Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-evaluation of Retrospective Pretests

George S. Howard
University of Houston

Kenneth M. Ralph
Lancaster County Office of Mental Health and Mental Retardation

Nancy A. Gulanick
University of Houston

Scott E. Maxwell
University of Houston

Don W. Nance
Wichita State University

Sterling K. Gerber
Eastern Washington State College

True experimental designs (Designs 4, 5, and 6 of Campbell & Stanley, 1963) are thought to provide internally valid results. This paper describes five studies involving the evaluation of various treatment interventions and identifies a source of internal invalidity when self-report measures are used in a Pretest-Posttest manner. An alternative approach (Retrospective Pretest-Posttest design) to measuring change is suggested, and data comparing its accuracy with the traditional Pretest-Posttest design in measuring treatment effects is presented. Finally, the implications of these findings for evaluation research using self-report instruments and the strengths and limitations of retrospective measures are discussed.

Campbell and Stanley (1963) assert that true experimental designs (Designs 4, 5, and 6) control for all sources of internal invalidity. One threat to internal validity is Instrumentation, which is defined as changes in the calibration of a measuring instrument or changes in raters' standards. Campbell and Stanley (1963) recommend the use of multiple "blind" raters and randomized rating materials to equalize Instrumentation effects for both the treatment and control groups. When self-report instruments are used, however, it is the research subjects themselves who serve as raters. Since treatment subjects have had different experiences than control subjects (i.e., the experimental treatment), the possibility of a confounding of Instrumentation with the experimental treatment exists. This constitutes a potential source of internal invalidity, even in true experimental designs. This potential is exacerbated when a purpose of the treatment is to change the subjects' understanding or awareness of the variable being measured, as seems to be the case with most treatment and training interventions.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 1 Winter 1979 pp. 1–23

Many researchers using a wide variety of self-report instruments have failed to find convincing support for the continuation of various psychological treatments (Bergin, 1971). In many such studies, these results were in stark contrast to both clients' and therapists' perceptions that treatment had been beneficial (Wolberg, 1960). This discrepancy between research findings and both client and therapist perceptions has led to the belief that present research methods and/or instruments are inadequate to assess psychological benefits.

In using self-report instruments, researchers assume that subjects have an internalized perception of their level of functioning with regard to a given dimension and that this internalized standard will not change from one testing to the next (pretest to posttest). As Cronbach and Furby (1970) note, researchers must be able to state the equivalent value on the posttest set of scores of each particular score on the pretest set of scores. That is, a common metric must exist between the two sets of scores. If the standard of measurement changes between the pretest and the posttest score, the two ratings will reflect this difference in addition to actual changes. Consequently, comparisons of the ratings will be invalid (Campbell & Stanley, 1963; Caporaso, 1973; Neale & Leibert, 1973).

Do treatments alter subjects' perceptions in a manner which contaminates self-report assessment of the treatment? If so, can these changes be measured and their deleterious effects removed from the assessments?

This paper presents the results of five studies in which self-report instruments are used to evaluate treatment interventions. Study 1 demonstrates a somewhat paradoxical finding, an apparent *increase* in dogmatism in subjects following a communications skills training workshop designed to reduce dogmatism. Discussions with later workshop participants suggested that changes during the workshop in their perception of their initial level of dogmatism were responsible for these confusing results. Study 2 examines this outcome more closely by employing an alternative approach to measuring change, the Retrospective Pretest-Posttest design. Conclusions about the effectiveness of the workshops were radically different for the two approaches (Pretest-Posttest vs. Retrospective Pretest-Posttest). Which approach, then, is more valid? Studies 3, 4, and 5 address this issue by comparing self-reported indices of change collected using both designs with more objective measures of change. The findings favor the Retrospective Pretest-Posttest design in providing a measure of self-reported change which is in closer agreement with the objective changes observed. These five studies suggest that when self-report measures are used, pretest-posttest comparisons might be contaminated by an instrumentation with experimental treatment confounding which constitutes a source of internal invalidity.

STUDY I

Method

As part of a program offered through Eastern Washington University, 48 communication skills workshops were conducted at Air Force bases across the country. All workshops were 30 hours in length and followed a predetermined course of topics and structured exercises.

Subjects

The subjects in this study were 704 male noncommissioned officers.

Facilitators

Twenty clinical psychologists served as workshop facilitators.

Instruments

The Rokeach Dogmatism Scale (RDS; Roakeach, 1960) is a 40-item self-report instrument designed to measure dogmatism. Possible scores range from 0 to 240, with the

higher scores representing a greater degree of dogmatism. The workshop coordinators selected the RDS because they felt it would be sensitive to the attitudinal changes reported by former workshop participants.

Procedure

The facilitators administered the RDS to all subjects at the beginning of the first workshop session (Pre). The workshop sessions were designed to increase the subjects' awareness of the factors which influence interpersonal communications. Subjects took part in structured group exercises wherein they gave and received feedback on the quality of their interactions. Interview and role-playing practice sessions were conducted which afforded the subjects an opportunity to practice more effective communication techniques. The facilitators readministered the RDS to all subjects at the end of the last workshop session (Post). All subjects completed a workshop evaluation form in which they were requested to comment on the workshop experience. The evaluation and unscored RDS forms were sent to Eastern Washington State University where the results were analyzed.

Results

Post minus Pre difference scores on the RDS were computed for all subjects. Figure 1 presents a histogram of the Pre/Post difference scores.

Sixty-two percent (434 subjects) of the program participants reported becoming more dogmatic during the course of the workshops, while 36% (253 subjects) became less dogmatic, and 2% (17 subjects) showed no change. A t-test (one tailed) for related measures revealed a marginally significant effect ($t = 1.65$ (703) $p < .05$) such that the mean posttest score was *more* dogmatic than the mean pretest score ($\bar{X}_{pre} = 156.72$, $S.D. = 70.12$; $\bar{X}_{post} = 162.25$, $S.D. = 77.37$). This finding was very surprising, since the differences were predicted in the opposite direction.

Discussion

One possible explanation of the results of this study is that a large number of workshop participants actually were more dogmatic upon completion of the workshop than they had been initially. However, all of the facilitators expressed extreme skepticism about the validity of this conclusion. Furthermore, in almost all instances, comments made on the workshop evaluation form reported changes in the subjects which are typically associated with becoming less dogmatic.

An alternative explanation involves the adequacy of the RDS as a measurement instrument. It may be that the RDS is a poor instrument and/or was not measuring the particular kinds of changes which were taking place in these subjects. However, the results were *not* "inconclusive," which is typically the sign of a bad measuring device or an inappropriate instrument. Rather, negative results were found, suggesting that the data might be providing evidence regarding a real phenomenon at work in these groups.

In an attempt to shed light on the causes of these disconcerting findings, a number of group members at a later workshop were interviewed regarding their responses on RDS. Many individuals reported that during the course of the workshop, they changed their perception of their initial level of functioning. A typical example of how subjects reported change in their pattern of thought while responding to an RDS item follows:

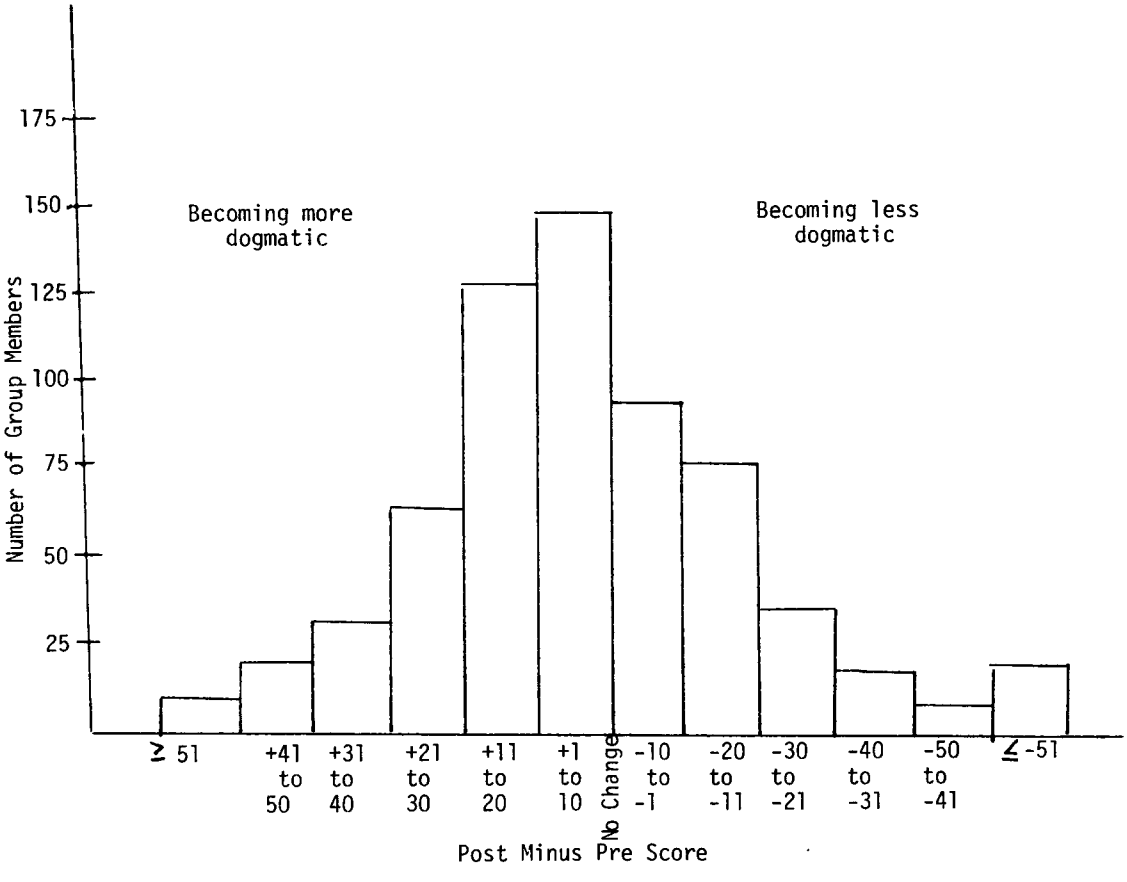
Item 13: In a heated discussion I generally become so absorbed in what I am going to say that I forget to listen to what others are saying.

Subject at Pretest: "I listen to what other people say when I'm talking to them. I'd say − 2 (I disagree on the whole)."

Subject at Posttest: "All those group exercises made me realize that I don't listen to people. I should have put +3 (I agree very much) the

Figure 1

Pre to Post Differences for the 1973-1974 Workshop Participants on the Rokeach Dogmatism Scale (RDS)



first time I filled this out. But the group really opened my eyes and helped me to try to be more of an active listener and so while I still sometimes forget to listen to people, overall I'm not doing nearly so badly now. I'll put −1 (I disagree a little)."

The data from these ratings were Pre "−2" and Post "−1." These ratings indicated that the individual had become more dogmatic. Clearly, however, the individual's perception was that the group helped him to become "less dogmatic" (i.e., Pre "+3," Post "−1"). The change in how he perceived his initial level of functioning on that dimension had confounded his report of improved functioning. This change in a subject's basis for determining his/her level of functioning on a given dimension is referred to as a "response-shift." The results of these post hoc interviews stimulated a second study to further examine the existence and impact of the response-shift.

STUDY II

Method

If the negative results of Study I were due to response-shifts, it was reasoned that by obtaining measures of subjects' pretest and posttest levels of dogmatism which were not contaminated by response-shifts, different conclusions might be reached regarding the effectiveness of the workshops. A method was sought whereby self-reported measures of pretest and posttest levels of dogmatism would be made with respect to the same internal standard. It was hypothesized that by substituting for the usual pretest a retrospective pretest (Campbell & Stanley, 1963; Deutsch & Collins, 1951; Information and Education Division, 1947; Woodruff & Birren, 1972) administered at the time of posttesting, the effects of treatment-produced response-shifts would be eliminated. Twelve workshops were conducted in which the RDS was administered using either a standard Pretest-Posttest or a Retrospective Pretest-Posttest procedure.

Subjects

The subjects in this study were 247 male noncommissioned officers.

Facilitators

Two clinical psychologists with extensive previous experience conducting the communications workshops served as group facilitators.

Procedure

The participants in each of the 12 communication workshops were randomly divided into two groups. Only the first group completed the RDS at the beginning of the first workshop session (Pre). The same 30-hour communication workshop as described in Study I was conducted. At the end of the last workshop session, this first group completed the posttest, while the other group was given the RDS and instructed to respond to each item twice. First, they were to report how they perceived themselves to be at present (Post). Immediately after answering each item in this manner, they were to answer the same item again, this time in reference to how they now perceived themselves to have been just before the workshop was conducted (Then). Subjects were instructed to make the Then response in relation to the corresponding Post response to insure that both responses would be made from the same perspective.

Results

Post minus Pre or Post minus Then difference scores were computed for all subjects. Table 1 presents the number of subjects in the Pre/Post and Then/Post conditions who reported becoming more or less dogmatic following the workshop.

A significantly greater number of group members using the Then/Post procedure reported becoming less dogmatic than group members using the Pre/Post procedure. (λ^2 (1) =11.17, $p <$

Table 1
Number of Subjects in each Condition Who Report Becoming More or Less Dogmatic

Condition	Number who reported becoming more dogmatic	Number who reported becoming less dogmatic
Pre/Post	53	51
Then/Post	37	89

.001). This result is evident for subjects regardless of which facilitator conducted their workshop (λ^2 (1) = 4.56, $p < .05$; λ^2 12.07, $p < .001$). Thirteen individuals in the Then/Post condition and four individuals in the Pre/Post condition reported no change and were thus excluded from this analysis.

Since subjects were randomly assigned to either the Pre/Post or Then/Post conditions, it was assumed that these groups are equivalent on their initial levels of dogmatism. If, however, the type of response-shift suggested by the interviews conducted with workshop participants was occurring, it would be expected that the individuals in the Then/Post condition would report more dogmatic Then scores than were reported as Pre scores by subjects in the Pre/Post group. Similarly, since subjects in both groups took part in the same workshops, no differences would be anticipated between the Post scores for the two groups. As expected, comparison of the Then and Pre scores revealed the Then scores to be reliably more dogmatic than the Pre scores (t (246) = 2.12, $p < .05$, $\bar{X}_{pre}$ = 145.18, $\bar{X}_{then}$ = 152.81); but no differences were found between the two sets of Post scores (t (246) = .27, *n.s.*, $\bar{X}_{post}$ (Pre/Post group) = 146.2, $\bar{X}_{post}$ (Then/Post group) = 144.9).

Discussion

The Then/Post procedure provided radically different results with which to evaluate the workshop compared to the Pre/Post procedure. Furthermore, this discrepancy can be attributed to differences between the retrospective pretest and the pretest. Similar results have been reported by Woodruff and Birren (1972).

Response-shift theory provides a plausible explanation for these findings. An increase in the subjects' understanding of the phenomenon under consideration or an increased appreciation of their initial level of functioning on that dimension could have caused them to report Then scores which were more dogmatic than their pretest scores might have been. However, other explanations are also possible. For example, these same results might have occurred if (1) subjects' memory of their pretest levels were inaccurate or (2) subjects biased their reports to provide the experimenters with favorable results.

These latter explanations reflect the concerns Campbell and Stanley (1963) express about retrospective measures. Campbell and Stanley note that such measures are susceptible to memory distortions. They conclude that pretests are more accurate than retrospective pretests and that retrospective pretest data should be collected only when pretest results are unavailable. However, Campbell (personal communication) stated that this conclusion was drawn without consideration of the possibility of an Instrumentation with experimental treatment confounding for self-reports and should be reconsidered if further research reveals that retrospective self-report pretests provide results which are more accurate than the usual self-report pretests.

The results of Studies I and II suggest that for self-report measures, a Then/Post procedure might yield more accurate change scores than a conventional Pre/Post procedure. The Then/Post procedure provided a set of results which agreed with the participants' perceptions of change, reported on their workshop evaluations, more closely than did the results obtained using the usual Pre/Post procedure. Since there is no generally accepted objective/behavioral measure of dogmatism, efforts to investigate the nonsubjective accuracy of Then/Post change scores compared to Pre/Post change scores required shifting to areas where both self-report and objective outcome measures were available. The following three studies reflect this shift.

STUDY III

Method

During spring semester 1976, two types of group programs for women were conducted. The groups were designed to promote androgyny in women by fostering the development of positive skills typically stereotyped as "masculine" in our society. In the first a discussion orientation (DO) similar to that utilized in consciousness-raising groups was employed. The second was a full treatment (FT) which utilized the assertiveness training technique of behavioral rehearsal along with discussion (Gulanick, 1976). The same topics were covered in both groups. In order to monitor the effectiveness of these groups, self-report and objective measures of assertiveness, sex-role orientation, and attainment of individual goals were obtained. Pretest, posttest, and retrospective pretest ratings were collected on all self-report measures, thus allowing generation of Pre/Post and Then/Post change scores which could then be compared to objective measures of change for all subjects. It was hypothesized that the retrospective pretest would be superior to the traditional pretest in providing a measure of self-reported change which agreed more closely with an objective index of change on each dimension.

Subjects

Serving as subjects for the study were 51 women who scored "feminine" on the Bem Sex-Role Inventory (Bem, 1974) and who were interested in participating in an experimental group aimed at fostering androgyny. The women were recruited through announcements made in a number of classes at a large southwestern university.

Facilitators

Four advanced graduate students in applied mental health programs who had prior experience as group facilitators served as facilitators for this study. They received 12 hours of didactic/experiential training in the specific techniques to be employed in the two groups. The facilitators worked in pairs, with each set of cofacilitators conducting one DO and one FT group.

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 repondents 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 (1976).

Objective Measure of Assertiveness (OMA). The objective measure of assertiveness consisted of each subject's verbal responses to eight taped stimulus situations. The stimulus situations were the same as or similar to those used by Eisler, Hersen, and Miller (1973). Each subject was instructed to listen to each stimulus statement and to respond verbally using the actual words she would use if the situation were really happening to her. All responses to the stimulus statements were audio-taped, coded,

and later rated for assertiveness by two trained raters using the Rathus Assertiveness Scale (Rathus, 1973). The interrater reliability (Pearson *r*) for this instrument in the present study was .92

Counseling Outcome Inventory (COI). The COI (Hill, 1975) is a self-report measure which provides an individualized measure of change on goals designated by a subject as personally relevant and important. In using the COI, the experimenter developed with each subject a list of six traits on which she would like to change and the specification of a behavioral definition for each (i.e., "assertion" may be defined as initiating conversations with co-workers before work). The experimenter insured that the traits listed by all subjects related to topics to be covered by the group 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 functioning on each, using a scale from "−3" (very dissatisfied) to "+3" (very satisfied). The product of the rank ordering provided a weighted score for each item, and the sum of the weighted scores yielded a total score.

Facilitator Ratings (FR). At the conclusion of the treatment groups, each facilitator was asked to rate "how much each member profited from the group experience" on a scale from "1" (not at all) to "5" (very much). The score for each subject was the sum of the ratings of the cofacilitators.

The Bem Sex-Role Inventory (BSRI). The BSRI (Bem, 1974) is a self-report measure of sex-role orientation which treats masculinity and femininity as separate dimensions, rather than as opposite ends of the same dimension. Respondents rate themselves on each of the 60 items using a 7-point scale. The Androgyny Score provides a measure of a person's relative masculinity and femininity, defined as Student's t-ratio for the difference between an individual's endorsement of the masculine versus feminine items. Data on the reliability and validity of the scale are presented by Bem (1974) and Gaudreau (1975). The recommendation made by Gaudreau (1975) to drop the items "masculine" and "feminine" from the inventory was incorporated into the instrument used in the present study. The BSRI was used as a selection measure in this study, and only those women with an Androgyny Score of +2.025 (feminine) or greater were eligible for participation.

Procedure

Thirty-five of the subjects were randomly assigned to either full treatment (FT) or discussion treatment (DO) condition such that there were two groups of 8 or 9 members in each condition. Seventeen subjects served as a waitlist control (WL). All subjects were pretested individually during the nine days prior to the first group session. The FT and DO groups met once a week for two hours for six sessions. During the week following the last group session, all subjects were given a posttest appointment at which the self-report Post and Then data were gathered and the OMA was readministered. Lastly, the facilitators were asked to rate how much each woman had profited from the group experience. Follow-up assessments were conducted two months after the posttest (mailed self-report instruments only) and one year after posttest (all self-report scales completed in an "as you are Now"/Then "as you were before the group" fashion). The OMA was administered at the one-year follow-up but not at the two-month follow-up.

Results

Change scores from pretest to posttest, pretest to two-month follow-up, and pretest to one-year follow-up were calculated twice: the first time using the self-report Pre scores and the second time using Then scores on all self-report measures. One-way analyses of variance were performed on these change scores. Mean ratings and the results of the analyses are reported in Table 2. Significant differences among groups

Table 2
Means and Summary Statistics for Dependent Measures
on Pre, Then, Post, Two-Month Follow-up and One-Year Follow-up

Measure	Group	Pre	Then	Post	2-Month Follow-up	1-Year Follow-up	Anova Pre or Then to Post Differences	Anova Pre or Then to 2-month Follow-up	Anova Pre or Then to 1-Year
BSFI t	FT	3.49	3.98	1.31	0.63	-0.99	P/Post 2.36	P/2 mo. 4.63*	P/1 yr. 5.19*
	DO	3.80	4.54	2.29	1.61	2.61			
	WL	3.54	3.87	3.77	3.17	2.59	T/p 8.22**	T/2 mo. 6.91	T/1 yr. 3.58*
BSRI	FT	3.91	3.58	4.60	4.72	5.14	P/p 2.60	P/2 mo. 1.06	P/1 yr. 2.45
MASC	DO	3.79	3.38	4.10	4.39	4.42			
	WL	3.95	3.76	4.06	3.89	4.17	T/p 9.10**	T/2 mo. 4.39*	T/1 yr. 2.80
CSES	FT	100.35	96.81	130.18	131.15	138.70	P/p 5.23**	P/2 mo. 3.09	P/1yr. 3.09
	DO	94.35	84.82	112.94	121.69	117.86			
	WL	101.65	103.47	110.82	116.80	121.38	T/p 11.58**	T/2 mo. 4.81*	T/1 yr. 1.98
COI	FT	-25.82	-32.12	28.76	34.00	35.00	P/p 9.35 **	P/2 mo. 2.62	P/1yr. .46
	DO	-31.06	-39.06	21.12	34.84	26.29			
	WL	-20.65	-26.89	-0.82	26.30	17.86	T/p 9.32**	T/2 mo. 1.38	T/1 yr. 2.73
OMA	FT	3.00		4.15		4.27	P/p 6.66**		P/1 yr. 5.21**
	DO	2.89		3.64		3.84			
	WL	3.24		3.56		3.71			

were found on the objective measure (OMA) from pretest to posttest and from pretest to one-year follow-up. When considering pretest to posttest treatment change on the self-report measures using the Pre scores, reliable differences among groups were found on the CSES and COI; but similar trends failed to reach significance on the BSRI androgyny and masculinity scales. However, when Then scores were used, all four scales showed significant differences. Similar comparisons substituting two-month follow-up scores for posttest scores yielded a similar pattern of results. When Pre scores were used, pretest to two-month follow-up differences among groups were found only on the BSRI androgyny scale. However, when Then scores were used, differences were noted for the BSRI androgyny, BSRI masculinity, and CSES measures. Conversely, analyses of one-year follow-up data using Pre and Then scores yielded findings which were in essential agreement with one another. To summarize, of the 12 analyses performed on self-report instruments using Pre scores, significant differences among groups were noted in 4 instances; whereas when Then scores were employed, 8 of the comparisons yielded statistically significant differences among treatment groups.

Inspection of mean ratings on self-report measures for the three groups revealed that differences between Pre and Then scores were larger for subjects in the two treatment groups than for their control group counterparts. This finding supports the intuitive hypothesis that "response-shift" effects are treatment dependent and are therefore potential contaminants in designs employing placebo or wait-list control groups.

As noted earlier, subjects completed a retrospective pretest at the one-year follow-up. Table 3 presents mean pretest and retrospective pretest ratings of treatment subjects who participated in the one-year follow-up. Inspection of Table 3 reveals that on three of the measures (BSRI *t*-Score, BSRI Masculinity Score, and COI), the retrospective pretest ratings of one-year follow-up were closer to the Then ratings made at posttest than they were to the Pre ratings. The ratings at posttest were not significantly different from retrospective ratings at one-year follow-up. However, in one instance (COI), Then scores at one-year follow-up were different from Pre scores (t (15) = 2.74, $p < .05$).

Discussion

While the objective measure of change (OMA) employed in this study was an obvious improvement over the anecdotal (facilitators' impressions, workshop evaluation form) evidence of

Table 3
Mean Pre, Then (At Posttest), and Then (At one-year Follow-up) Ratings on Self-Report Measures for Treatment Group Subjects. (N = 17)

Measure	Pre	Then (At Posttest)	Then (At one-year follow-up)
BSRI t	3.56	4.41	4.37
BSRI Masc.	3.87	3.63	3.71
COI	-26	-33	-35
CSES	100	92	96

change in Studies I and II, it did utilize a role-play format and hence should not be viewed as a "true" behavioral measure. Nevertheless, regarding the effectiveness of the intervention, the Then/Post analysis was generally more in agreement with the analysis of OMA results than the Pre/Post self-report analysis.

A number of writers have drawn attention to problems in measuring change (Cronbach & Furby, 1970; Linn & Slinde, 1977). These fundamental problems are exacerbated when change scores are correlated to obtain information regarding the relative validity of Then/Post versus Pre/Post self-report approaches. Hence, the following comparisons should be viewed with caution. First, for treatment subjects, objective behavioral change was more closely related to Then/Post difference scores than to Pre/Post difference scores. This is evidenced by correlations between self-report and judges' ratings of assertiveness (CSES with OMA: $r_{T/P} = .54$, $r_{P/P} = .41$) and attainment of individual goals (COI with FR: $r_{T/P} = .25$, $r_{P/P} = .15$). However, the differences between these pairs of correlations were not statistically significant. Second, control subjects Then/Post with judges' ratings of change correlations were no higher than their Pre/Post with judges' ratings of change correlations. This is consistent with the response-shift hypothesis, since control subjects were not exposed to an intervention which would alter their basis for responding.

To obtain further anecdotal data about response-shift phenomena, subjects were asked immediately after completing the Then measure to give their reactions to completing a retrospective pretest. The differences between the responses of treatment and control subjects were striking. Control subjects reported that their pretest responses still seemed valid and hence saw no reason for altering those ratings. Treatment subjects, on the other hand, were extremely articulate in documenting the differences between their Pre and Then ratings and in pinpointing the specific events within the group which caused them to doubt the validity of the Pre ratings. Finally, on all of the self-report measures employed in this study, most individuals reported lower levels of functioning on the Then scores than on the Pre scores. This same pattern was apparent in the response shifts for dogmatism reported in Studies I and II.

STUDY IV

Method

Response to the FT program of Study III was so strong and positive, it was decided to offer a modification of the program in fall 1976. A research component was included in an effort (1) to replicate the findings of Study III and (2) to introduce other self-report measures in order to ascertain if they were also subject to response-shift effects.

Recruitment, selection, pretesting, and posttesting procedures were the same as those employed in Study III. No discussion-only groups were included, and the treatment period was extended to eight weeks in order to include two additional topics not covered in Study III.

Subjects

Eighteen FT subjects were divided into two FT groups, and 13 subjects served as no-treatment controls. Twelve FT subjects completed the program and supplied complete data to be used in subsequent analyses.

Instruments

The Adult Self-Expression Scale (ASES). The ASES developed by Gay, Hollandsworth, and Galassi (1975) was used in the place of the CSES to measure assertiveness. The ASES is a 48-item self-report measure of assertiveness designed for adults in which respondents describe themselves using a 5-point scale. Scores can range from 0 to 192, with higher scores reflecting a more assertive response pattern. Gay, Hollandsworth, and Galassi (1975) report reliability and validity data on the ASES.

The Personal Attributes Questionnaire (PAQ). The PAQ, developed by Spence, Helmreich, and Stapp (1974), was used along with the BSRI as a measure of sex-role orientation. The PAQ consists of 55 bipolar adjectives which fall into three subscales: male-valued, female-valued, and sex-specific. Subjects rate themselves for each item on a 5-point scale as to how descriptive each trait is of them. Since the aim of the treatment program was to increase the participants' masculine behavior potential, the male-valued subscale was used as a measure of masculinity in the replication study.

Treatment Program

The treatment program was expanded from 6 to 8 weeks. The additional topics, "Dealing with conflict/criticism from others" and "Support systems" were included in order to aid the generalization of changes made during the treatment.

Results and Discussion

Pretest to posttest change scores were computed twice: the first time using Pre scores and the second time using Then scores for all self-report dependent measures. One-way analyses of variance were performed on the change scores. Mean ratings and results of the analyses are reported in Table 4.

A significant difference between groups on pretest to posttest change on the objective measure of assertiveness (OMA) was found. Regarding the effectiveness of the intervention, similar conclusions would be drawn whether the self-reported Pre or Then scores were used on the BSRI Masculine Score, PAQ Masculine Score, and COI. However, when Pre scores were employed for the BSRI Androgyny and ASES measures, the analyses failed to show significant differences between groups; whereas when Then scores were employed, statistically significant differences between groups were observed.

A problem encountered when using pretest to posttest change scores is that they are negatively correlated with pretest scores (Cronbach & Furby, 1970). This problem is highlighted when there are sizable discrepancies in pretest scores. While this was not the case in Study III, there was great variability among pretest ratings (ranging from 2.3 to 4.5) for treatment subjects on the OMA in Study IV. Given that a maximum posttest score is 5.0, one subject had a potential for change which was 5.4 times that of another subject. To lessen the impact of this imbalance, Pre/Post OMA scores were converted to "G" statistics by the formula (Post−Pre)/(Scale Max−Pre) (McGuigan, 1967). Then/Post ASES change scores correlated more highly with OMA change ("G" statistic) than did Pre/Post ASES (ASES with OMA: $r_{T/P}$ = .28, $r_{P/P}$ = .12).

As with Study III, these results should be viewed with caution. Yet the Then/Post analyses do appear to provide a somewhat clearer picture of the effectiveness of the treatment than the Pre/Post self-report approach. Further, anecdotal evidence does suggest somewhat greater concurrent validity for a Then/Post approach relative to a Pre/Post approach. Thus, while these findings in themselves are not conclusive, they do raise questions that warrant further investigation.

STUDY V

Method

Independent of the development of Studies III and IV, Study V was also conducted to determine the validity of Then/Post measures relative to Pre/Post self-report indices of change. In addition to exploring response-shift biases in a totally new training/education setting, Study V attempted to ascertain if Pre/Then differences (attributed to response-shifts in this paper) might be due to systematic memory distortion. Specifically, the discussion of response-shifts assumed that at posttesting, subjects remembered their

Table 4
Mean Pre, Post, and Then Ratings
and Summary Statistics for Dependent Measures

Measure	Condition	Pre	Then	Post	Analysis of Difference Scores	
BSRI t	Treatment	2.42	3.22	.25	P/P	2.49
	Control	3.41	3.83	2.70	T/P	5.44*
BSRI	Treatment	4.05	4.18	5.01	P/P	4.79*
MASC	Control	3.79	3.48	3.92	T/P	4.76*
PAQ	Treatment	44.08	46.83	56.50	P/P	2.79
MASC	Control	41.46	40.31	45.62	T/P	1.83
ASES	Treatment	100.50	93.17	131.83	P/P	3.35
	Control	81.30	75.92	94.17	T/P	5.11*
COI	Treatment	-35.84	-29.92	29.42	P/P	7.26*
	Control	-30.00	-28.54	3.54	T/P	8.93**
OMA	Treatment	3.27		4.53	P/P	9.82**
	Control	3.13		3.58		

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

pretest level of functioning, remembered their Pre rating of their level of functioning, and consciously *chose* to provide a different and more accurate set of ratings (Then). Study V attempted to determine if subjects' memory of their Pre ratings is accurate.

Subjects

Serving as subjects for this study were 51 undergraduate students enrolled in a credited course entitled "Communication in Helping Interviews" at a midwestern university. Two sections of the course were offered, meeting at different times with different instructors. Within each section, the students were randomly divided into three groups of 17 subjects each, named respectively "Pre/Post" (PP), "Then/Post" (T/P), and "All Test" (AT) groups. The instructor did not know to which group each subject was assigned.

Thirty-seven subjects completed all testing required for the group to which they were assigned and thus were included in this study. Fourteen subjects were excluded because of incomplete sets of data (four from the PP group, eight from the TP group, and three from the AT group).

Instruments

The Helping Questionnaire (HQ). The HQ is a 12-item self-report instrument designed for this study to assess students' perceptions of their own helping skills levels. Items tap such skills as attending, reflecting feelings, reflecting content, and goal setting included in the course content. Examples of typical items include the following:

Item 7: In general, when you try to help someone with a concern or problem, how well do you understand which feelings the helpee is experiencing?

Item 8: In general, when you try to help someone with a concern or problem, how well do you tell the helpee how you understand his/her feelings?

Responses range from "no understanding" (1) to "complete understanding" (9). Items 1 and 12 are identical and assess students' perceptions of their overall helping ability (i.e., "In general, when you try to help someone with a concern or problem, how helpful are you?")

Procedure

The course utilized a lecture/lab method to teach students the art of helping (Carkhuff, 1969), including the skills of attending, reflecting feelings, reflecting content, summarizing, and goal setting. In their first class period, subjects in the PP and AT groups were asked to complete the HQ. After completing the instrument, all the subjects were randomly paired. Each pair was then assigned to a small recording room. They were instructed to conduct two 15-minute helper-helpee interviews. One student (the helpee) was to talk about a real problem or concern that he or she was experiencing, and the partner (the helper) was to be helpful. After about 15 minutes, the pairs were to give feedback to each other and then switch roles for the other 15-minute helper-helpee interview. All interviews were tape recorded.

Just prior to the final exam, the PP subjects were asked to complete another HQ, while the TP subjects were asked to complete the HQ in a Then/Post manner. The AT subjects also completed the HQ in this same Then/Post manner. However, after completing the Then/Post HQ, the AT subjects were also requested to complete a memory HQ. The memory HQ instructed the AT subjects to record what they remembered their precourse HQ ratings to be.

After these posttest self-report ratings were completed, all subjects were again randomly paired to conduct two additional 15-minute helper-helpee interviews. The same precourse instructions and procedures were used for these postcourse helper-helpee interviews. Samples taken from the 72 precourse and postcourse interviews which were sufficiently audible to be rated were coded and randomized. These

samples consisted of either the first 10 minutes or the entire interview if it was shorter than 10 minutes, as was the case for some precourse interviews.

Ratings

The raters were two psychology graduate students who had previously been trained to rate on both Carkhuff (1969) and Truax and Carkhuff (1967) rating scales and had over 60 hours of prior rating experience. They were "blind" concerning the design and hypotheses of the study and whether they were rating Pre or Post tapes. The raters were instructed to rate the helper in each sampled interview on three scales: Feeling, Content, and Global. The Feeling and Content scales were derived from Carkhuff (1969) and Truax and Carkhuff (1967) helper scales but were altered to become parallel forms of Item 8 (communicating empathic feelings) and Item 9 (communicating understanding of content) on the HQ. Raters were also instructed to record on the Global scale their impressions of "how helpful" each helper would be in general. The Global scale thus parallels HQ Items 1, 12, and the total HQ score. Interrater reliabilities for the Feeling, Content, and Global scales were, respectively, .95, .94, and .94. In view of these high correlations, the mean ratings of the two raters was considered a reliable measure of a subject's helper-helpee interview helping skill level.

Results and Discussion

T-tests between Pre and Post scores for subjects in the PP group, AT groups, and judges' ratings were performed. Similar analyses between Then and Post scores were calculated for AT subjects and NT subjects. Data relevant to the response-shift phenomenon in AT group subjects are presented in Table 5. As hypothesized, mean Then ratings were significantly lower than either mean Pre or mean Memory ratings. The Then with Memory and Then with Pre large *t*-values suggest that subjects' postcourse estimates of their precourse levels of functioning (Then) were systematically lower than their actual precourse estimate and also their memory of those precourse ratings. Biases or inadequate memory of the actual precourse ratings do not explain these results, since mean differences between subjects' memory ratings and precourse ratings were small. Thus, these results support the contention that a response-shift occurred between pretest and posttest administration and that these differences in pretest and retrospective pretest ratings are due to something other than systematic distortion of subjects' memory of their precourse level of functioning.

HQ Items 1 and 12 ask the subjects to rate their overall helping ability. The intervening items (Items 2–11) might serve to make the subjects more aware of the expertise possible in helping, causing a mini-response-shift. For the 27 subjects who were administered HQ pretests, the mean Item 12 ratings ($\overline{X} = 6.6$, $S.D. = 1.2$) were smaller than mean Item 1 ratings ($\overline{X} = 6.9$, $S.D., = .8$). While these differences were in the hypothesized direction, they did not reach statistical significance ($t\,(26) = 1.44$).

Perhaps the most important comparisons are whether Then/Post self-report measures of change were more in agreement with judges' ratings of change in their interviewing behavior measures than were self-report measures used in a Pre/Post manner. Combining data for all subjects, the Feeling ($t\,(35) = 4.81$, $p < .01$), Content ($t\,(35) = 4.10$, $p < .01$), and Global ($t\,(35) = 4.21$, $p < .01$) posttreatment judges' ratings were significantly higher than the pretreatment ratings, suggesting that students did improve their levels of helping skills. Table 6 presents means and summary statistics for the three groups on HQ scales and judges' ratings. Conclusions concerning the effectiveness of the course would clearly be different if the TP rather than PP group is considered, since Then and Post scores for HQ Items 8 and 9 (concerning the heavily emphasized abilities of the subjects to communicate empathic feeling and content to helpees) and

Table 5
Means and Summary Statistics for Pre, Then, and Memory Ratings on Target HQ Items of AT Group Subjects (N =14)

HQ Items	$\bar{X}$ Pre	$\bar{X}$ Memory	$\bar{X}$ Then	Pre with Memory t	Memory with Then t	Then with Pre t
8	6.3	6.4	5.5	-.17	2.75**	1.59
9	6.4	6.5	5.9	-.19	2.28**	1.39
1	6.9	6.5	5.8	1.31	2.69**	3.04**
12	6.7	6.8	6.1	-.29	2.86**	2.28*
Total HQ	77.6	77.2	68.3	.15	3.23**	2.63*

*p < .05
**p< .01

total HQ scores were significantly different for the TP group. Comparable Pre and Post scores, however, were not reliably different for the PP group. The AT group provided for comparisons of both Then/Post and Pre/Post scores within the same sample. While both analyses revealed significant differences, the Then scores were consistently lower than the comparable Pre scores, as predicted. Finally, as can be seen in Table 6, Then/Post change scores were consistently higher than Pre/Post change scores, which is consistent with the previous studies reported in this paper. Given substantial Pre to Post changes in judges' ratings, the Then/Post scores reflected these changes more clearly than did self-reported Pre/Post comparisons.

GENERAL DISCUSSION

Taken together, these five studies lend strong support to the contention that when self-report measures are used in a Pre/Post manner, the results might well be confounded by a response-shift. Evidence for this phenomenon was found using measures of dogmatism (Studies I and II), assertiveness (Studies III and IV), and helping skills (Study V). Because of the broad range of settings and instruments in which the response-shift has been observed, it seems likely that a sizeable portion of the literature on program evaluation, counseling outcome, group, attitude, and personality research might be influenced by this confounding. Since the validity of a large quantity of data is at stake—data upon which important scientific and practical conclusions are drawn—the issues presented and supported in these studies warrant further extensive and objectively critical investigation. In every instance, the bias operated to increase the probability that the experimental hypothesis would be rejected. Therefore, it is likely that existent literature which relies upon Pre/Post self-report measures may also contain errors of conservativism. This concurs with the contention of many practitioners who assert that research studies often erroneously fail to document the benefits of their interventions.

One might be concerned that in all the studies reported herein where Then/Post data was gathered, the treatment was found to be effective. Perhaps whenever subjects are given the expectation that change has occurred and asked to respond to a self-report instrument in a Then/Post manner, they will automatically report changes. Two other studies were conducted (Ralph, 1975) using the Tennessee Self-Concept Scale in a Then/Post manner to evaluate the im-

Table 6
Means, Standard Deviations, and Change Statistics for
Judges' ratings and Self-Reported Change for All Groups in Helping Skills

Pre/Post Group (N = 13)

	Pre		Post				Pre		Post		
HQ Items	X	s.d.	X	s.d.	t	Judges' Ratings	X	s.d.	X	s.d.	t
18	6.5	.8	6.6	1.1	.56	Feeling	3.54	1.7	5.15	1.68	2.83*
9	6.8	.9	7.0	1.2	.81	Content	4.38	1.4	5.50	1.1	2.03*
1	6.8	.7	6.9	1.0	.56	Global	4.65	1.4	5.50	1.3	2.45*
12	6.5	1.0	7.3	1.1	2.51						
Total HQ	79.4	9.9	84.2	10.7	1.99						

Then/Post Group (N = 9)

	Then		Post				Pre		Post		
HQ Items	X	s.d.	X	s.d.	t	Judges' Ratings	X	s.d.	X	s.d.	t
8	3.9	1.1	6.9	.3	8.05**	Feeling	3.38	1.2	5.27	.9	4.98**
9	4.6	.7	7.1	.6	10.55**	Content	4.44	1.2	5.50	.8	2.40*
1	5.3	1.1	7.1	.8	5.49**	Global	5.06	1.0	5.61	.9	1.36
12	5.3	1.1	7.1	.9	4.88**						
Total HQ	60.2	8.1	85.8	7.2	10.31**						

All Tests Group (N = 14)

	Then		Post				Pre		Post		
HQ Items	X	s.d	X	s.d.	t	Judges' Ratings	X	s.d.	X	s.d.	t
8	5.5	1.2	7.8	.6	8.00**	Feeling	3.57	1.3	4.93	1.2	2.21*
9	5.9	.7	7.7	.5	10.48**	Content	3.93	1.2	5.07	1.0	2.87*
1	5.8	1.0	7.4	.5	6.27**	Global	4.21	1.2	5.71	.9	4.53**
12	6.1	.8	7.5	.5	6.03**						
Total HQ	68.3	9.6	89.6	3.8	9.00**						

	Pre		Then		
HQ Items	X	s.d.	X	s.d.	t
8	6.3	1.1	7.8	.6	5.97**
9	6.4	.8	7.7	.6	4.18**
1	6.9	.8	7.4	.5	1.58
12	6.7	.8	7.5	.5	2.69*
Total HQ	77.6	7.4	89.6	3.8	4.76**

* p .05; **p .01

pact of semester-long growth groups. In one instance subjects reported slight, nonsignificant positive change, while absolutely no change was reported in the other study. The same evaluations conducted in a Pre/Post manner yielded the same findings, suggesting that Then/Post change reflects more than subject compliance in providing favorable evaluations. (Descriptions of these studies can be obtained upon request).

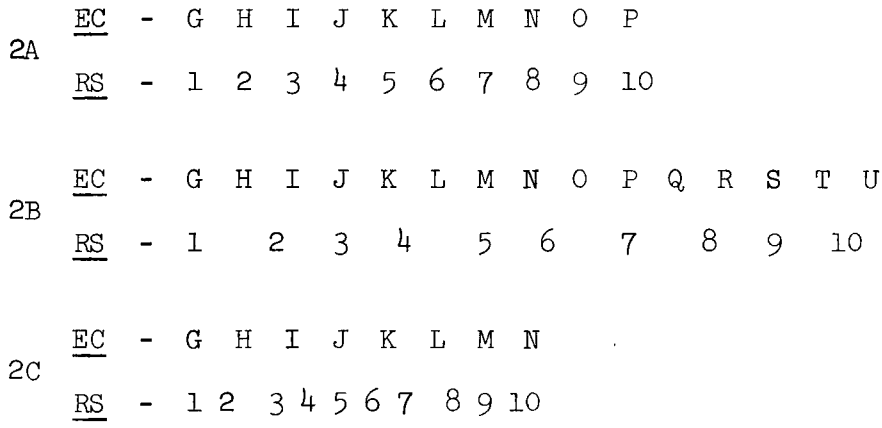
Theory of Response Shifts

Subjects' selections of self-report responses are intended to identify the nature of their experiences. Subjects form an understanding of what the points on the response scales represent in terms of possible types and degrees of particular experiences or perceptions they might have. The process subjects use to match their awareness to the responses provided is like the overlaying of a stable continuum (e.g., marks on a piece of paper) on a varying continuum (e.g., marks on a partly stretched piece of elastic material). The stable continua represent the actual types and degrees of awareness that the subjects might experience. The varying continua represent the subjects' understanding of the available self-report responses. The response scale continua are stretched to fit the subjects' experience continua.

Figure 2 illustrates how these two continua might be related. The letters of the experience continuum represent possible intensities or degrees of a subject's experience. If a subject's particular initial internal state was *K*, then "5" would be the corresponding response scale rat-

Figure 2

Examples of How a Subject's Experience Continuum (EC) Might Correspond to His/Her Understanding of a Self-Report Response Scale (RS)



2A - The initial correspondence

2B - After a positive expanded shift

2C - After a negative contracting shift

ing. However, due to a subject's experiences over time, his/her awareness of possibilities might increase from, for instance, *P* to *U*. In turn, the subject's perception of the response scale would stretch to fit the experience continuum. *G* and "1" would still correspond, but the intervals between alternative responses would increase. Thus, as can be noted in Figure 2B, although the subject's perception of the variable may have actually increased from *K* to, for example, *M*, the response corresponding to *M* on the second assessment might still be "5."

In any case, such a positive expanding shift of the response scale to fit an enlarged experience continuum would produce systematic errors of measurement. In the communication workshops, for example, a positive expanding shift may have occurred, since group members were made aware of their typically low level of functioning on the variables measured by the Rokeach Dogmatism Scale. Such a response-shift would account for the perplexing findings of increased dogmatism ratings after the treatment, while facilitators and participants verbally reported decreased dogmatism. This type of response-shift and the resultant systematic error tend to work against experimenters who are trying to find positive changes on self-report instruments. Furthermore, this type of shift might be expected whenever subjects' perceptions of positive possibilities increase.

On the other hand, it is possible that an individual's subjective perception of positive possibilities could decrease. This is more likely for subjects in no-treatment or placebo-treatment conditions. Due to the absence of true change, subjects might become more pessimistic about the possibilities of positive change. The effect, rather than an expansion, might be a contraction of the experience continuum.

In the example used above, experiences *O* and *P* would be excluded. *G* and "1" would still correspond, but the interval between alternative scale values on the response scale would decrease. Thus, as noted in Figure 2C, a subject who on pretest indicates a "5" (corresponding to experience *K*) may on posttest indicate a "6," although no true gain on the variable of interest occurred.

The contraction of positive possibilities thus yields systematic errors which may partly account for slight gains in control groups where no change was expected. Consequently, gains in control groups (a common occurrence) make it more difficult for researchers to find significantly greater gains in experimental groups.

Although only two types of shifts have been discussed up to this point (i.e., the positive expanding and the positive contracting shifts), other types of shifts are quite conceivable. For example, if the subjects believed that their initial perceptions of the low possibilities were incorrect, negative expanding shifts (in which the high end of the scale is fixed but the intervals between scale values is increased) or negative contracting shifts (in which again the high end of the scale is fixed but the intervals between scale values is decreased) are both possible. Indeed, shifts in the relative position of the experience continua and response scales can occur in as many ways as two scales can be related.

Method of Analysis

There are several alternative methods of analyzing data from a study utilizing pretest, posttest, and retrospective pretest information. Consider a study in which an experimental group is to be compared with a control group when group assignment is random. Five possible comparisons between the two groups are:

1. Compare mean posttest scores.
2. Compare mean posttest-pretest difference scores.
3. Compare mean posttest-retrospective pretest difference scores.
4. Compare posttest means adjusted by pretest means through analysis of covariance (ANCOVA).
5. Compare posttest means adjusted by retrospective pretest means through analysis of covariance.

Huck and McLean (1975) show that for most repeated measures designs, the analysis of covariance is preferred. They do not discuss, however, either the possibility of a response-shift with self-report measures or the inclusion of a retrospective pretest. Thus, it is not clear whether their demonstration of the superiority of ANCOVA will hold when there is response-shift.

Following Huck and McLean's formulation of structural models for pretest-posttest designs, the following models are appropriate when there is a response-shift:

Group *Pretest*

Control

$$Y_{11m}=\mu+\pi_{m(1)}+B_1+B\pi_{1m(1)}+\varepsilon_{11m} \quad (1)$$

Treatment

$$Y_{21m}=\mu+\pi_{m(2)}+B_1+B\pi_{1m(2)}+\varepsilon_{21m} \quad (2)$$

Group *Retrospective*

Control

$$Y_{12m}=\mu+\pi_{m(1)}+B_2+B\pi_{2m(1)}+\varepsilon_{12m} \quad (3)$$

Treatment

$$Y_{22m}=\mu+\pi_{m(2)}+B_2+\gamma+B\pi_{2m(2)}+\varepsilon_{22m} \quad (4)$$

Group *Posttest*

Control

$$Y_{13m}=\mu+\alpha_1+\pi_{m(1)}+B_3+B\pi_{3m(1)}+\varepsilon_{13m} \quad (5)$$

Treatment

$$Y_{23m}=\mu+\alpha_2+\pi_{m(2)}+B_3+\gamma+B\pi_{3m(2)}+\varepsilon_{23m} \quad (6)$$

where:

Y_{ijm} is the score on Y for the m^{th} *person in group* i on measure j,

μ is the grand mean parameter,

$\pi_{m(i)}$ is the effect associated with the m^{th} person in group i,

B_j is the effect for the j^{th} measure,

$B\pi_{jm(i)}$ is the effect associated with the interaction of measure j and person m in group i,

α_i is the effect of treatment i,

γ is the response-shift parameter, and

ε_{ijm} is a random error component for the m^{th} person in group i on measure j (with the usual ANOVA assumptions about error).

These models differ from Huck and McLean's only in that allowance is made here for the possibility of a response-shift for treatment subjects on the retrospective pretest and the posttest because after receiving the treatment, these subjects may shift the scale of response by γ units, either up or down. Although γ can be either positive or negative in the model, it seems that treatment subjects typically have lower scores than controls on the retrospective pretest (when the effect of the treatment itself is to increase scores, as in the present studies), so that γ probably tends to be negative.

The five possible analysis procedures can now be compared, since structural models have been specified. A desirable procedure must provide an unbiased estimate of $\alpha_2 - \alpha_1$, the treatment effect; otherwise, the procedure is biased, and estimates as well as tests of significance will be misleading.

The expected value of the comparison of treatment with control when group assignment is random for each of the five procedures is as follows:

1. Comparison of mean posttest scores

$$E(\bar{Y}_{23.}-\bar{Y}_{13.})=(\alpha_2-\alpha_1)+\gamma \quad (7)$$

2. Comparison of mean posttest-pretest difference scores

$$E[(\bar{Y}_{23.}-\bar{Y}_{21.})-(\bar{Y}_{13.}-\bar{Y}_{11.})]=(\alpha_2-\alpha_1)+\gamma \quad (8)$$

3. Comparison of mean posttest-retrospective pretest difference scores

$$E[(\bar{Y}_{23.}-\bar{Y}_{22.})-(\bar{Y}_{13.}-\bar{Y}_{12.})]=\alpha_2-\alpha_1 \quad (9)$$

4. Comparison of posttest means adjusted by pretest means (the regression coefficient is denoted by b)

$$E[(\bar{Y}_{23.} - b\bar{Y}_{21.}) - (\bar{Y}_{13.} - b\bar{Y}_{11.})] = (\alpha_2 - \alpha_1) + \gamma \quad (10)$$

5. Comparison of posttest means adjusted by retrospective pretest means (the regression coefficient here is also denoted by *b*)

$$E[(\bar{Y}_{23.} - b\bar{Y}_{22.}) - (\bar{Y}_{13.} - b\bar{Y}_{12.})] = (\alpha_2 - \alpha_1) + (1 - E(b))\gamma \quad (11)$$

The only procedure that leads to an unbiased estimate of the treatment effects is the third, the comparison of mean posttest-retrospective pretest difference scores. The other four approaches will, on the average, underestimate the true treatment effect if $\alpha_2 - \alpha_1$ and γ are opposite in sign, as current evidence suggests they may often be (for the fifth treatment, $E(b) < 1$ must also hold). Only the third approach, the analysis of mean posttest-retrospective pretest differences, provides a test of the null hypothesis of real interest, namely that $\alpha_1 = \alpha_2$. Surprisingly, then, the analysis of covariance is not the method of choice in this situation, contrary to its status in the traditional conception of pretest-posttest designs. For this reason, treatment effects in the studies described in this paper have been tested by comparing mean posttest-retrospective pretest difference scores rather than by using analysis of covariance.

Toward the Future

While the studies reported herein provide a substantial beginning to an understanding of response-shifts, this is clearly but the beginning. Further documentation and clarification of the prevalence and impact of response-shift phenomena is needed. How should research on this problem proceed?

Although the present studies favored the Then/Post approach in providing a more accurate estimate of a treatment effect, future research should assess the conditions under which a Pre/Post design would be more appropriate. Howard and Tinsley (in prep.) suggest that researchers begin incorporating collection of Then data into their present Pre/Post designs. A preliminary analysis could then be made to determine whether a significant response-shift has occurred between pretest and posttest and which measures (Pre or Then scores) are appropriate for the subsequent data analysis. Several empirical methods for determining whether a significant shift has occurred have been offered (Howard & Tinsley, in prep.) but have yet to be investigated and compared in order to assess the limiting conditions (i.e., scaling assumptions) under which they might be employed.

When increasing subjects' understanding of their level of functioning on a specific dimension is one goal of an intervention, making a comparison of Pre and Then scores on that dimension might provide researchers with a means of assessing whether or not that goal has been met. Ironically, the same response-shift which, if ignored, serves to bias outcome research, might have the potential, when measured, to provide desirable outcome information.

Further research is needed to identify and clarify the various causal determinants of the response-shift. One factor which may be involved is the level of information subjects have at the pretest regarding the dimension on which they are asked to self-report. As illustrated in the dogmatism example described in the introduction, the treatment intervention may actually be providing subjects with information which will enable them to better assess their pretest level of functioning. Use of an "informed pretest," wherein a thorough description of the variable being measured is provided to a subject prior to administration of the self-report pretest instruments, may thus serve to lessen response-shift bias.

A second hypothesis to be studied suggests that the amount of time and effort which a subject spends in a treatment might conflict with his/her belief that no real change had occurred. This dissonance could be allayed by changing her/his initial self-assessment to one that would yield results more fitting to the effort spent. A

similar phenomena which may be operating to produce response shifts is subject acquiescence. This involves the perceived demand characteristics of the experimental situation and the subjects' desire to please the experimenter. Further research is needed to verify or refute the role of these factors in response-shifts.

Finally, it is obvious that the adequacy of the measures used in evaluation research affects the quality of the findings. The integration of self-report, objective, and behavioral measures has long been recognized as the most complete way to evaluate a treatment intervention. Use of pretest, posttest, and retrospective pretest self-report data, by providing a more sensitive assessment of a subject's perspective of personal change, will add yet another valuable dimension to evaluation research endeavors.

References

Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 155–162.

Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: John Wiley & Sons, 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.

Carkhuff, R. R. *Helping and human relations: A primer for lay and professional helpers* (Vols. 1 & 2). New York: Holt, Rinehart, & Winston, 1969.

Caporaso, J. A. Quasi-experimental approaches to social science: Perspective and problems. In J. A. Caporaso & L. L. Roos, Jr. (Eds.), *Quasi-experimental approaches: Testing theory and evaluating policy*. Evanston, IL: Northwestern University Press, 1973.

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

Deutch, M., & Collins, M. E. *Interracial housing: A psychological evaluation of a social experiment*. Minneapolis: University of Minnesota Press, 1951.

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

Galassi, J., Delo, J., Galassi, M., & Bastein, S. The college self-expression scale: A measure of assertiveness, *Behavior Therapy*, 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.

Gaudreau, P. *Bem sex-role inventory validation study*. Paper presented at the American Psychological Association Convention, Chicago, 1975.

Gay, M. L., Hollandsworth, J. G., & Galassi, J. P. An assertiveness inventory for adults. *Journal of Counseling Psychology*, 1975, *22*, 340–344.

Gulanick, N. A. *A group program for highly feminine women aimed at increasing androgyny*. Unpublished doctoral dissertation, Southern Illinois University, 1976.

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

Howard, G. S., & Tinsley, H. E. A. *Use of a retrospective pretest to accomodate the effects of response-shift bias with self-report measures*. University of Houston, in preparation.

Huck, S., & McLean, R. Using a repeated measure ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, *82*, 511–518.

Information and Education Division, U. S. War Department. Negro infantry platoons in white companies of seven divisions. In T. M. Newcomb & E. L. Hartley (Eds.), *Readings in social psychology*. New York: Holt, 1947.

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.

McGuigan, F. J. The G-statistic: An index of amount learned. *National Society for Programmed Instruction Journal*, 1967, *6*, 14–16.

Neale, J. M., & Liebert, R. M. *Science and behavior: An introduction to methods of research*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Ralph, K. M. *The self-report response-shift theory of now/then self-report procedure*. Unpublished doctoral dissertation, Southern Illinois University, 1975.

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

Rokeach, M. *The open and closed mind*. New York, 1960.

Spence, J. T., Helmreich, R., & Stapp, J. The personal attributes questionnaire: A measure of sex-role stereotypes and masculinity-femininity, *JSAS Catalogue of Selected Documents in Psychology,* 1974, *4,* 43.

Truax, C. B., & Carkhuff, R. R. *Toward effective counseling and psychotherapy.* Chicago: Aldine Publishing Co., 1967.

Wolberg, L. R. The evaluation of psychotherapy. *Acta Psychotherapeutica,* 1960, *12,* 262–283.

Woodruff, D. S., & Birren, J. E. Age changes and cohort difference in personality. *Developmental Psychology,* 1972, *2,* 252–259.

Acknowledgments

We thank our many colleagues for their helpful comments on earlier drafts of this paper. The comments of Jerome L. Myers were particularly substantial and most helpful in forming the final product.

Author's Address

Requests for reprints should be sent to George S. Howard, Department of Psychology, University of Houston, Houston, TX 77004.