Implications of the Pre/Post/Then Design for Evaluating Social Group Work

Elizabeth A. R. Robinson
Case Western Reserve University
Howard J. Doueck
State University of New York at Buffalo

Pre/post designs that rely on self-report measures are potentially contaminated by response shift bias, a change in respondents' understanding of the phenomena being measured between the pre- and posttests. Retrospective pretests, or then tests, have been shown to be useful in identifying response shifts. Pre/post/then designs are highly utilitarian designs that have the potential of providing more accurate assessments of respondent change. This article reviews the literature on response shift bias and the utility of retrospective pretests in detecting such bias. Three studies that used the pre/post/then design to evaluate self-reported changes in caseworker abilities (knowledge or skill) relative to practice evaluation, basic interviewing skills, and/or assessment are described.

The most typical design used in analyzing the effectiveness of training and intervention groups is the pre/post design, which obtains information from participants before and after the training or intervention and assumes that changes in the two measures reflect the impact of the training or intervention. The more sophisticated, and fortunate, researcher may add a control group that also provides pre/post data on the issue of concern (Campbell & Stanley, 1966; Cook & Campbell, 1979). An assumption of both of these designs when self-report instruments are used is that the measurement instrument is static—that a common metric exists for the two sets of scores. However, it is clear from the literature and the research presented here that this is an assumption that must be examined, particularly when the purpose of the treatment whose effectiveness is being examined is to change group members' understanding or awareness of the variable being measured.

Authors' Note: Correspondence may be addressed to Elizabeth A.R. Robinson, Mandel School of Applied Social Sciences, Case Western Reserve University, Cleveland, Ohio 44106. The authors acknowledge the helpful feedback and comments of participants at the Eighth Annual Symposium for Empirical Group Work, May 22-24, 1992, Madison, Wisconsin, at which this article was first presented.

Research on Social Work Practice, Vol. 4 No. 2, April 1994 224-239

The work of Howard and others (Bray, Maxwell, & Howard, 1984; Doueck & Bondanza, 1990; Howard, Milham, Slaten, & O'Donnell, 1981; Howard, Ralph, et al., 1979; Sprangers, 1988; Sprangers & Hoogstraten, 1989) has documented changes in the internal standards of respondents from pretest to posttest that render those two sets of scores no longer comparable, and therefore less valid indicators of change. For example, women who go through an assertiveness training program have a different understanding of assertiveness than they did before the training (Howard, Ralph, et al., 1979). Similarly, Air Force officers given communication skills training have a different understanding of their previous use of dogmatic responses than they had prior to that training (Howard, Ralph, et al., 1979). In both of these cases, the validity of a pre/post comparison is threatened by changes in instrumentation. The respondents have changed the calibration of the measurement instrument as they have changed their understanding of the variable being measured. This phenomenon has been labeled a response shift that introduces unwanted bias (termed response shift bias) into the research.

Response shift can be documented by using a then test, or a retrospective pretest, in which subjects are asked after treatment to rate their perceptions of themselves on the variables of concern prior to treatment. Significant pre/then differences suggest that a response shift has taken place (Howard, 1980; Sprangers, 1988).

The pre/post/then design asks subjects to complete a typical pretest and posttest, either a standardized instrument or a measure designed for the study. Following the treatment condition, and either before or after completion of the posttest, participants are asked to reflect back to the start of treatment and rate themselves on a second posttest, the then test. These then test scores allow the respondents to provide an assessment of their pretreatment condition based on a greater knowledge and awareness of the variables the treatment addressed. The retrospective pretest scores are compared to posttest scores to assess treatment efficacy and to pretest scores to determine the degree of response shift.

Areas in which a response shift bias has been documented include efforts to reduce social prejudice and dogmatism, assertiveness training, drug and alcohol use, personnel screening, interviewing skills, interviewing anxious patients, problem-solving skills training, training graduate assistants for teaching, and training caseworkers in practice evaluation (Doueck & Bondanza, 1990; Howard & Dailey, 1979; Howard et al., 1981; Howard, Ralph, et al., 1979; Howard, Schmeck, & Bray, 1979; Sprangers, 1988).

Researchers have tended to avoid retrospective data on the grounds that they are more likely to be distorted by memory lapses, cognitive errors, and

subject bias (Bloom & Fischer, 1982; Campbell & Stanley, 1966). However, post/then difference scores have been found to correlate more highly with non-self-report measures of change than pre/post difference scores (Howard & Dailey, 1979; Howard et al., 1981; Howard, Ralph, et al., 1979; Howard, Schmeck, et al., 1979). Howard, Ralph, et al. (1979) tested assertiveness skills and the attainment of individual counseling goals using pretest, posttest, and then test self-report measures. They also taped and coded verbal responses to stimulus situations and had facilitators, blind to subject conditions, rate assertiveness and goal attainment. They found then/post change scores on standardized measures of assertiveness and individual counseling goals were more in agreement with changes in the more objective pre/post ratings than they were with the pre/post self-report change scores. A later study by Howard and his colleagues (1981) reported that changes in facilitator rating scores correlated .52 with then/post change scores, compared to a correlation of .25 with pre/post change scores. An even more dramatic case is that reported by Howard and Dailey (1979), in which personnel screening interviewing skills (behavioral counts and ratings) correlated .33 and .43 with then/post scores, compared to –.05 and –.06 for pre/post scores. These studies suggest that retrospective pretests may not only be more valid than traditional pretests, but may also more accurately reflect subjects' status on the variables of concern.

Bray et al. (1984) analyzed the statistical power lost in pre/post designs when response shift bias is present and found it to be substantial. They found that the most powerful and unbiased estimate of treatment effect when response shift bias is present is the difference between posttest and retrospective pretest scores. Such response shift bias might well explain the lack of a relationship often found between self-report measures and other more objective indices of the same construct (Nisbett & Wilson, 1977). It probably also explains the lack of significance often found in studies of the effectiveness of intervention and training tested with self-report measures in pre/post designs, including group intervention and training that are perceived as more effective by their designers, group leaders, and participants than the data indicated. Howard (1980) has suggested that many treatment effectiveness studies using pre/post designs may be underestimates of effect, causing the experimental hypothesis to be rejected unnecessarily.

Are there other options for resolving this problem besides the retrospective pretest? The use of an informed pretest design in which respondents were provided information on the topic of concern has been investigated in several studies (Benjamin, 1982; Howard, Dailey, & Gulanick, 1979). These investigators found that providing information on the construct of interest prior to

the pretest did not improve the accuracy of self-reported assertiveness and interviewing skills.

Research has been carried out to disentangle the cause of such response shifts from pretest to then test. The issue here is whether such response shifts were due to changes in respondents' standards as suggested above, to response style bias (such as social desirability), or to subject bias, that is, to subjects attempting to provide positive results by adjusting their then test scores downward. Sprangers (1988) recently reviewed the existing studies that allowed for such an analysis by using placebo control groups. Because response-shifts are treatment dependent (that is they only occur in the treatment condition), a placebo control design should illuminate whether subject bias operates. A significant pre/then difference was not found in the placebo control conditions, but was found in the experimental conditions. It appears that subject bias is not a significant factor in response shift. Further, the results on response style suggest that social desirability is no more or less a factor in then tests than it is in pre- or posttests and may be less of an issue after treatment, at posttest and then test, than it is prior to treatment (Aiken, 1986; Howard et al., 1981; Sprangers & Hoogstraten, 1987, 1988).

Response shift bias is a potential source of internal invalidity in studies that rely solely on self-report pre/post tests to determine the effectiveness of an intervention or training. When response shift operates, the treatment is by definition effective, but uncontrolled and undetected changes in instrumentation are confounded with the experimental treatment. Response shifts are most likely to occur when the purpose of treatment is to change subjects' understanding, awareness, or knowledge of the variable being measured, which is the case in many interventions and training programs. The issue for a particular study using a pre/post design is to determine if subjects' perceptions have been altered by the treatment condition in a manner that contaminates self-report assessments collected prior to the treatment. Second, can these changes in subjects' perceptions of their status on the variables at issue be measured and a more accurate assessment of status prior to intervention be obtained?

Researchers in a number of fields, including social work, medicine, substance abuse treatment, education, management, and counseling psychology, have started to use the retrospective pretest to capture change. For example, Doueck and Bondanza (1990) used the pre/post/then design to evaluate a caseworker training group. In addition, it has been used to evaluate an interviewing skills training group (Howard & Dailey, 1979) and drug abuse treatment program (Aiken, 1986; Rhodes & Jason, 1987). This article reviews and reanalyzes the Doueck and Bondanza pre/post/then study re-

ferred to above (Study 1) and reports the findings from two additional studies using the same design with smaller samples and in different contexts and training content.

STUDY 1

Method and Procedures

Doueck and Bondanza (1990) trained 18 preventive services staff from a large urban multipurpose service agency. The training program focused on the use of single-systems designs to monitor and evaluate practice. Four weekly sessions were held, each session meeting for 3 hours, for a total of 12 hours of training. In accordance with the recommended procedure for a pre/post/then evaluation, Doueck and Bondanza evaluated their program using an instrument consisting of 11 items measured along a continuum from 0 (low level of current ability) to 10 (high level of current ability). A trained research assistant administered the instrument three times, once before the training and twice at the conclusion of the training. After the posttest was completed and returned, participants were given the following instructions:

> Now we would like to ask you to reflect back on your skills before the training began. Given your current understanding of the concepts and skills related to practice evaluation, we would like you to assess your level of ability at that time.

Participants were given the same instrument for the third and final time, the then test.

Results

For purposes of comparison with Studies 2 and 3, we reanalyzed the data using the Wilcoxon Signed Rank nonparametric procedure (Brown & Hollander, 1977). Like all nonparametric procedures, the Wilcoxon Signed Rank Test does not have the same assumptions about distributions that its parametric counterpart would have and is particularly appropriate when sample sizes are small. In addition, the output from the Wilcoxon test indicates the number of individuals who have either positive or negative differences, making it particularly informative when using the pre/post/then procedure. The Wilcoxon Signed Rank Test compares favorably to the *t* test and "may even be superior" to it when the assumptions for a *t* test are not met (Hays, 1973, p. 782). By comparison, the results from the Wilcoxon Signed Rank Test analyses were slightly more conservative than the *t* tests

reported in the original article. The interested reader is referred to Doueck and Bondanza (1990) for comparison.

In Table 1, a comparison of the pretest scores with the posttest scores indicated that 9 of the 11 items were statistically significant, with a 10th item, ability to measure client change during treatment (Item 7), approaching significance, $Z = -1.88$, $p = .06$.

When the then test scores were compared with the posttest scores, 10 items were found to be statistically significant, with the 11th item approaching significance, $Z = -1.91$, $p = .056$. None of the items were statistically significant during the pre/then comparisons, illustrating, perhaps, that a response shift was minimal.

Discussion of Study 1

The results from Study 1 provide only limited support for the potential usefulness of the pre/post/then approach to evaluate training. Results from the then/post analysis revealed that self-reported ability on two items (2 and 7) had been overestimated by workers during the pretest. Both items reached statistical significance in the then/post analyses. However, because of the large number of significant differences found in the pre/post analyses, the use of the then test did not substantially change the overall results. Nonetheless, if only a pre/post test design had been used, the additional changes found in the then/post analyses may have been overlooked.

STUDY 2

Method and Procedures

Sample. Seven preventive services and child protection staff from a rural department of social services participated in a training program, which focused on assessment and the use of practice evaluation methods. The sample was predominantly female (86%) and Caucasian (100%), with an average age of 32. Of the sample, 28% had a graduate degree in either social work or education. The remainder had bachelor's degrees.

Design and procedure. Participants received 12 hours of training administered by the project director, two 6-hour sessions meeting 1 day per week for 2 weeks. The instrument developed for use in this study consisted of 12 items measured along a similar continuum to Study 1, from 0 (low level of current

TABLE 1: Study 1: Wilcoxon Signed Rank Comparisons of Worker Ability Pre/Post/Then Training

	Pre/Post (n = 15)			*Then/Post (n = 15)*			*Pre/Then (n = 15)*		
	Cases			*Cases*			*Cases*		
Ability to	−	+	Z	−	+	Z	−	+	Z
1. Select an appropriate design	1	14	−3.23**	1	14	−3.29***	5	2	−.76
2. Specify measurable problems	4	6	−.97	2	9	−2.49*	7	3	−1.38
3. Specify measurable goals	1	8	−2.31*	1	10	−2.71**	5	2	−1.10
4. Specify a treatment plan	3	9	−2.12*	1	7	−2.24*	6	4	.00
5. Evaluate outcomes	3	8	−2.00*	4	10	−1.91	5	7	−.59
6. Use graphs and tables to monitor progress	1	12	−2.90**	2	11	−2.69**	5	5	−.82
7. Measure client change	5	9	−1.88	0	12	−3.06**	7	5	−1.10
8. Develop a measurement package	1	13	−3.11**	1	14	−3.24**	2	6	−.98
9. Use statistics to analyze progress	1	11	−2.70**	0	10	−2.80**	6	3	−.83
10. Develop a measurement tool	2	12	−2.95**	3	12	−2.73**	4	7	−1.11
11. Assess empirically based questionnaire	1	14	−3.15**	2	11	−2.87**	4	8	−1.45

*$p < .05$; **$p < .01$; ***$p < .001$.

ability) to 10 (high level of current ability). With the single exception that all portions of the evaluation were administered by the trainer, the precise procedure described above was followed during administration of the instruments.

Results

Table 2 presents the results. In the pre/post comparisons, 4 of the 12 items were found to be significant at the .05 level. However, the then/post comparison found 11 of the 12 items statistically significant. When the pretest scores were compared with the then test scores, 4 items were statistically significant and 2 items, specify client problems in measurable terms (Item 5) and measure client change (Item 8), approached significance, $Z = -1.83$, $p = .068$, and $Z = -1.89$, $p = .059$, respectively. The findings in the pre/then analyses indicated that a substantial response shift occurred on 8 of the 12 items. Further, the pre/then data indicated that the participants tended to overestimate their pretest scores, doing so on 11 of the 12 items. Only 1 item, evaluate the results of an intervention (Item 12), was underestimated by the participants.

Discussion of Study 2

Study 2 provides much stronger evidence of the potential usefulness of the pre/post/then approach to evaluate casework training groups. The differences between the results of the pre/post analyses and the then/post analyses are particularly striking. Seven items that were not significant in the pre/post analyses became significant in the then/post analyses. Further, the use of the Wilcoxon Signed Rank Test is a relatively easy method for identifying the number of individuals who had overestimated or underestimated their knowledge and skills on each item. The results from the pre/then analyses provided some support for the assumption that participants tend to be more likely to overestimate self-reported skills and knowledge rather than underestimate them.

STUDY 3

Method and Procedures

Sample. Ten child welfare staff (child protection, preventive services, and adoptions) from another rural county department of social services participated in a training program that focused on basic interviewing skills and

TABLE 2: Study 2: Wilcoxon Signed Rank Comparisons of Worker Ability Pre/Post/Then Training

	Pre/Post (n = 7)			*Then/Post (n = 7)*			*Pre/Then (n = 7)*		
	Cases			*Cases*			*Cases*		
Ability to	–	+	Z	–	+	Z	–	+	Z
1. Do an ecological assessment	0	6	–2.20*	0	7	–2.37*	5	1	–1.26
2. Do an individual assessment	1	3	–1.28	0	7	–2.37*	6	0	–2.20*
3. Develop a problem inventory	1	6	–1.44	0	7	–2.37*	6	0	–1.28
4. Prioritize client problems	0	3	–1.60	0	6	–2.20*	5	0	–2.02*
5. Specify client problems in measurable terms	2	3	–.67	0	5	–2.02*	4	0	–1.83
6. Specify client goals in measurable terms	2	1	.00	0	3	–1.60	4	1	–1.08
7. Specify an intervention	1	4	–1.62	0	5	–2.20*	3	2	–.67
8. Measure client change	2	4	–1.15	0	7	–2.37*	5	1	–1.89
9. Develop client-specific measures	0	7	–2.37*	0	7	–2.37*	6	0	–2.20*
10. Assess an empirically based questionnaire	1	4	–1.62	0	7	–2.37*	5	1	–1.68
11. Develop a measurement package	0	7	–2.37*	0	7	–2.37*	4	1	–1.48
12. Evaluate the results of an intervention	0	7	–2.37*	0	7	–2.37*	0	5	–2.02*

*$p < .05$.

techniques. The sample was 70% female and 100% Caucasian. Average age of participants was 35, with a range from 23-47. They came from a variety of academic backgrounds (human services, sociology, criminal justice, English literature, and labor relations). None of the participants had a graduate degree.

Design and procedure. Participants received 12 hours of training administered by the project director, two 6-hour sessions meeting 1 day per week for 2 weeks. Similar to the prior studies, the pre/post/then evaluation procedure was followed. The instrument developed for use in this study consisted of 20 items measured along a continuum from 0 (low level of current ability) to 10 (high level of current ability). Participants were asked to self-assess their ability to accomplish each of the tasks identified in the items. The procedure described in Study 2 was followed during administration of the instrument.

Results

Table 3 presents the results. In the pre/post comparisons, six items were found to be significant, five at the .05 level and one at the .01 level. In contrast, the then/post comparison found all remaining items with the exception of paraphrase content (Item 3), statistically significant. Item 3 approached significance, $Z = -1.95$, $p = .051$. In the pre/then comparisons, one item was statistically significant, use of genuineness as a response (Item 16), and one approached significance, use of open and closed questions (Item 1), $Z = -1.84$, $p = .066$. In only one instance, provide advice (Item 11), pretest scores were underestimated by participants compared to then test scores. All other items were overestimated at pretest.

Discussion

Study 3 provides even stronger support for the potential usefulness of the pre/post/then approach to evaluate caseworker training. As in Study 2, there were marked differences between the results of the pre/post analyses and the then/post analyses. Further, Study 3 provides additional support for the proposition that participants are more likely to overestimate self-reported skills and knowledge rather than underestimate them.

GENERAL DISCUSSION

The results from all three studies indicate the utility of the pre/post/then design for evaluating social work training groups, with the strongest evidence

TABLE 3: Study 3: Wilcoxon Signed Rank Comparisons of Worker Ability Pre/Post/Then Training

	Pre/Post (n = 10)			*Then/Post (n = 10)*			*Pre/Then (n = 10)*		
	Cases			*Cases*			*Cases*		
Ability to	−	+	Z	−	+	Z	−	+	Z
1. Use open and closed questions	2	6	−1.40	1	9	−2.50*	8	1	−1.84
2. Reflect feelings	2	7	−1.66	0	9	−2.67**	6	1	−1.69
3. Paraphrase content	3	6	−.65	1	8	−1.95	6	3	−1.60
4. Interpret nonverbal behavior	2	5	−1.27	0	9	−2.67**	7	2	−1.72
5. Respond to nonverbal behavior	1	7	−1.47	0	8	−2.52*	5	3	−1.26
6. Interpret client behavior or content	1	7	−1.40	0	10	−2.80**	6	2	−1.26
7. Use summarization	3	6	−1.18	0	9	−2.67**	4	2	−1.47
8. Use clarification	1	8	−2.37*	0	10	−2.80**	4	3	−.34
9. Use confrontation	2	8	−1.53	0	7	−2.37*	6	4	−.76
10. Provide information	1	7	−2.17*	0	9	−2.67**	4	4	−.21
11. Provide advice	1	9	−2.65**	0	9	−2.67**	2	5	−1.69
12. Use concreteness as a response	3	7	−.92	0	7	−2.37*	4	4	−1.12
13. Use structuring as a response	3	6	−1.36	0	8	−2.52*	5	5	−.41
14. Use empathic understanding	1	7	−2.17*	0	9	−2.67**	6	3	−1.60
15. Use unconditional positive regard	3	6	−1.42	0	9	−2.67**	4	4	−.70
16. Use genuineness as a response	2	4	−1.36	0	8	−2.52*	8	1	−2.37*
17. Establish trust	1	7	−2.10*	0	9	−2.67**	7	1	−1.75
18. Use power in a relationship	1	7	−2.24*	0	6	−2.20*	4	3	−.17
19. Use self-disclosure	2	7	−1.54	1	8	−2.37*	4	3	−.85
20. Use intimacy in a relationship	1	9	−1.78	1	7	−2.24*	3	3	−.31

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

coming from Studies 2 and 3. There were stronger findings in the then/post analyses for each study, compared to the pre/post comparisons. On the majority of items, workers tended to overestimate their abilities at pretest compared to their more informed retrospective ratings. As a result, the percentage of items that were statistically significant in the then/post analyses increased dramatically. Few items were underrated during the pretests. In only one instance, Item 5 in Study 1, did underrating lead to a reduction in significance level.

These findings generally indicate that an important use of the pre/post/then design would be to reduce the probability of a Type II error, rejection of a group training program that was potentially beneficial. It is possible that for some group interventions or training programs the opposite effect might occur. This outcome, however, appears to be much less likely as evidenced by our findings and those of others who have investigated response shift bias.

The use of nonparametric statistics with this design can make it particularly user friendly. In addition to the fact that such statistics are not dependent on rigorous assumptions relative to the nature of the data, the Wilcoxon Signed Rank Test is simple to use and the results indicate the number of individuals who tended to underestimate or overestimate each item (Brown & Hollander, 1977). Researchers and evaluators who have larger sample sizes and control groups may wish to use more sophisticated tests such as those advocated by Koele and Hoogstraten (1988a, 1988b) and Bray et al. (1984).

These studies are limited by the lack of non-self-report measures that could corroborate whether the pretest or then test measure was more reflective of workers' abilities prior to training. We therefore cannot confirm the assumption that then test scores are more accurate, although we feel fairly confident in that assumption, given previously published research and our own experience implementing these groups. The studies are also limited in their generalizability by their small sample sizes and the nature of the training provided.

The above studies and the cited literature suggest that using a pre/post/then design can add to our knowledge about an intervention and reduce the likelihood of rejecting an effective intervention. Most group interventions are designed to alter participants' understanding of a common issue—whether practice evaluation among protective workers, substance abuse among addicted clients, or battering among spouse abusers. A critical part of group work is increasing participants' willingness to share that understanding. If group workers and researchers rely solely on pre/post indicators of change, substantial changes will be missed because pretests are inherently

limited to participants' beliefs and knowledge prior to the group and their willingness to share those cognitions. The classic pretest is limited not just by deficiencies in knowledge and skills, but also by social desirability and social presentation issues, as well as the more fundamental issues of defensiveness, trust, and self-esteem. Particularly in the case of treatment groups, where trust, cohesiveness, and self-disclosure are fostered by group leaders, one expects both understanding of the common issue and willingness to share that understanding to increase over the course of the group. Both of these issues affect pregroup measures, making them less valid than retrospective measures. The idea of a retrospective pretest is not new or unique. Howard, Ralph, et al. (1979) refer to a 1951 study on prejudice that used a then test. Retrospective pretests hold a lot of promise for researchers, evaluators, trainers, or social group workers. They are particularly useful for those who run small groups, often with low budgets, and who want to obtain a complete sense of the extent of change their clients or participants have experienced. The pre/post/then design appears to be a relatively cost effective, easily implemented method for evaluating such programs.

Despite the benefits of the design, it cannot substitute for a more rigorous approach to evaluation with no-treatment control groups, placebo control groups, random assignment, and the inclusion of non-self-report pre/post measures when such opportunities exist. In addition, other threats to internal validity besides response shift bias are not adequately controlled for in the pre/post/then design and the issue of generalizability to in vivo situations is limited. Nevertheless the pre/post/then design is an improvement over typical pre/post or post-only designs for evaluating group treatments. The methodological issues raised by this body of research are intriguing. Response shift bias is a phenomenon whose effects on a pre/post design can easily confound the conclusions we might draw, particularly if the design relies heavily on self-report measures. The use of a retrospective pretest in a pre/post design can help us estimate the extent of the response shift should it exist. Because response shift may be dependent on treatment efficacy—that is, a pre/then difference when no such difference exists in a no-treatment or placebo control group—finding a response shift indicates that such treatments are effective whereas finding no pre/then difference is a strong indication that the treatment was ineffective. This conclusion would correspond with what Howard and colleagues reported. However, determining the precise cause of the pre/then shift becomes far more difficult without the use of controls, as such differences may or may not be the result of response style, social desirability, or other factors (see Howard & Dailey, 1979; Howard et al., 1981; Sprangers,

1988; Sprangers & Hoogstraten, 1989). The alternative hypothesis of social desirability has been offered as an explanation for marked pre/then differences in the alcohol and drug abuse literature (Aiken, 1986; Rhodes & Jason, 1987) and may be evident in other highly stigmatized treatment fields such as family violence.

Researchers may wish to vary the timing of the pretest and retrospective pretest. Anticipatory effects have been reported in which the decision to enter treatment has been sufficient to initiate client change even prior to the inception of treatment (Rubin, 1992). As a result, a pre/post finding would be attenuated if the pretest were administered at the beginning of the intervention. Alternatively, if building trust is a major issue with a particular group, obtaining a pretest after 3-4 weeks of intervention may provide a more accurate perspective on pretreatment levels than a typical pretest administered at the inception of treatment (S. Rose, personal communication, May 22-24, 1992).

Recognizing the limits of any single measure in accurately capturing a complex phenomenon such as change, social science researchers, including social work researchers, have advocated the use of multiple measures. It was the incongruence between the results obtained from traditional self-report measures and other, more objective measures, such as behavioral indices and observer ratings, as well as anecdotal participant and trainer information, that prompted some of the original work on response shift bias. Including a then test in a pre/post design provides another measure that helps triangulate toward a better estimate of the actual pretreatment status of clients or group participants. In short, we recommend the use of retrospective pretests to increase the validity of self-report measures in social group work practice, evaluation, or research.

Research questions remain on the circumstances in which a response shift that threatens the validity of a study is most likely to occur. For example, is response shift bias more likely to occur when participants are not aware or knowledgeable about the treatment in question or when participants are reluctant to self-disclose at pretests? Is response shift more likely when the treatment attempts to change knowledge, skills, attitudes, or unrecognized feelings (e.g., prejudice) but not recognized feelings or beliefs? Would group processes such as cohesiveness, self-disclosure, or decision making, have positive or negative effects on the validity of the measures, compared to traditional pretest measures? Research in social group work would be advanced by a greater understanding of response shift bias and the circumstances under which it may or may not operate.

REFERENCES

Aiken, L. S. (1986). Retrospective self-reports by clients differ from original reports: Implications for the evaluation of drug treatment programs. *The International Journal of the Addictions, 21*(7), 767-788.

Benjamin, E. R. (1982). Using the post-then method of evaluation. *Training, 19*, 72.

Bloom, M., & Fischer, J. (1982). *Evaluating practice: Guidelines for the accountable professional*. Englewood Cliffs, NJ: Prentice-Hall.

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement, 44*, 781-804.

Brown, B. W., Jr., & Hollander, M. (1977). *Statistics: A biomedical introduction*. New York: Wiley.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Doueck, H. J., & Bondanza, A. (1990). Training social work staff to evaluate practice: A pre/post/then comparison. *Administration in Social Work, 14*(1), 119-133.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart and Winston.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review, 4*(1), 93-106.

Howard, G. S., & Dailey, P. R. (1979). Response shift bias: A source of contamination of self report measures. *Journal of Applied Psychology, 64*(2), 144-150.

Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response shift bias. *Applied Psychological Measurement, 3*(4), 481-494.

Howard, G. S., Milham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subject response style effects on retrospective measures. *Applied Psychological Measurement, 5*(1), 89-100.

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

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement, 16*, 129-135.

Koele, P., & Hoogstraten, J. (1988a). A method for analyzing retrospective pretest/posttest designs: I. Theory. *Bulletin of the Psychonomic Society, 26*(1), 51-54.

Koele, P., & Hoogstraten, J. (1988b). A method for analyzing retrospective pretest/posttest designs: II. Application. *Bulletin of the Psychonomic Society, 26*(2), 124-125.

Nisbett, R. E., & Wilson, T. D. (1977). "Telling more than we can know:" Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Rhodes, J. E., & Jason, L. A. (1987). The retrospective pretest: An alternative approach in evaluating drug prevention programs. *Journal of Drug Education, 17*(4), 345-355.

Rubin, A. (1992, May). *Methodological issues in using single-case designs to evaluate groups for battered women*. Paper presented at the 8th Annual Symposium on Empirical Foundations of Group Work, Madison, WI.

Sprangers, M. (1988). Subject bias and the retrospective pretest in retrospect. *Bulletin of the Psychonomic Society, 27*(1), 11-14.

Sprangers, M., & Hoogstraten, J. (1987). Response-style effects, response-shift bias, and a bogus-pipeline. *Psychological Reports*, *61*, 579-585.

Sprangers, M., & Hoogstraten, J. (1988). Response-style effects, response-shift bias, and a bogus-pipeline: A replication. *Psychological Reports*, *62*, 11-16.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*(2), 265-272.