Journal of Agricultural Education
Volume 40, Number 4, pp. 28-37
DOI: 10.5032/jae.1999.04028

RESPONSE SHIFT BIAS:A PROBLEM IN EVALUATING LEADERSHIP DEVELOPMENT WITH SELF-REPORT PRETEST-POSTTEST MEASURES

Frederick R Rohs, Professor

University of Georgia

Abstract

This study sought to investigate the effects of response shzft bias on outcomes using a self-report measure in a leadership development course. While students in this study rated themselves as having a "high " level of leadership skill at the end of the course, significant differences were found between their self-report ratings using the pretest/posttest and the then/posttest approach. The degree of response shift (then/post pre/post comparison) was also significant. The findings from this study together with other studies cited suggest that when employing self-report measures, the then/post approach provides a less conservative and more accurate means of assessing leadership skill development than would the traditional pretest/posttest approach.

Introduction

The accurate evaluation of instructional programs and activities is an area of concern to a broad spectrum of educators and social scientists. By continuously monitoring and improving the efficacy of research methodologies, educators can assess the impact of their programs with greater precision and sensitivity.

Instructional programs in leadership development have proliferated in recent years in academic and informal settings. Although it is widely accepted that these activities possess considerable potential for producing change, documenting these changes and benefits have haunted many educators.

Many evaluation studies of leadership development programs have employed some form of introspective self-report measure. If such programs attempt to identify impacts in behavioral change, a typical approach has been to use a pretest-posttest evaluation design to document change. However, this procedure possesses some potential problems. To compare pretest and posttest scores, a common metric must exist between two sets of scores (Cronbach & Furby, 1970). In using self-report measures, educators assume that a person's standard for measurment of the dimension being used will not change from pretest to posttest. If the standard of measurement were to change, the posttest ratings would reflect this shift in addition to the actual changes in the person's level of functioning. Consequently, comparisons of pretest with posttest ratings would be confounded by this distortion of the internalized scale yielding an invalid interpretation of the effectiveness of the program (Campbell & Stanley, 1963, Caporaso, 1973, Neal & Leibert, 1973).

One consequence of most leadership development programs is to change a person's understanding of the leadership skill being measured. One might contend that to the extent the program meets this goal of greater understanding, it will alter each person's perspective in his or her self-evaluation. For example, program participants might feel at pretest that they are "average" leaders with "average" leadership skills. The program changes their understanding of the skills involved in being a leader; after the workshop they understand that their level of functioning was really below average at the pretest. Suppose they improved their leadership skills as a result of their participation in this leader development program and moved from below average to average with respect to their new understanding of leadership. Then their pretest and posttest ratings would be average. If we do not consider that these ratings are based on different understandings of the dimension of leadership, we might erroneously conclude that they had not benefitted from the leadership program. Whenever such shifts in understanding occur, conventional self-report pretest/posttest designs are unable to accurately gauge the impact of instructional programs. Literature reviews cited by Pohl(1982) indicate that often when self-report measures are used, there is a lack of findings of significant differences between pre and posttest measurements.

Several studies (Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Pohl, 1982; Sprangers & Hoogstraten, 1988; Rockwell & Kohn, 1989; Rohs & Langone, 1997) have documented the "response shift bias" phenomenon as a source of contamination of self-report measures that result in inaccurate pretest ratings. To correct this problem, Howard, et al. (1979) recommended that at the posttest session participants are asked to respond twice to each item on the self-report measure. The first asks participants to report their behavior or understanding as a result of the program (post). The second asks participants to report their behavior before the program ("then" rating). The difference between the then and pre self-report ratings is referred to as response shift. Because then ratings and post ratings are made in close proximity, it is more likely that both ratings will be made from the same perspective and thus be free of response-shift bias. Response shift biases have been found in educational settings dealing with knowledge of subject matter and the learning ofbasic helping skills. Studies by Howard (1980) investigating actual changes in amount of material acquired in a college course using self-reports of content learned revealed the then-post approach reflected more accurately the changes in students' knowledge of the subject matter from before to after the class than did the pre-post approach. Similar results were also found by Bray and Howard (1980) with teachers when evaluating a teaching skills training program. In no study comparing then-post and pre-post self-report methods was the pre-post measure superior or even equivalent to the then-post approach in reflecting behavioral indices of change.

Purpose

The overall purpose of this study was to investigate the effects of response shift bias on outcomes using a self-report measure in a leadership skills development course. More specifically 1) did a response shift occur? 2) what was the extent of that shift and 3) what difference in scores occurred between the traditional pretest-posttest design and the then-posttest design method of collecting data?

Methods And Procedures

The data were obtained from students enrolled in two sections of AGR 300, a college-wide undergraduate course in agricultural leadership skills. To increase the internal validity of the study, the classes were randomly assigned to one of two treatment groups (pre-post or then-post design). Students from a college-wide agriculture orientation class (AGR 10 1) served as a control group. Each leadership class was 10 weeks in length and covered such topics as leadership theory, group development and maintenance, conflict management as well as group decision making and consensus building techniques. Throughout the classes students participated in various exercises that allowed them to practice various skills and techniques discussed in class. The instructor and course content remained the same for both leadership classes.

The Youth Leadership Life Skills Development Scale (YLLSDS), developed by Dormody, Seevers and Clason (1993), was used to measure students leadership skill development. The YLLSDS is a 30 item paper and pencil instrument which asks individuals to indicate on a four point scale (O=none, 3=much) the degree to which they posses each skill or characteristic. Total scale values can range from a low of 0 to a high of 90. For descriptive purposes Dormody, et. al (1993) suggest scale values of 0 to 30 to be considered "no to slight leadership skills development," from 31 to 60 "moderate development" and from 61 to 90 "high" development.

According to Dormody, et al. (1993) the YLLSDS was assessed for face and content validity by a panel of faculty from New Mexico State University and field tested with a stratified random sample of 262 New Mexico senior 4-H and FFA members. The Cronbach's alpha reliability coefficient for the scale was .98.

Students in the first AGR 300 group received the YLLSDS on the first and last day of class asking them to rate themselves on each of the 30 items (pre-post group). Students in the second AGR 300 group receive the YLLSDS on the last day of class asking them to respond twice to each item(then-post group). First they were asked to report how they perceive themselves currently (post). Immediately after answering each item in this manner, they were asked to answer the same item again, this time in reference to how they perceive themselves at the beginning of the course (then). Students comprising the control group(AGR 101 class) received the YLLSDS on the first and last day of class as did the first AGR 300 group. A total of 30 students comprised the pre-post group, 28 comprised the then-post group, and 32 comprised the control group.

The data were coded and entered into a computer file on the mainframe. Using SAS 608 the data were summarized and analyzed. Statistical tests were employed to determine if differences existed between groups on pretest measures. Significant differences (p<. 0 1) were found to exist. An analysis of co-variance model was then used to measure statistical differences with the appropriate pretest score as the co-variate. The then score was used as the co-variat for the then-post group since this score represents their beginning course ratings. To establish an overall significance test for each question of .05, the Bonferroni method was employed to determine the significance for each paired test within each question. Since there were three paired tests per question, each pair was tested at the probability level of .05 divided by 3 which for 27 to 3 1 degrees of freedom computes to t-values of 2.53 to 2.55.

Results

Significant differences in mean scores were found between the pre-post, then-post and control groups in 15 of the 30 scale items and in the overall scale score (table 1). This suggests that the leadership course did influence a student's self-reported leadership skill level. Adjusted posttest mean scores indicate that the pre-post and then-post groups, those who participated in the leadership course, achieved higher levels of skill development than those who comprised the control group with the then-post group achieving the highest score (table 2). However, a closer inspection of the group means reveals some striking contrasts (table 3).

While the control group's pretest and posttest scores revealed no significant differences, as was expected, several differences were noted between the pre-post and then-post groups in their pretest and posttest scores. The pre-post group of students completed the leadership self-report measure at the beginning and end of the course and reported no significant differences in scores on any of the scale items or total scale score. While both total scale scores (pre=70.00, post=70.36) suggest a high level of leadership skill development the lack of significant differences between these scores indicates "no change" in their leadership skill development between

Table 1. F- Values and Probabilitv for Posttest Leadership Life Skill Develonment Scores with Pretest Scores (Covariate)

Variable	F-Value	F-Probability
Can determine needs	4.73	,010
Have a positive self-concept	0.90	.410
Can express feelings	0.38	.685
Can set goals	1.99	.145
Can be honest with others	1.92	.159
Use information to solve problems	7.69	.001
Can delegate responsibility	5.84	.004
Can set priorities	14.06	.001
Am sensitive to others	0.36	.696
Am open-minded	0.03	.996
Consider the needs of others	0.75	.474
Show a responsible attitude	1.04	.359
Have a friendly personality	2.35	.102
Consider input from group members	5.55	.005
Can listen effectively	8.34	.001
Can select alternatives	15.59	,001
Recognize the worth of others	10.48	.001
Create atmosphere of acceptance	2.43	,095
Can consider alternatives	10.48	.001
Respect others	8.61	.001
Can solve problems	9.55	.001
Can handle mistakes	7.19	.001
Can be tactful	6.96	.001
Can be flexible	1.83	.169
Get along with others	6.71	.002
Can clarify my values	1.87	.165
Use rational thinking	1.97	.147
Am open to change	4.02	.022
Have good manners	1.13	.329
Trust other people	2.17	.122
Total scale score	18.87	.001

Table 2. Mean Posttest Scores for Leadership Life Skill Development by Group

Variable	Adjusted Mean Score		
	Then/Post	Pre/Post	Control
Can determine needs	2.60	2.30	2.08
Have a positive self-concept	2.43	2.22	2.29
Can express feelings	2.49	2.29	2.46
Can set goals	2.67	2.59	2.39
Can be honest with others	2.35	2.23	2.00
Use information to solve problems	2.42	2.39	1.91
Can delegate responsibility	2.35	2.18	1.71
Can set priorities	2.70	2.37	1.92
Am sensitive to others	2.14	1.94	1.91
Am open-minded	2.10	1.82	1.92
Consider the needs of others	2.51	2.32	2.37
Show a responsible attitude	2.47	2.28	2.44
Have a friendly personality	2.60	2.37	2.26
Consider input from group members	2.77	2.69	2.32
Can listen effectively	2.51	2.52	1.99
Can select alternatives	2.65	2.48	1.80
Recognize the worth of others	2.64	2.55	2.16
Create atmosphere of acceptance	2.61	2.32	2.22
Can consider alternatives	2.69	2.65	2.12
Respect others	2.58	2.50	2.00
Can solve problems	2.46	2.39	1.88
Can handle mistakes	2.52	2.33	1.80
Can be tactful	2.49	2.36	1.91
Can be flexible	2.58	2.42	2.31
Get along with others	2.71	2.57	2.24
Can clarify my values	2.20	1.83	2.01
Use rational thinking	2.28	1.92	1.96
Am open to change	2.26	2.06	1.71
Have good manners	2.19	1.98	1.87
Trust other people	2.31	1.95	2.00
Total score	74	70	61

Table 3. Mean Scores and Test of Significance for Leadership Life Skill Development Scores by Croup

Variable	Then/Post (N=28)			Pre/Posttest (N=30)			Control (N=32)		
	Then	Post	i-value	Pre	Post	i-value	Pre	Post	t-value
Can determine needs	1.48	2.32	7.60*	2.40	2.30	-.520	2.20	2.25	0
Have a positive self-concept	1.72	2.28	4.09”	2.28	2.26	0	2.12	2.27	1.732
Can express feelings	1.80	2.28	3.76”	2.12	2.26	.460	2.36	2.50	2.000
Can set goals	1.92	2.56	3.34”	2.44	2.58	.237	2.52	2.50	0
Can be honest with others	1.68	2.20	2.78”	2.36	2.18	-.646	2.20	2.12	-1.444
Use information to solve problems	1.72	2.32	3.76”	2.28	2.32	.463	2.04	1.95	- .901
Can delegate responsibility	1.72	2.16	3.13”	2.20	2.16	-.188	1.88	1.75	-2.449
Can set priorities	1.80	2.52	6.55*	2.44	2.51	1.000	2.12	1.95	-1.732
Am sensitive to others	1.72	2.00	2.02”	2.28	1.98	-1.780	1.92	1.87	-0.327
Am open-minded	1.40	1.92	3.02”	1.92	1.92	0	1.88	2.00	.901
Consider the needs of others	1.84	2.36	4.49”	2.28	2.37	.213	2.28	2.40	.768
Show a responsible attitude	1.88	2.40	3.13”	2.48	2.39	-.900	2.20	2.35	1.237
Have a friendly personality	2.04	2.52	4.49”	2.52	2.50	-.526	2.04	2.12	1.072
Consider input from group members	1.72	2.60	5.97*	2.40	2.60	1.155	2.52	2.50	0
Can listen effectively	2.00	2.52	3.16”	2.12	2.50	1.621	1.88	1.85	0
Can select alternatives	1.28	2.52	6.99*	2.28	2.52	1.044	2.12	2.07	-1.444
Recognize the worth of others	2.00	2.56	4.21”	2.44	2.54	.224	2.24	2.20	-.810
Create atmosphere of acceptance	1.72	2.40	3.76”	2.56	2.38	-1.365	2.56	2.52	-1.809
Can consider alternatives	1.68	2.56	5.56*	2.40	2.56	1.095	2.48	2.40	-1.444
Respect others	2.04	2.44	3.76”	2.44	2.42	-.253	2.36	2.15	-2.317
Can solve problems	1.80	2.36	4.09”	2.40	2.40	0	2.08	1.92	-1.549

(table continues)

Variable	Then/Post (N=28)			Pre/Posttest (N=30)			Control (N=32)		
	Then	Post	t-value	Pre	Post	t-value	Pre	Post	t-value
Can handle mistakes	1.64	2.40	4.41"	2.52	2.40	-.569	2.08	1.95	-1.000
Can be tactful	1.88	2.40	3.76"	2.36	2.39	.253	2.00	1.87	-1.162
Can be flexible	1.92	2.44	5.34"	2.44	2.44	0	1.92	2.02	1.809
Get along with others	1.84	2.56	4.71"	2.60	2.55	-.271	2.04	2.10	.810
Can clarify my values	1.44	1.96	3.34"	2.36	1.96	-2.76	2.04	2.15	1.000
Use rational thinking	1.40	2.00	4.09*	2.32	1.98	-1.55	2.12	2.20	0
Am open to change	1.52	2.20	3.34"	2.08	2.21	.492	1.84	1.82	-9.011
Have good manners	1.72	1.96	1.81	2.20	1.96	-1.06	2.08	2.07	-1.072
Trust other people	1.64	2.00	2.91"	2.28	2.00	-1.371	2.32	2.22	-1.809
Total scale score	52.04	68.4	6.68"	70.0	70.36	-.079	64.56	63.97	-1.149

* EC.05

administration of the pre and posttests. However, the then-post group reported significant differences in all items on the scale and total scale score. The then-post group reported an overall 16 point gain with a total posttest scale score of 68.40 indicating a "high" level of leadership skill development (table 3). When the then total scale score (x=52.0) from the then-post group and the pretest total scale score (x=70.0) from the pre-post group were compared, significant differences (t=-4.46, p<.001) were found indicating a response shift effect totaling 18 points.

Discussion

Both groups of students experienced the same educational program and instructions. They were measured using the same leadership assessment instrument yet reported very different levels of impact. While both groups' posttest scores were roughly equal (Table 3), a comparison of the two groups' pretest scores indicates that the pre-post group rated themselves higher in the beginning on all 30 leadership skill items than did the then-post group. This comparison suggests a response shift may have taken place in the participants of the pre-post group since no change between pre and posttest scores was reported, giving the impression that the course was not effective. Students comprising the then-post group reported more dramatic changes in scores since they were evaluating themselves with the same standard of measurement or level of understanding on both their responses, how they felt now (post) and how they felt at the beginning of the course (then). Thus, their pretest rating and difference between their pretest (then) and posttest rating of leadership skill level reflects a more accurate assessment of their change in leadership skill level than did those students who rated themselves at the beginning and at the end of the course (pre-post group).

This study provides additional evidence of the impact of response shifts on self-report ratings. The then-post procedure provided radically different results with which to evaluate the leadership class compared to the pre-post procedure. The response shift effects, differences between the then and pretest scores, are treatment dependent. While the use of a control group can address extraneous variables not accounted for by the treatment, it cannot eliminate the danger of such an instrumentation effect. The score on a given scale may have a different meaning for the treatment group than for those in the control group. Response shift theory provides a plausible explanation for these findings. An increase in the students' 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 leadership posttest scores which were lower than their pretest scores. However, other explanations are also possible. For example, these same results might have occurred if (1) students remembered their pretest rating and level of functioning and consciously over represented their posttest level rating or under rated their pre course level on the then pretest to report a positive experience or (2) biased their reports to provide the instructors with more favorable results. However, the time period between the administration of the instrument, 10 weeks in this study, would not enhance the students memory. Students in the pre-post and control groups were also asked on their posttest to record what they thought was their pretest score. No accurate recalls occurred. All students participating in the study were assured that their responses were confidential and would not influence their class grade. Other studies (previously cited) also refute these alternative explanations.

While this study took place in the college classroom where other objective measures could be employed, the then-post analysis yielded a drastically different set of conclusions regarding the effectiveness of the leadership class from the pre-post approach. The then-post data revealed that the course produced major changes in the leadership skills of students verses a "no change' conclusion using pre-post data.

Although there may be alternative explanations for then-post pre-post differences, the position taken in this study is that response shifts are the result of changes in a student's understanding or standard of measurement regarding leadership skills. Since leadership courses seek to enhance a student's understanding and leadership skill level, we can be fairly confident that changes in a student's standard of measurement or level of understanding will be affected. For example, students in this study discovered through group activities, discussion and practical exercises that they were not as skilled as previously thought. They were not as open minded and subject to change, they were unable to delegate tasks to others in their group, they lacked the ability to listen effectively, be sensitive to others and recognize the worth or needs of group members. These shortcomings, as well as deficiencies in other leadership skill areas as measured by the YLLSDS, led to several problems within their groups requiring students to draw on additional leadership skills to solve these problems. The skills needed in these situations tested their tact, flexibility, trust and rational thinking abilities. When students were made aware of these deficiencies and how it affects one's ability to provide leadership, a change in understanding or standard of measurement occurred. This change in standard of measurement or level of understanding, manifested as response shift, greatly influenced the level and accuracy of outcome measures and the effect of a leadership development program.

Conclusions

Regardless of evaluation design, posttest total mean scores from students in all three groups reported high levels of leadership skill development (61 or above). Although all were in the "high" category, those comprising the control group scored the lowest with a posttest mean score of 63.9 versus a posttest mean scale score of 70.3 for the pre-post group and a score of 70.0 for the then-post group.

Students in the pretest-posttest group reported no significant differences between pre and posttest scores while those in the then-posttest group reported significant differences between then and posttest scores, the then score being significantly lower than the posttest score.

When comparing self-report pretest scores and then mean scores of both treatment groups it appears that students in the pre-posttest group initially tended to over estimate their level of leadership skill. Upon analysis, the then-post groups' mean was significantly different (lower) than the pre-post groups' mean suggesting that the difference was likely due to a response shift. These results support previous studies by Howard and Dailey (1979), Bray and Howard (1980), Pohl (1982), Sprangers and Hoogstraton (1988), and Rohs and Langone (1997) which document the effects of response shift when using self-report pretest-posttest measures.

A response shift of 18 points was observed and thus served to provide a less conservative then-post assessment of the change in leadership skill development than did the pre-post method.

Recommendations and Implications

Given the extent and pervasiveness with which response shift bias has been documented and the superiority of then-post over pre-post methodology in evaluating various educational programs and activities, it is recommended that researchers collect then pretest data along with traditional pre and post self-ratings. If other objective and behavioral measures are available integrating them will help to provide a more complete assessment of change.

The adequacy of the measure used affects the quality of the findings. This study employed a measure with established validity and reliability. However, this researcher's experience suggests that many leadership self-report measures lack this credible foundation. While valid and reliable leadership measures exist they are often more difficult and costly to obtain, thus, our propensity to rely on self-report measures to collect impact data continues. Reasonable efforts should be made to establish the validity and reliability of all self-report measures used.

Further clarification is still needed regarding the contexts in which then pretest measures might be inappropriate as well as the use, analysis, and interpretation of these measures. Research is lacking that identifies and clarifies the various causal determinants of the response shift. One factor may be the level of information individuals have at the pretest regarding the dimension, in this case leadership skills, on which they are asked to self-report. Another might be the measurement instrument itself. Questions or subscales within the instrument (i.e. cognitive verses attitudinal items) might be more affected by or sensitive to response shift. It would seem, intuitively, that cognitive variables would be less subject to response shift than attitudinal variables.

References

Bray, J. H. & Howard, G. S. (1980). Methodological considerations in the evaluation of a teacher-training program. Journal of Educational Psychology 72(1): 62-70.

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

Caporaso, J.A. (1973). Quasi-experimental approaches to social science: perspectives and problems. In: A.J. Caporaso & L.L. Rooss, Jr. (Eds.), Quasi-Experimental Approaches: Testing Theorv and Evaluating Policy Evanston, Ill. :Northwestern University Press.

Cronbach, L.J. & Furby, L. (1970). How we should measure "change''--or should we? Psvcholoaical Bulletin 74:68-80.

Dormody, T., Seevers,, B.S. & Clason, D. L. (1993). The Youth Leadership Life skills Development Scale: An Evaluation and Research Tool for Youth Organizations. (Research Report

672) New Mexico State University, Agricultural Experiment Station.

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 in self report measures. Journal of Applied Psychology 64 (2): 144-150.

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 pre-tests. Applied Psvcholonical Measurement 3 : l-23.

Neale, J. M. & Leibert, R.M. (1973). Science and Behavior: An Introduction to Research Methods. Englewood Cliffs, N. J.: Prentice Hall.

Pohl, N.F. (1982). Using retrospective pre-ratings to counteract response-shift confounding. Journal of Experimental Education 50 (4):21 1-214.

Rockwell, S. K. & Kohn, H. (1989). Post-then pre evaluation. Journal of Extension 27 (2): 19-21.

Rohs, F.R. & Langone, CA. (1997). Increased Accuracy in Measuring Leadership Impacts. Journal of Leadership Studies 4(1): 150-159.

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