who participated in this study for course credit were randomly divided into two groups. The first group simply answered a few questions that asked for demographic information, the College Self-Expression Scale (CSES; Galassi, Delo, Galassi, & Bastien, 1974), a slight revision of the Learning Styles Questionnaire (LSQ; Schmeck, Ribich, & Ramanaiah, 1977), and the Jacobson-Kellogg social desirability scale (J-K; Jacobson, Kellogg, Maricauce, & Slavin, 1977). These instruments were chosen because of their inclusion in the major study reported in this paper. The second group was informed that the study involved validating a new voice analyzer, which was capable of determining the accuracy of a person's response by analyzing its emotional content.

One student had been contacted prior to the study and had been asked to serve as a confederate. At the beginning of the study, a volunteer was requested from the group to demonstrate the voice analyzer procedure. The confederate was selected and was asked to answer five questions, purposely giving incorrect responses to two or three questions. However, since it was a test, the confederate was to try to conceal the incorrect responses. The five questions were (1) What is your astrological sign? (2) How many brothers and sisters do you have? (3) Are you currently enrolled in a history course? (4) How many odd digits are there in your social security number? and (5) What is your class (e.g., freshman, sophomore)? The confederate was asked to respond to each question in sentence form and the responses were audiotaped. The experimenter's assistant then left the room to "have the tape analyzed." Meanwhile, the experimenter asked the confederate which questions had been answered incorrectly. The assistant returned shortly thereafter and identified the three questions that the confederate had indicated had been answered incorrectly.

The experimenter then explained the sequence of events that all subjects would complete for the experiment. Subjects would complete the three questionnaires. When they finished, they would answer the same questions the subject in the demonstration had answered and their voices would be analyzed to be certain that the technique would work for them. After this confirmation, they would then answer each question on the tests and have the responses audiotaped. These audiotapes would later be analyzed as part of the study. Subjects were asked if they had any questions, told to begin completing the questionnaires, and when finished to leave the room. Research assistants would then direct them to individual rooms to audiotape their responses.

When subjects completed the questionnaires they were simply sent to another room where the questionnaires were collected and the subjects debriefed, given course credit, and dismissed. Subjects in the bogus pipeline group were amused by the deception, and all indicated that they had believed the deception when they completed the questionnaires.

Results

Table 1 presents the data for the two groups on the three questionnaires. Scores on all three scales were significantly lower for subjects in the bogus pipeline condition than for the control group. The mean difference between CSES scores for the two groups was 15.9 points, while the difference on the J-K was almost 9 points. It is possible that the differences in CSES scores might simply reflect a tendency for subjects to rate themselves unrealistically highly under normal conditions. This finding might be particularly important since in their Study 3 Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) reported Then scores that were consistently lower than Pre scores for the treatment subjects. These Pre/Then differences might have been due in large part to the effect of the other deceptive component of social desirability on CSES Pre scores, rather than to the increases in awareness hypothesized by response-shift bias explanations.

Table 1
Results of CSES, LSQ, and J-K Questionnaires for Bogus Pipeline and Control Groups

	Bogus Pipeline (N=18)		Control (N=18)		
Scale	Mean	S.D.	Mean	S.D.	t
CSES	107.3	111.4	123.2	128.9	2.32*
LSQ	97.1	120.6	117.8	102.1	2.30*
J-K	18.0	20.3	26.9	29.9	2.80**

*p < .05; **p < .01

If this were the case, a high positive correlation would be expected between CSES and J-K scores for the control subjects but a substantially lower correlation would be expected between the two scales for the bogus pipeline group in which the other deceptive effects had been removed from each measure. The correlation between CSES and J-K scores for control subjects was .14, whereas the correlation for bogus pipeline subjects was .07. These weak correlations suggest that partialling out the effects of social desirability from CSES Pre scores would not alter the CSES Pre scores enough to account for more than a fraction of the Pre/Then differences which Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) attributed to response-shift bias effects. It was concluded that the bogus pipeline procedure eliminated the other deceptive component of social desirability and would thereby attenuate subject impression management and compliance with task demands to "improve" in the principal study to be conducted.

Method

Subjects and Experimenters

Forty subjects were chosen from respondents to an offer of training in assertiveness in exchange for participating in this study and payment of a nominal fee ($5). The request for subjects was made to a sample of undergraduate courses at a large southwestern university. Twenty subjects were randomly assigned to one of two assertiveness training groups, and the remaining 20 subjects served as a waiting list control group. Control group subjects were offered an assertiveness training group immediately after termination of the study. The facilitator of the assertiveness training groups was a part-time faculty member at the university where the study was conducted, who also had a private practice which included offering assertiveness training workshops. She had been conducting assertiveness training groups for over 4 years. Pretesting and posttesting activities were conducted by one of the junior authors who was unaware of whether subjects were in the treatment or control condition.

Instruments

The College Self-Expression Scale (CSES). The CSES (Galassi, Delo, Galassi, & Bastien, 1974) is a 50-item self-report measure of assertiveness in which respondents describe themselves using a 5-point scale. The 21 positively worded items are summed, and the 29 negatively worded items are reverse-scored and summed to yield a total assertiveness score. High scores reflect an assertive response pattern, whereas low scores indicate nonassertive responses. Extensive data on the reliability and validity of the scale are reported by Galassi et al. (1974) and Galassi, Hollandsworth, Radecki, Gay, Howe, and Evans (1975). Test-retest reliability coefficients over a 2-week period were .89 and .90. Construct validity was established with the Adjective Check List; concurrent validity with supervisors' ratings and behavioral measures of assertiveness in role-play situations were also obtained.

Counseling Outcome Inventory (COI). The COI is a self-report measure described by Hill (1975) as a process-orientated approach to the evaluation of the attainment of goals designated by a client as personally relevant and important. One of the weaknesses of group research is the tendency to utilize only global measures of change and thus to overlook the unique goals of each individual group member (Kiesler, 1971). A modified form of the COI was employed in this study to provide an individualized measure of change.

In using the COI, the experimenter developed with each subject a list of six traits on which he/she would like to change and the specification of a behavioral definition of each (i.e., "assertion" may be defined as initiation of conversations with co-workers before work). The experimenter insured that the traits listed by all subjects related to the topics to be covered by the treatment program.

The subjects ranked the chosen traits in the order of importance to them from "6" (most) to "1" (least) and gave a self-rating of their level of present (Pre) functioning on each, using a scale from "−3" (very dissatisfied) to "+3" (very satisfied). The product of the rank ordering and the self-rating provided a weighted score for each item, and the sum of the weighted scores yielded a total score.

The Learning Skills Questionnaire (LSQ). The LSQ is a 50-item self-report inventory using behaviorally oriented statements to assess important learning processes in the academic setting. Items are worded similarly to those in the CSES; they are keyed both positively and negatively; and the rating format (5-point scale) and scoring is the same. Hence, the LSQ resembles the CSES on all aspects except content. Data on the reliability and validity of the LSQ are reported by Schmeck, Ribich, and Ramanaiah (1977). Test-retest reliabilities over a 2-week period ranged from .79 to .88 for the various subscale scores. Also, the relationship between the LSQ and external tests of knowledge and performance in a paired-associate experiment were also investigated and found to provide substantial support for the validity of the LSQ.

The Jacobson-Kellogg Social Desirability Scale (J-K). The J-K scale (Jacobson, Kellogg, Maricauce, & Slavin, 1977) is a 68-item questionnaire that employs a true/false format and that measures need for approval. It was chosen because its more recent construction better reflects current social desirability values and because its greater length results in improved reliability over the Marlow-Crowne Scale. Further, it has been shown (Millham & Kellogg, in press) to be sensitive to bogus pipeline manipulation reflecting impression management.

Facilitator Ratings

Following the completion of the treatment, the facilitator was presented with the individual goals of each subject. She was not, however, given any information regarding the subject's self-ratings on those goals. The facilitator then estimated the amount each subject profited from the training on his/her own unique set of goals. Ratings were made on a scale ranging from 1 (the subject did not profit from this group) to 5 (the subject made substantial gains toward reaching his/her goals).

Experimental Treatment

The assertiveness training group met once a week for 2 hours for five sessions. There were two groups of 10 members each. Both groups covered identical topics and were run by the same facilitator. The facilitator's role at each session was to introduce the topic; to give a short explanation of its relevance to assertiveness; to facilitate the sharing of ideas, feelings, and experiences among group members; and to incorporate behavior rehearsal principles of modeling, practice, feedback, and reinforcement. Behavior rehearsal or small group exercises relating to the week's topic supplemented group discussion of the topic. Homework assignments included goal setting, practice of skills learned in the group, and frequency counts of various behaviors. In general, each session included the following:

Report of homework (after the first week),
Introduction to the week's topic by the therapist,
General group discussion of the topic,
Experimental and/or role-playing component,
Homework-generalization to outside group.

The topics included definitions of assertiveness and the concept of assertive human rights, assertive refusal, assertive initiation, self-esteem, expressing negative feelings, achievement, and competition.

Bogus Pipeline

When subjects were recruited, they were asked to supply their university student identification number, from which demographic data on each subject could be obtained through the university registrar's office. When subjects in the bogus pipeline condition came for their posttesting session, the experimenter gave them the following instructions:

> We are in the process of obtaining norms on a new way of administering self-report tests. Considerable work has been completed already and now we want to get data on a large group of college students. This technique differs from previous ones in that the person completes the rating scale while being monitored by a voice analysis device which can determine if the person is being completely accurate in his or her answers. As you may know, these types of devices are not always reliable. We've picked an instrument which is limited in that it doesn't work for everybody, but when it does work for a person, it does so completely. So before we go any further, we would like to see if the device will work for you. Here is a list of questions I'd like you to look over.

The experimenter then handed the subject a list of questions, the correct answers already having been gathered on the subject. She allowed the subject about 30 seconds in which to examine the question and then continued:

> After I switch on the microphone, I would like you to read each question distinctly into the microphone, followed by your answer. At the same time, mark your answers in pen or pencil in the appropriate spaces on the sheet. Answer truthfully to most of the items but deliberately make inaccurate statements on 2 or 3 of them without telling me which they are. When you have finished reading and answering the questions, I'll go down the hall to monitor and get your results.

The experimenter "switched on the microphone," which in fact terminated in an empty wall socket, and signaled the subject to begin. After the subject had completed the questionnaire, the experimenter took the answer sheet to another room where she checked the answers against the demographic data sheet and determined those items the subject had answered falsely. The experimenter then returned and informed the subject of the findings, thereby convincing him/her that the voice analyzer can distinguish between "true" and "false" responses.

The experimenter then proceeded to say:

> Since this experiment requires that you give answers to a questionnaire under conditions where inaccurate responding will show up on our monitor, it is necessary that you be given the opportunity to withdraw from the experiment without penalty; you will be given the same amount of course credit (if appropriate) and will, instead, complete the questionnaire without being monitored, under instructions to be as accurate as possible. Do you have any objections to performing the experiment? Now we will continue as before. After I switch on the microphone, I will turn on the tape

> recorder. Each question will be read twice so that if you don't understand it the first time, it will be clear the second. Be sure to be as accurate as possible in your responses, as discrepancies will show up on our monitor. Do you have any questions?

After giving the instructions, the experimenter switched on the microphone and the tape recorder and left the room. This constituted the bogus pipeline administration of the CSES, LSQ, J-K, and COI.

Procedures

Pretesting sessions for all subjects included the administration of the CSES, J-K, LSQ, and COI (Pre). All subjects were scheduled for an individual posttesting session within a week after the conclusion of the treatment groups. In the posttesting, through random assignment, half of the treatment subjects and half of the control subjects were exposed to the bogus pipeline deception, and the rest were not. All subjects were given the CSES, COI, LSQ, and J-K under the instructions to evaluate their present functioning on these scales (Post). Immediately afterward, subjects were asked to rate how they believed they were functioning when the study began for the LSQ, COI, and CSES (Then).

At the conclusion of the posttesting session, subjects were debriefed. At this time they were asked (when relevant) whether they believed the bogus pipeline deception. They were also informed that the bogus pipeline does not actually work, and their reactions to it were discussed. Completed data were obtained for 36 subjects. Two treatment subjects dropped out of school during the course of the study, and the investigators were unable to reach two of the control subjects to arrange for posttesting.

Results

One-way ANOVAS[2] of Post/Pre CSES scores ($F(1, 34) = 13.40$, $p < .001$) and, Post/Pre COI scores ($F(1, 34) = 2.66$, *n.s.*) yielded one significant treatment effect and a second effect which approached significance. These analyses suggest that treatment subjects self-reported greater increases in assertiveness and their own individual goals than did their control group counterparts. The same analysis substituting Then scores for Pre scores yielded significant treatment effects for the CSES ($F(1, 34) = 12.67$, $p < .01$) and marginally significant effects for the COI ($F(1, 34) = 2.92$, $.10 > p > .05$). Again, treatment subjects reported that they profited more than did control subjects.

Considerable evidence exists which suggests that Then/Post change scores correlate more highly with objective measures of change than do Pre/Post self-report measures (Howard & Dailey, 1979; Howard, Dailey & Gulanick, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck & Bray, 1979). The correlation of the facilitator rating of change with self-reported COI Pre/Post change was .25 (*n.s.*). Facilitator ratings of change were correlated with Then/Post self-reported change .52 ($p < .05$). A Hotelling-Williams test of the equality of two Pearson correlations computed among three variables in a single sample found the Then/Post self-reported change to be significantly more highly correlated ($Z = 3.86$, $p < .05$) with the facilitator's ratings than was Pre/Post self-reported change.

Thus, in the present study the pattern of results reflecting superior validity of the retrospective method of evaluating treatment change was

[2] A nontreatment group was included in this study to demonstrate and to insure the presence of a treatment effect. This finding permits analysis of the potential operation of response bias in treatment-dependent response-shifts. Although the design could be constructed as a 2 × 2 factorial, it is conceptually meaningful only as two one-way analyses of variance: one which analyzes treatment effects and the other which investigates the influence of response bias for treatment subjects. This is due to the fact that response-shift theory involves treatment-dependent changes, while it does not predict the reactions of control subjects. Therefore, predictions based upon a simultaneous consideration of treatment conditions with bogus pipeline conditions are inappropriate.

consistent with previous findings comparing a traditional Pre/Post self-report methodology with the Then/Post analysis. The effectiveness of treatment and the comparability of the present results to those obtained previously permit a direct test of the possible operation of response style differences impacting the retrospective measurement.

One indication that treatment subjects were simply complying with implicit task demands in making retrospective ratings would be treatment subjects demonstrating greater Pre/Then differences on the LSQ than their control group counterparts. Mean Pre/Then difference on the LSQ was 3.53 for treatment subjects and 9.68 for control subjects. Control subjects actually gave retrospective ratings that were slightly more in the direction of complying with implicit task demands than treatment subjects; however, these differences did not reach significance ($F(1, 34) = 2.22$, *n.s.*).

Table 2 presents the correlations (Spearman rho) of Pre J-K scores with Pre and Then CSES, COI, and LSQ scores for all subjects, treatment subjects alone, and control subjects alone. For those measures related directly to the treatment intervention (CSES and COI), the results indicated a low to moderate relationship between social desirability responding and self-reports of assertiveness. Following treatment, retrospective self-reports (Then scores) of pre-intervention assertiveness demonstrated a diminished relationship to social desirability responding, indicating that such retrospective measures of pretreatment states (Then/Post method) were not more biased by self-deceptive and impression management responding than those obtained prior to treatment. In fact, they appear less biased than the pretreatment measures. In addition, the relationship between social desirability responding and retrospective self-report of assertiveness without a treatment intervention did not differ from those obtained on initial testing.

Taken together these results indicate that the effect of treatment in the present study not only increased assertiveness but also reduced social desirability responding in retrospective measures of pretreatment assertiveness. Analysis of a nontreatment-related self-report measure (LSQ) reflected no impact of treatment on the relationship between the LSQ and social desirability responding on the retrospective measure and no difference in this relationship for the Pre and Then measures. This provides further evidence for the specificity of the treatment effects in reducing social desirability responding on retrospective measures. However, the restricted sample size of these groups virtually precluded the possibility of the differences between these correlations reaching statistical significance, and consequently these comparisons were not attempted.

Several analyses were undertaken to ascertain if response-shifts were influenced significantly

Table 2

Correlations (Spearman Rho) of Pre J-K Scores with Pre and Then Self-Ratings on the CSES, COI, and LSQ[a]

Pre J-K with:	N	CSES	COI	LSQ
Pre Scores, All Subjects	36	.34	.21	.37
Then Scores, All Subjects	36	.21	.22	.38
Then Scores, Treatment Subjects	18	.21	-.03	.29
Then Scores, Control Subjects	18	.33	.18	.47

[a] While it would have been appropriate and informative to compare correlations from each experimental condition (e.g. treatment/bogus pipeline subjects), the limited number of subjects in such an analysis precluded interpretable analysis.

by impression management, reflecting compliance with implicit task demands to demonstrate improvement to the evaluator. The bogus pipeline technique was utilized to investigate such impression management, impacting observed response-shifts in pretreatment to posttreatment (retrospective) evaluation of pretreatment states.

The J-K was administered at posttest to ascertain if the bogus pipeline deception was effective. A one-way ANOVA of Post J-K scores found a significant ($F(1, 34) = 6.30, p < .05$) effect such that bogus pipeline subjects endorsed reliably fewer socially desirable responses (mean = 21.0) than their non-bogus pipeline counterparts (mean = 29.7). This finding demonstrates the effectiveness of the bogus pipeline manipulation, since comparison of the groups' Pre J-K scores revealed no differences ($F(1, 34) = .30$, *n.s.*). Mean Pre J-K rating for subjects who were later assigned to the bogus pipeline was 23.1, while the mean Pre rating for the non-bogus pipeline group was 25.1

Given the effectiveness of the bogus pipeline manipulation in demonstrating the operation of impression management responding for the subjects in this study, an analysis was undertaken to determine if such responding was operating and influencing response-shift effects. Pre/Then differences were calculated for treatment subjects on the COI, CSES, and LSQ. Mean Pre/Then differences and results of tests of differences between bogus pipeline and non-bogus pipeline treatment subjects are presented in Table 3.

There was no evidence for an effect of bogus pipeline manipulation on response-shifts for the treatment subjects. In other words, there is no evidence for a significant operation of impression management influencing shifts in retrospective evaluation of pretreatment states. Such impression management would be expected to be operating on the retrospective evaluations if subjects were yielding to task demands in order to appear improved following treatment.

Discussion

It has been suggested that enhanced familiarity with the goals of intervention and personal effort and involvement in treatment might accentuate social desirability responding-compliance with implicit task demands to demonstrate improvement on retrospective self-reports following treatment. The results of the present study do not support that hypothesis. The correlations of social desirability scores with pretreatment self-reports of treatment-related measures (CSES and COI) were higher than those obtained between social desirability scores and retrospective self-reports on the same measures. It would appear, within the context of the intervention procedures employed in the present study, that social desirability responding is actually diminished in utilizing the Retrospective-Pre (Then) methodology. This positive effect of treatment can be seen in two other sets of results where, for nontreatment control subjects and for measures on a nontreatment-related

Table 3
Pre-Then Differences on the COI, CSES, and LSQ Bogus Pipeline and Non-Bogus Pipeline Treatment Subjects

	Bogus Pipeline (N=18)		Non-Bogus Pipeline (N=18)		
Scales	Mean	S.D.	Mean	S.D.	t
CSES	-2.0	12.3	-6.0	14.2	.44
COI	4.9	16.5	10.6	18.6	-.59
LSQ	2.1	15.1	5.1	14.0	-.42

variable (LSQ), the correlation between social desirability scores and self-reports on the various measures remained very similar to those obtained between initial (pre-treatment; pre-waiting list) measurements and social desirability scores. Although the sample size did not permit an unequivocal statement that the retrospective measures following treatment were statistically less significantly biased with general social desirability responding than traditional pre-measures, the pattern of findings supports that interpretation. The results do indicate clearly that the assumption of greater social desirability bias in retrospective self-reports is not tenable.

An additional set of analyses was conducted to investigate further the impact of response bias on retrospective measures of change. Differences between scores obtained under bogus and non-bogus pipeline reflect the operation of impression management and hence should be indicative of attempts to meet implicit task demands to demonstrate improvement to the evaluator. The shifts in self-report measures of pretreatment states that occur using the retrospective methodology were found to be no different when obtained under bogus pipeline than under non-bogus pipeline conditions. These findings indicate that there is no evidence for the operation of impression management influencing the shifts in evaluation obtained in employing the retrospective methodology.

The final indication that retrospective measures were not unduly influenced by generalized compliance with task demands came from the results of the LSQ. The fact that there was no response-shift effect on this measure that was unrelated to the content of the treatment intervention suggests that subjects were reporting treatment-induced changes rather than simply providing the experimenter with a favorable set of results when making Then/Post ratings.

The present findings add to the existent literature that finds Then/Post self-report indices of change to be more highly correlated with objective measures of change than are Pre/Post self-report indices. These demonstrations of superior concurrent validity of Then/Post ratings demand that researchers engaged in measuring change with self-report instruments, especially in the areas of program effectiveness, consider the probable impacts of response-shift bias and adjust their research strategies accordingly. Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) recommend the use of a Retrospective Pretest-Posttest design that allows the investigator to determine if a substantial response-shift has occurred and, if so, to employ the appropriate procedures to attentuate the source of bias in the results.

More broadly, theorists (Cronbach & Furby, 1970; Linn & Slinde, 1977) have noted that the measurement of change is a complex and problematic endeavor. Cronbach and Furby have recommended as a viable alternative that, with random assignment of subjects to conditions, an analysis of posttest scores will avoid the difficulties associated with measuring change. Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) have demonstrated that when response-shift bias is present, posttest-only comparisons are inappropriate. Consequently, one is forced to measure change, not because retrospective measures allay Cronbach and Furby's concern any better than traditional pretests, but rather because their alternative to measuring change is no longer viable.

Finally, given the extent and pervasiveness with which response-shift bias has been documented and the superiority of Then/Post over Pre/Post methodology in evaluating training interventions, it is strongly recommended that researchers begin to collect retrospective pretest data along with the traditional Pre and Post self-ratings. Use of retrospective measures, which provide a more sensitive assessment of a subject's perspective of personal change, will add yet another valuable dimension to current research efforts. That is, when one goal of a treatment intervention is that of increasing participants' understanding of their level of functioning on a specific dimension, making a comparison of pretest and retrospective pretest scores on that dimension might provide researchers with the means to assess whether that goal has been met. Ironically, the same response-shift which, if ignored, serves to bias evaluation research, has the potential, when measured, to provide further useful outcome information.

References

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Cronbach, L. J., & Furby, L. How we should measure "change"—or should we? *Psychological Bulletin*, 1970, *74*, 68-80.

Galassi, J., Delo, J., Galassi, M., & Bastien, S. The college self-expression scale: A measure of assertiveness. *Behavior Theory*, 1974, *5*, 165-171.

Galassi, J., Hollandsworth, J., Radecki, J. C., Gay, M., Howe, M. R., & Evans, C. Behavioral performance in the validation of an assertiveness scale. *Behavior Therapy*, 1976, *7*, 447-452.

Hill, C., A process approach for establishing counseling goals and outcomes. *Personnel and Guidance Journal*, 1975, *53*, 571-576.

Howard, G. S., & Dailey, P. R. Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, 1979, *64*, 144-150.

Howard, G. S., Dailey, P. R., & Gulanick, N. A. The feasibility of informed pretests in attenuating response-shift bias. *Applied Psychological Measurement*, 1979, *3*, 481-494.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, 1979, *3*, 1-23.

Howard, G. S., Schmeck, R. R., & Bray, J. H. Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, 1979, *10*, 305-315.

Jacobson, L. I., Kellogg, R. W., Maricauce, A., & Slavin, R. S. A multidimensional social desirability inventory. *Bulletin of the Psychonomic Society*, 1977, *9*, 109-110.

Jones, E. E., & Sigall, H. The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, 1971, *76*, 349-364.

Kiesler, D. K. Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: John Wiley, 1971.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, 1977, *47*, 121-150.

Millham, J., & Kellogg, R. W. Need for social approval: Impression management or self-deception? *Journal of Research in Personality*, in press.

Schmeck, R. R., Ribich, F., & Ramanaiah, N. Development of a self-report inventory for assessing individual differences in learning processes. *Applied Psychological Measurement*, 1977, *1*, 413-431.

Author's Address

Request for reprints or further information should be sent to George Howard, Psychology Department, University of Houston, Houston, TX 77004.