The Feasibility of Informed Pretests in Attenuating Response-Shift Bias

George S. Howard, Patrick R. Dailey, and Nancy A. Gulanick
University of Houston

Response-shift bias has been shown to contaminate self-reported pretest/posttest evaluations of various interventions. To eliminate the detrimental effects of response shifts, retrospective measures have been employed as substitutes for the traditional self-reported pretest. Informed pretests, wherein subjects are provided information about the construct being measured prior to completing the pretest self-report, are considered in the present studies as an alternative method to retrospective pretests in reducing response-shift effects. In Study 1 subjects were given a 20-minute presentation on assertiveness, which failed to significantly improve the accuracy of self-reported assertiveness. Other procedural influences hypothesized to improve self-report accuracy—previous experience with the objective measure of assertiveness and previous completion of the self-report measure—also were not related to increased self-report accuracy. In a second study, information about interviewing skills was provided at pretest using behaviorally anchored rating scales to participants in a workshop on interviewing skills. Response-shift bias was not attentuated by providing subjects with information about interviewing prior to the intervention. Change measures which employed retrospective pretest measures demonstrated somewhat higher (although nonsignificant) validity coefficients than measures of change utilizing informed pretest data.

Although the measurement of change is important in virtually all areas of psychological research, it is an endeavor fraught with problems (Cronbach & Furby, 1970; Linn & Slinde, 1977). In the evaluation of training and treatment interventions, change is frequently measured by means of subject self-reports in pretest-posttest designs, such that the degree of change from pretest to posttest for treatment subjects, relative to their control group counterparts, is assumed to reflect the value of an intervention. With random assignment of subjects, this design (Design 4; Campbell & Stanley, 1963) was thought to provide internally valid results. However, Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) have recently reported the problem of response-shift bias, which is a threat to internal validity in evaluation studies employing self-report measures. The problem of response-shift bias is handled by substituting retrospective pretest (Then) ratings for traditional pretest (Pre) ratings in the analysis of change (see Howard et al., 1979, for an explanation of response-shift effects and of how retrospective ratings are obtained).

Four potential causes for the Then–Pre self-report differences which have been attributed to response-shift bias are (1) memory distortion, e.g., forgetting; (2) subject's response-style effects, e.g., subject acquiescence, social desirability; (3) insufficient insight or awareness of one's own level of functioning with respect to a particular construct; and (4) insufficient under-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 481–494

standing of the construct, e.g., assertiveness, interviewing skills, dogmatism.

Howard et al. (1979) in Study 5 investigated the potential for memory distortion effects and concluded that at posttest (Post), subjects' memory of their pretest ratings was not systematically different from their Pre ratings but that Then ratings were reliably different from both Pre and Memory ratings. Howard and Dailey (1979) in Study 2 arrived at similar conclusions regarding the influences of memory distortions. Regarding subject response style, Howard, Millham, Slaton, and O'Donnell (under review) considered the possibility that response-shift effects may simply be subject acquiescence artifacts. That is, subjects may not believe they have changed a great deal, but their desire to provide the experimenter with a favorable set of results leads them to lower their Then ratings. However, no support for this explanation was found.

The other two possible causal determinants suggest that it is because the intervention enables subjects to attain a greater understanding of either or both the constructs being measured and/or their level of functioning on those dimensions that their Then self-report ratings differ from their Pre self-report ratings. Gaining a greater awareness of one's level of functioning might be truly treatment dependent and thus may require a retrospective approach to counteract response-shift bias. However, in the case of insufficient understanding of the construct, investigators might provide subjects with enough information about the construct prior to the pretest to allow them to make a more accurate assessment of their pretest level of functioning. This method of rapprochement would be welcomed by many investigators, since utilizing "informed" pretests to lessen response-shift bias would eliminate the need for retrospective ratings. The two investigations reported below take differing approaches to providing subjects with information about the construct designed to improve the accuracy of their self-report pretest ratings, in an attempt to attenuate response shift bias effects.

STUDY 1

This study sought to determine if a short information-giving session can improve the accuracy of self-report assessments. The construct utilized in this study was assertiveness. The effects of two procedural influences that may affect self-report pretest accuracy (previous experience with the self-report instrument or the behavioral measure) were also investigated.

Method

Subjects

Eighty-eight undergraduates participated in the present study for credit toward the requirements of their introductory psychology course. Complete data sets were obtained for 83 students who were included in the analyses.

Instruments

The College Self-Expression Scale (CSES). The CSES (Galassi, Delo, Galassi, & Bastien, 1974) is a 50-item self-report measure of assertiveness on which respondents describe themselves using a 5-point scale. Scores can range from 0 to 200, with higher scores reflecting a more assertive response pattern. Extensive data on reliability and validity of the scale are reported by Galassi et al. (1974) and Galassi, Hollandsworth, Radecki, Gay, Howe, and Evans (1975).

Objective Measure of Assertiveness (OMA). The objective measure of assertiveness consisted of each student's verbal responses to 8 taped stimulus situations. The stimulus situations were the same as or similar to those used by Eisler, Miller, and Hersen (1973). Each student was instructed to listen to each stimulus statement and were to respond verbally using the actual words he/she would use if the situation were really happening. All responses to the stimulus statements were audiotaped, coded, and later rated for assertiveness by two trained raters using the Rathus Assertiveness Scale (Rathus,

1973). The mean assertiveness rating for each student was employed in the present study.

Raters

Two senior psychology majors with previous experience using the Rathus Assertiveness Scale (Rathus, 1973) were used as raters of the OMA tapes. The interrater reliability for this instrument in the present study was .89.

Assertive Information and Placebo Sessions

Each student attended either a session which provided information about assertiveness (Assertive Information) or an attention placebo session wherein a number of social activities on campus were discussed (Placebo). Both presentations lasted about 20 minutes. In the Assertive Information session, the experimenter presented didactic information about assertiveness and nonassertiveness, answered students' questions, and gave examples of assertive and nonassertive behavior. In order to check that students had achieved an accurate understanding of the concept of assertiveness, the experimenter asked each student to give an example from his/her own life illustrating assertive and nonassertive behavior and provided further clarification if necessary. The Placebo session followed a similar format of presenting the student information on campus activities, answering questions, giving several examples of possible activities, and asking the students to share some activities in which they might engage that semester. An advanced graduate student conducted all Assertive Information and Placebo sessions.

Procedure

Students were randomly assigned to one of eight groups. These groups differed from one another in the order in which the measures were administered and whether students attended the Assertiveness Information or the Placebo session. Table 1 shows the order in which various activities were completed by the students in each condition. All students were instructed to carefully and honestly complete the tasks of the study. For those who completed the CSES a second time, further instructions were given after the Assertive Information or Placebo ses-

Table 1
Number of Subjects and Order of Activities in the Experimental Conditions

Condition	N	Order of Activities				
		1	2	3	4	5
1	11	CSES	OMA	Assertive Information	CSES	___
2	10	CSES	OMA	Placebo	CSES	___
3	11	___	OMA	Assertive Information	CSES	___
4	10	___	OMA	Placebo	CSES	
5	10	CSES	___	Assertive Information	CSES	OMA
6	10	CSES	___	Placebo	CSES	OMA
7	10	___	___	Assertive Information	CSES	OMA
8	11	___	___	Placebo	CSES	OMA

sion to respond as they now felt was true, whether that meant their responses were similar to or different from their earlier responses. Total testing time, which included Assertive Information or Placebo for each student, was approximately 50 minutes. Students were then debriefed and given course credit.

Data Analysis

Four possible influences that might increase the accuracy of self-report ratings can be identified: (1) student having received Assertive Information; (2) student having received an exhortation to try to provide accurate ratings on the CSES, which took place at the end of each Assertive Information and Placebo session; (3) student having previously completed the objective measure; and (4) student having previously completed the self-report measure. Each self-report rating may be characterized as either possessing or not possessing each of these four potential influences. For example, the first time a student in Condition 1 completed the CSES he/she would not have had the benefit of any of these four potentially helpful influences. However, the second time that student completed the CSES he/she might have been influenced by all four effects. The crucial question, then, was whether any of these factors helped to improve the accuracy of CSES ratings.

The correlation coefficient is a standard index of agreement. Correlations represent the degree to which scores on two measures are proportional when expressed as deviations from their means. Accuracy implies a further condition that the absolute values of the scores on these measures be equivalent. Since simple correlation coefficients would not test for this latter condition, correlation was deemed an inadequate analysis, and the authors developed the following procedure wherein a subject's OMA rating was used as a basis from which to predict an expected CSES score.

The functional relationship between the CSES and OMA needed to be determined; this was accomplished by using data from treatment subjects in a previous study (Study 3 of Howard et al., 1979). The relationship between Post-CSES and Post-OMA scores was found to be CSES = 13.8 OMA + 69.44. Similarly, the line of best fit for the Then-CSES with Pre-OMA scores was described by CSES = 13.3 OMA + 52.04. The average of these two regression lines was CSES 13.5 OMA + 60.74, which was assumed to be the relationship between these two measures. These comparisons assumed that OMA ratings for the data utilized to compute Predicted scores were equivalent to the OMA ratings in the present study.

Since the raters and times of rating were different, some indication of the equivalence of scales was necessary. One author randomly selected 20 taped responses from each study, and the two remaining authors rated them without knowing from which study they came or how they had been scored. The mean OMA rating for Study 3 of Howard et al. (1979) was 3.78; the authors' rating of these same tapes was 3.70. In the present study, however, the mean OMA rating for selected tapes was 3.98, whereas the authors' mean rating of these responses was 3.87. There were no statistical differences between judges' and authors' OMA ratings in the two studies. By entering each student's OMA score, a Predicted CSES score was obtained for that student.[1] The difference (D-score) between this Predicted CSES score and the student's observed CSES score served as the measure of CSES accuracy. Obviously, the smaller the value of D, the greater the accuracy of the CSES rating. If a sample of D-scores were accurate, the

[1]It is important to use both pretreatment and posttreatment ratings in generating the equations to yield Predicted CSES scores (1) because subjects for Study 3 of Howard et al. (1979) were selected for that study if they were highly feminine on the Bem Sex-Role Inventory and (2) due to the relationship between sex-role orientation and assertiveness, the distribution of Pre-OMA scores was skewed toward the nonassertive end of the scale. Since no selection technique was employed in the present study, OMA scores covered a wide range.

mean of the sample would be zero (i.e., positive and negative scores tending to cancel out). If, however, some source of systematic bias were present, mean D-scores would be positive if students' CSES ratings were overestimates of their assertiveness and negative if they were underestimates.

Results

CSES scores were overestimates of the Predicted CSES ratings. (Mean D-score = 19.53; $t(81) = 9.31, p < .001$). Table 2 presents a breakdown of the D-Scores for CSES ratings after the Assertive Information and Placebo sessions grouped by the presence or absence of assertive information, prior OMA experience, and prior CSES experience. Statistical comparisons could not be made for the effect of the accuracy instruction.

CSES accuracy was not significantly improved by assertive information or prior experience with either the OMA or CSES. Paradoxically, one factor (OMA prior experience) was actually related to a slight, though nonsignificant, decrease in CSES accuracy. If subject response-style effects were feared to contaminate Then scores in Study 3 of Howard et al. (1979), one might choose to generate Predicted CSES employing the equation CSES = 13.8 OMA + 69.44, which is obtained by using Post-CSES and Post-OMA scores only. If this formula were to be employed, the CSES scores in the present study would still represent overestimations of Predicted CSES scores, although they would be viewed as somewhat more accurate.

Discussion

The CSES scores in this study represented overestimates of the predicted CSES values. This is in agreement with the findings of Studies 3 and 4 of Howard et al. (1979), which found that self-reported Pre assertiveness scores were inflated relative to response-shift-free Then scores. These inaccuracies were present in spite of a 20-minute information session on assertiveness. Although the information session might have taught the students about the concept of assertiveness, it may have failed to sensitize them enough to their own level of assertiveness to enable them to provide an accurate assessment of their functioning. Similarly, asking the students for one example in the information session may not have had enough impact for them to relate personally to the construct. In addition,

Table 2
Mean D-Scores for CSES Ratings Grouped by the Three Potential Accuracy Influencing Factors

Factor and Group	Mean	S.D.	t
Assertive Information			
Yes	16.37	22.61	
No	20.54	26.10	-1.05
Prior OMA Experience			
Yes	21.05	25.16	
No	15.85	24.87	1.32
Prior CSES Experience			
Yes	15.37	24.42	
No	21.54	25.24	1.57

the information session may not have allowed students to generalize the concept of assertiveness from the specific examples given during that session to a wide range of situations included in the self-report and the objective measures of assertiveness. Gormally, Hill, Otis, and Rainey (1974) noted that generalization is difficult to achieve without sufficient practice across various situations to enable subjects to deduce practical guidelines in assertiveness. Finally, there may have been variance in the presentation on assertiveness provided individually to the students.

Ironically, prior experience with the OMA was related to a slight decrease in CSES accuracy. When completing the OMA, students received no feedback on the appropriateness of their responses. In the absence of such feedback, students may have inappropriately evaluated their nonassertive responses as "correct," thus erroneously overestimating their assertiveness.

In order to further investigate the potential of an "informed pretest" in eliminating response-shift bias, a second study was undertaken wherein the "information" was given in a standardized behaviorally anchored form, and an actual analog interview pretest rather than a taped stimulus was employed. The relative efficacy of an informed pretest versus the conventional pretest in evaluating the effectiveness of a training intervention was also observed.

STUDY 2

This study investigated the influence of information about interviewing skills in attenuating response shift in the self-reported evaluation of a workshop designed to improve interviewer skills in selection interviews in industrial contexts. The program consisted of a week-long seminar that heavily emphasized interviewing skills practice plus didactic training.

In the industrial appraisal literature two general methods for collecting performance data might be classified as the summative and behavior-specific approaches (Schwab, Heneman, & DeCotis, 1975). The summative approach defines performance using poorly delineated levels of functioning (e.g., below average, excellent, poor). While variants of this approach may be designed to collect performance using a more multidimensional approach rather than a single global composite assessment, basically these exist at a rather ambiguous level by virtue of the different interpretations of behavior used by the raters.

The behavior-specific approaches include more attention to actual examples of behavior, are more multidimensional in their approach, and attempt to reduce the ambiguity or variance in interpretations that are possible across raters through empirical categorization and scaling. Thus, users of this type of performance appraisal method are provided information from which they may infer appropriate performance category definitions and more reliably assess functioning levels.

In previous studies, Howard and Dailey (1979) utilized a seven-item self-report measure to assess participants' skill levels on six target interviewer dimensions (questioning techniques, structured approach, supportive attitude, rapport building, active listening, and relevant material) plus a rating of overall effectiveness. Subjects had responded to each item by utilizing a scale from 1 (to a very little extent) to 9 (to a very great extent) to indicate the extent to which they felt they possessed the six types of interviewer skills and were capable of conducting an effective interview. However, the subjects had received no explanation of the dimensions, nor were they provided with examples of what might be appropriate or inappropriate behaviors for any dimension. Using this summative style self-report instrument, Howard and Dailey (1979) found substantial response-shift contamination that significantly lowered the concurrent validity of the self-reported indices of change. It was speculated that if behaviorally anchored scales were to be employed, subjects would be provided with appropriate information about the six interviewer performance dimensions at pretest,

thus allowing them to operationally define each construct. This procedure would then be expected to improve the accuracy of their pretreatment ratings and thereby attentuate response-shift effects.

Method

Subjects

Twenty participants in a week-long interviewing institute served as subjects in this study. They possessed from 1 month to 18 years prior interviewing experience (median = 6 months) and were heterogeneous with respect to parent company and geographic location. All were engaged in jobs which involved conducting job interviews with potential job candidates or subordinates.

Program

The Institute was conducted over 5 days (36 hours of training) and involved training in the interviewer skills mentioned above. The program was focused around three videotaped practice interviews with undergraduates who were currently seeking employment and three small group play-back sessions wherein three interviewers reviewed the videotapes of their interviews with a member of the Institute staff whose duty it was to critique each interviewer's performance.

Instrument Development

Eleven college seniors in an undergraduate interviewing course took part in the development of the behaviorally specific performance appraisal scales (BSS) for use in the present study. The procedure followed was that of Smith and Kendall (1963). The process of scale development included (1) identification of criterion dimensions; (2) collection of critical interviewing behaviors representing these dimensions; (3) retranslation; (4) scaling; and (5) final editing and scale construction. Specifically, the students together with one of the authors first identified and agreed upon the six dimensions of interviewer behavior mentioned above. Critical incidents (Flanagan, 1954) were collected by the students from personal experience and from viewing 40 archival Institute tapes of selection interviewing. The focus was on the behaviors of the interviewers in these six criterion areas. It was the task of the students to describe the behaviors observed on the tapes that were particularly noteworthy (either positive or negative). Included in the description of an incident was a description of the circumstances, the behavior of the interviewer, its consequences to the interview, and an indication of the criterion area in which the behavior fit. This phase resulted in more than 300 critical interviewing behaviors (172 usable) across the six interviewer performance dimensions.

The retranslation step involved feeding back to the students the incidents they had previously collected, with their task being to read the incident and individually to assign it to the most appropriate of the six dimensions. This represented a quality control step in BSS development; a cutoff of 55% (6 of 11 students) or greater agreement was used to insure that there was reasonable agreement regarding the dimension to which each particular incident belonged. Items were discarded when the judges were unable to agree on appropriate category assignment. The judges were again presented with the list of interviewer behaviors ($N = 110$) grouped according to the six performance dimensions. Each rater's task this time was to independently provide a scale value representing the favorability of each incident using a 9-point scale ranging from 1 (extremely unfavorable) to 9 (extremely favorable). Means and standard deviations were calculated for each incident, and items whose standard deviations exceeded 1.75 were discarded. The mean for each of the remaining incidents ($N = 52$) served to anchor the dimension; and when fully constructed, a dimension had behavioral examples at several scale points across the dimension.

Procedure

During the introductory session of the Institute, participants were asked to examine the BSS scales and to indicate their level of interviewing skill on the seven items (Pre). Before the actual training began, each participant interviewed a job applicant and all interviews were recorded on videotape. This served as the behavioral Pre measure (Beh Pre). Interviews were conducted, with undergraduates serving as interviewees, and lasted for 30 minutes or less. The Institute proceeded as planned, and part of the conclusion of training was the completion of a final videotaped interview, serving as the behavioral Post Measure (Beh Post). Following all training and feedback sessions, the participants again completed the BSS, answering each item once as they felt they were at that point in time (Post) and once as they felt they had been at the beginning of the workshop (Then). Subjects were instructed to feel free to agree with their Pre self-report ratings if they felt that they were accurate or to disagree with those ratings if they now saw them as being inaccurate. The participants were provided with a brief explanation of the purpose of the study, and the Institute was concluded.

Six upper level undergraduates, who themselves were in a similar interviewing class, served as judges to rate the videotapes. An 18-hour training period was conducted by one of the authors wherein the scale dimensions were explained and discussed. Archival Institute tapes were viewed, rated, and discussed until all raters were comfortable that they understood the dimensions to be assessed. The 40 taped interviews were coded, randomized, and shown to the six raters, who independently assessed each interviewer's skill along the seven criteria (six specific plus one overall) using a scale from 1 (to a very little extent) to 9 (to a very great extent). These were identified as judges' skill ratings.

Additionally, based upon recommendations by Fear (1973) and Banaka (1971), the authors developed a set of behavioral composite variables with favorable and unfavorable interviewer behaviors pertaining to each interviewing dimension. Judges were instructed to tabulate incidents of these behaviors while viewing the tapes, and these entries were used to form a linear composite for each of the six interviewing dimensions (behavioral incidents). For example, the supportiveness composite was formed by totaling the number of appropriate interruptions plus the number of agreement statements minus the number of inappropriate interruptions.

Reliabilities for the seven judges' skill ratings ranged from .89 to .96, and reliabilities for composites of the six behavioral incidents ranged from .74 to .95 (see Table 4). The mean ratings of the judges were employed as the unit of analysis for both the skill ratings and behavioral incidences.

Results

Due to the number of dependent variables (seven BSS scales, seven judges' skill ratings, and six behavioral incidents scales) and their pattern of high intercorrelations, a multivariate analysis of variance (MANOVA) was employed to test for treatment effects. Table 3 presents mean Pre, Post, and Then self-report scores along with the results of univariate F-tests of Pre/Post and Then/Post differences for this study. Subjects reported significant before- to after-workshop changes, whether measured by the Pre/Post ratings (multivariate F (7, 13) = 16.92, $p < .001$) or the Then/Post ratings (multivariate F (7, 13) = 9.11, $p < .001$). Significant differences were found for all subsequent univariate Pre/Post comparisons and all Then/Post comparisons. A significant treatment effect was found for Pre/Post comparisons of judges' skill ratings (multivariate F (7, 13) = 18.85, $p < .001$) and behavioral incidents (multivariate F (6, 14) = 26.96, $p < .001$). Table 4 presents mean Pre and Post judges' skill ratings and behavioral incidents in addition to the results of 13 univariate F-tests.

In an attempt to ascertain the relative effectiveness of the Then/Post self-report relative to the Pre/Post self-report approach, both

Table 3
Mean Pre, Post, and Then Ratings and Results of Univariate F-Tests of Pre/Post and Then/Post Differences for Each Self-Report Item

Dimensions	Pre	Post	Then	F-Ratio (1,19 df) Pre/Post	Then/Post
Questioning Techniques	6.20	7.85	5.10	15.09**	53.47**
Structured Format	4.60	8.20	4.45	76.00**	54.67**
Supportiveness	6.50	7.45	6.00	9.28**	11.59**
Rapport	7.10	7.85	6.55	4.67**	5.73**
Actively Listened	6.95	7.70	5.50	10.82**	25.13**
Relevant Material	5.30	6.90	4.80	10.05**	16.79**
Overall	4.50	6.50	3.95	45.56**	49.29**

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

Pre/Post and Then/Post self-report change scores were correlated with Pre/Post changes in judges' skill and behavioral incidents ratings. Table 5 presents the results of these correlations and tests (Hotelling-Williams test of $\varrho_{12} = \varrho_{13}$) of the equality of two Pearson correlations computed among three variables in a single sample. On the judges' skill ratings, six of the seven comparisons favored the Then/Post approach, with these differences reaching significance (one tailed) in one instance. The mean correlation of changes in judges' skill ratings with Pre/Post self-report change was .17, whereas the mean correlation of changes in judges' skill ratings with Then/Post self-report change was .38. With regard to the correlations with the behavioral incidents in three cases (none of which would have reached significance had the test been two tailed), the Pre/Post self-report approach was superior; this trend was reversed in the other three cases (one significant). The mean correlation of changes in behavioral incidents with Pre/Post self-report change was .04, and the mean correlation of changes in behavioral incidents with Then/Post self-report change was .08.

Finally, the magnitude of response shift (Pre–Then self-report differences) was compared for a behaviorally specific scale (present study) versus summative self-report measures (Studies 1 and 2 of Howard & Dailey, 1979).[2] Pre–Then differences would be reduced if (1) information supplied to subjects about the characteristics of selection interviewing skills was helpful in reducing response shift and (2) the BSS adequately supplied that information. Table 6 presents mean Pre–Then differences and the results of univariate *F*-tests comparing the three aforementioned studies. As expected, there was

[2]The Institute programs were essentially identical, and data collection procedures differed only with respect to the use of the BSS in the present study. Since the assignment of participants to programs was in no way random, concern should be with the initial equivalence of the three samples for these comparisons. Since the Pre self-report was a behaviorally specific scale for the present study but summative for the other two, comparisons along these dimensions would be suspect. Similarly, no judges' ratings were available from Study 1 of Howard and Dailey (1979), whereas different sets of raters were employed for Study 2 of Howard and Dailey (1979) and the present study. Consequently, comparisons of Pre scores of judges' ratings would be of dubious value. Hence, the initial equivalence of the three samples will remain untested. The authors recommend that subsequent comparisons among studies be viewed as anecdotal only, and all conclusions will be stated with extreme caution.

Table 4
Mean Pre and Post Scores and Results of Univariate F-Tests for Judges Skill Ratings and Behavioral Incidents of Videotaped Interviews

Dimensions	Pre	Post	F(1,19)	Rating Reliability[a]
Questioning Techniques				
Skill Ratings	4.16	6.56	65.31**	.96
Behavioral Incidents	8.82	2.67	52.36**	.91
Structured Format				
Skill Ratings	3.59	6.55	42.83**	.93
Behavioral Incidents	6.42	25.55	70.56**	.92
Supportiveness				
Skill Ratings	4.75	6.97	35.84**	.91
Behavioral Incidents	25.23	37.87	28.92**	.84
Rapport				
Skill Ratings	4.38	5.92	26.01**	.87
Behavioral Incidents	7.07	13.75	5.33*	.74
Actively Listened				
Skill Ratings	4.30	6.56	75.44**	.94
Behavioral Incidents	-.70	1.25	12.69**	.82
Relevant Material				
Skill Ratings	4.18	6.29	35.51**	.89
Behavioral Incidents	21.43	39.67	25.21**	.95
Overall Skill Ratings	4.01	6.28	53.13**	.91

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

[a] Reliabilities were corrected for attenuation using the Spearman-Brown procedure.

no difference in the magnitude of response shift when the data from the two studies which employed the summative self-report index were compared (multivariate F (7, 54) = .698, p = .67). Similarly, the differences between Study 1 of Howard and Dailey and the present study (multivariate F (7, 54) = 1.51, p = .18) were not significant, nor did the differences between Study 2 of Howard and Dailey and the present study reach significance (multivariate F (7, 54) = 1.26, p = .28).

DISCUSSION

All measures of change (Pre/Post self-report, Then/Post self-report, Pre/Post judges' skill ratings, Pre/Post behavioral incidents) found significant before-to-after changes in the work-

Table 5
Correlations of Change in Judges Skill Ratings and Behavioral Incidents with Pre/Post and Then/Post Self-Report Change and Results of Tests Differences Between Correlations

Dimensions	Self-Report Pre/Post	Self-Report Then/Post	Hotelling-Williams
Questioning Techniques			
Skill Ratings	.14	.36	.66
Behavioral Incidents	-.04	-.46	3.22
Structured Format			
Skill Ratings	.66	.43	1.37
Behavioral Incidents	.61	.37	1.26
Supportiveness			
Skill Ratings	.07	.47	3.05*
Behavioral Incidents	-.06	.38	1.90
Rapport			
Skill Ratings	.11	.36	.99
Behavioral Incidents	-.09	-.22	.27
Actively Listened			
Skill Ratings	-.34	-.04	.67
Behavioral Incidents	-.17	.14	.65
Relevant Material			
Skill Ratings	.14	.41	1.55
Behavioral Incidents	.00	.38	3.07*
Overall	.42	.64	1.34

*$p < .05$

shop participants. These findings replicate both studies reported by Howard and Dailey (1979).

Regarding the anticipated higher validity of Then/Post indices of change to the Pre/Post self-report measures for predicting criterion change, partial support was found. As found previously by Howard and Dailey (1979), mean correlation of change in judges' skill ratings with Then/Post self-reported change was superior to the same comparison using the Pre/Post self-report changes (.38 versus .17). Change in behavioral incidents was not related to either Pre/Post or Then/Post self-reported change. This finding is somewhat surprising, given Howard and Dailey's relationships between Pre/Post self-reported change with change in behavioral incidents (−.05) and Then/Post change with change in behavioral incidents (.33). Several reasons for the failure to replicate might be that (1) different judges were used in the present study; (2) several new behavioral incidents were added to Howard and Dailey's list; or (3) changes from global to behaviorally specific scales might have been responsible for attenuating the relationship between self-report measures and behavioral incidents. However, the authors find none of these potential causes compelling.

The question of whether providing information to subjects via behaviorally specific scales would remove response-shift effects must be answered in the negative. There was a significant response shift noted in the present study. While the magnitude of response shift in this study was

Table 6
Mean Pre-Then Differences and Results of
Univariate F-Tests for Studies 1 and 2
of Howard and Dailey (1979) and the Present Study

	Behaviorally Specific	Summative		Results of Univariate F-Tests (1,60 df)		
				Study 1		
Dimensions	Present Study	Study 1 H&D	Study 2 H&D	H+D with Study 2 H+D	H+D with Present Study	Study 2 H+D with Present Study
Questioning Techniques	1.10	1.13	.53	1.21	.06	.70
Structured Format	.15	1.45	1.57	.04	4.71*	5.47*
Supportiveness	.50	.95	.86	.05	.44	.20
Rapport	.55	.50	.39	.03	.05	.01
Actively Listened	1.45	.95	1.19	.23	1.17	.36
Relevant Material	.50	1.38	.71	1.58	3.74	.47
Overall	.55	1.28	.71	1.04	1.31	.02

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

slightly smaller than was observed in the two studies by Howard and Dailey, the differences did not approach statistical significance.

More generally, there are many reasons why it is difficult to measure change using gain scores (Cronbach & Furby, 1970; Linn & Slinde, 1977). The primary problem is that there is greater error associated with difference scores than with single measurements. Therefore, researchers generally prefer posttest comparisons between experimental and control groups rather than comparisons of change scores. Unfortunately, response-shift bias effects render posttest-only comparisons invalid, since treatment and control subjects' ratings are made with respect to different scales (see Howard et al., 1979, pp. 20–21). In this case retrospective measures are suggested to gain control for the effects of response-shift bias. The approach of the authors is preferred, not because it allays the problems identified by Cronbach and Furby, but because their solution to the problem is no longer appropriate. Consequently, it is necessary once again to measure change as the lesser of two evils.

However, if treatment and control subjects provide ratings with respect to different scales, does this not pose a threat to construct validity (Cook & Campbell, 1976)? Regrettably, no clear-cut resolution to this question can be given at this time. However, progress has been made by Golembiewski, Billingsley, and Yeager (1976), who identified three conceptually different types of change: (1) alpha (true change; changes in level or state over time taken on a constantly calibrated instrument); (2) beta (observed variation, where apparent change is due to recalibration of the instrument between assessments); and (3) gamma (reconceptualization by the participant of the phenomenon that is measured). Beta change would seem to be related to response-shift effects and is handled by the use of retrospective measures. With the control of beta effects, the assessment of alpha change becomes straightforward. However, the quantification of gamma effects, which would constitute the threat to construct validity, is difficult. Golembiewski et al. (1976) recommended the use of factor analysis with a comparison of factor structures at pretest and posttest representing an estimate of gamma change. This approach is unsatisfactory for several reasons. Terborg, Howard, and Maxwell (under review) suggested the use of profile analysis as a means of assessing gamma change independently of alpha and beta change (the independence of assessment issue was one of the major problems of the factor analytic approach). However, the newer approach for gamma assessment requires substantial additional computations. Terborg et al. (under review) summarized by pointing out that "the measurement of change is a complex and problematic endeavor. . . . Once again, we find that human beings are complex and cognitive beings. Our suggestions are intended to enable us to appreciate further human change in its complexities" (p. 24).

CONCLUSIONS

Data from the two studies reported herein suggest that providing subjects with information at pretest about the target construct (assertiveness or interviewing skills) will not substantially alter pretest ratings nor attenuate response-shift effects. For the present, it must be concluded that response-shift effects appear to be truly treatment dependent; hence, a retrospective approach may be required to remove their deleterious effects. The major untested explanation for response-shift effects is that subjects change their awareness of their level of functioning with regard to the construct.

References

Banaka, W. H., *Training in depth interviewing.* New York: Harper & Row, 1971.

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.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field setting. M. B. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand-McNally, 1976.

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

Eisler, R. M., Miller, P. M., & Hersen, M. Components of assertive behavior. *Journal of Clinical Psychology.* 1973, *29,* 295-299.

Fear, R. A. The evaluation interview. New York: McGraw-Hill, 1973.

Flanagan, J. C. The critical incident technique. *Psychological Bulletin,* 1954, *51,* 327-358.

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

Galassi, J., Hollandsworth, J. G., 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.

Golembiewski, R. T., Billingsley, K., & Yeager, S. Measuring change and persistence in human affairs; Types of change generated by OD designs. *Journal of Applied Behavioral Science,* 1976, *12,* 133-157.

Gormally, J., Hill, C., Otis, M., & Rainey, L. A microtraining approach to assertion training. *Journal of Counseling Psychology,* 1974, *22,* 299-303.

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., Millham, J., Slaton, S., & O'Donnell, L. Influence of subject response-style effects on retrospective measures. *Journal of Research in Personality* (under review).

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

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.

Rathus, S. Instigation of assertive behavior through videotape-mediated assertive models and directed practice. *Behavior Research and Therapy,* 1973, *11,* 57-65.

Schwab, D. P., Heneman, H. G. III, & DeCotis, T. A. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology,* 1975, *28,* 549-562.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology,* 1963, *47,* 149-155.

Terborg, J. T., Howard, G. S., & Maxwell, S. E. Evaluating planned organizational change: A proposed method for the assessment of alpha, beta, and gamma change at the individual and group level. *Academy Management Review,* (under review).

Acknowledgments

The authors thank Robert Pritchard and Scott Maxwell for their comments on earlier drafts of this manuscript.

Author's Address

Send requests for reprints or further information to George S. Howard, Department of Psychology, University of Houston, Houston, TX 77004.