Improving the Evaluation of Leadership Programs: Control Response Shift

Frederick R. Rohs
Professor and Extension Staff Development Specialist
109 Four Towers
Department of Agricultural Leadership,
Education and Communication
The University of Georgia
Athens, Georgia 30602
frrohs@uga.edu

Abstract

The Cooperative Extension Service has been a key partner in the design, implementation and evaluation of leadership development programs. To evaluate the effectiveness of their training and the effects of response shift bias on outcomes using a self-report measure, one hundred forty-seven County Extension Agents participated in this leadership study. Participants were randomly assigned to one of two treatment groups or to a control group. Two different evaluation designs (pre-posttest and then-post) were used. The then-post design asks participants to first report their behavior or understanding as a result of the training (post) and then to retrospectively report this behavior before the training. The then-post evaluation design provided more significant change data than did the traditional pre-posttest design indicating a response shift occurred. Such differences in evaluation findings suggest that the educational benefit of such trainings may be underestimated when using the traditional pre-post evaluation design.

Introduction

The planning, implementation and evaluation of educational programs represents a substantial portion of the services provided by Cooperative Extension Service. Although it is widely accepted that these programs possess considerable potential for increasing knowledge and producing change, documenting these changes have haunted many educators. This disparity between the willingness to design and implement programs and the reluctance to expend equal energy and resources in their evaluation is understandable. Organizational incentives in training and development programs have favored design and implementation rather than evaluation (Campbell, 1971). One educator aptly commented "we favor program takeoffs, not their endings."

Many evaluation studies of educational and training programs have employed some form of introspective self-report measures. If such programs attempt to identify impacts in behavioral change, the typical approach has been to use a pretest-posttest evaluation design. 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 and practitioners assume that a person's standard for measurement of the dimension being assessed 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, Neale & Leibert, 1973).

One consequence of most leadership development programs is change in a person's understanding of the leadership skills 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, a participant 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 pretests. Suppose they improved their leadership skills as a result of their participation in this leadership 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 new ratings are based on different understandings of the dimension of leadership, we might erroneously conclude that they had not benefited from the leadership program. Whenever such shifts in understanding occur, conventional self-report pretest-posttest designs are unable to accurately gauge the impacts of these programs. Literature reviews 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 have documented the "response shift bias” phenomenon as a source of contamination of self-report measures that result in inaccurate pretest ratings (Rockwell & Kohn, 1989, Howard & Daily, 1979, Howard, Ralph, Gulanick, Maxwell, Nance & Gerber, 1979, Pohl, 1982, Sprangers & Hoogstraten, 1988, Rohs & Langone, 1996). To correct this problem it has been recommended that at the posttest session participants be asked to respond twice to each item on the self-report measure (Howard et al, 1979). 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) thus eliminating the pre-test.

The difference between the "then" and "pre" self-report ratings is referred to as a 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.

This paper presents results from an internal Extension leadership development program employing the post-then-pre method of evaluation.

SELD: Southern Extension Leadership Development

During the past decade, the Cooperative Extension System has faced an era of economic scarcity and has been impacted by a number of internal and external challenges (Ladewig & Rohs, 2000). Many of these changes and challenges have changed the nature of work and relationships. Organizations that respond to the changing nature of work and authority relationships are learning organizations (Senge, 1990).

A major challenge impacting the transition to a learning organization is that few Extension administrators are professionally trained in competencies and styles of leadership appropriate for learning organizations. Rather, they have been promoted to leadership positions because they excelled in their subject-matter discipline, and they learn their new craft by emulating those who proceeded them. While this practice is commonplace throughout the industrialized world, these administrators often lack the necessary leadership competencies necessary to truly transform their organizations to compete in the information technology era (Patterson, 1998).

In response to the growing need to understand and cope with the many changes currently and potentially impacting the Extension System, Cooperative Extension Directors and Administrators of the Southern Region called for the establishment of a regional leadership development program. The result was the formation of Southern Extension Leadership Development (SELD).

The SELD program is unique in that the competency-based approach builds around the skills individuals and groups in Cooperative Extension need to be effective in the future. With such knowledge, Extension educators can design professional development plans that are relevant, useful, and customized to their needs. While regional workshops were conducted individual states were encouraged to implement their own leadership development program.

The centerpiece of SELD is the Managerial Assessment of Proficiency (MAP), developed by Training House, Inc. of Princeton, NJ. The assessment portion is a video-driven, competency-based, computer-scored simulation consisting of 200

items that assesses a participant's proficiency in 12 competencies. The twelve competencies are: time management, setting goals, planning and scheduling work, training, coaching and delegating, appraising people and performance, disciplining and counseling, listening and organizing, giving clear information, getting unbiased information, solving problems, making decisions and weighing risk, thinking clearly and analytically. The assessment portion was followed with a series of competency building workshops to strengthen participants weaker competency areas.

Evaluation Methodology

The data for this study were obtained from county Extension Agents participating in the Georgia SELD program between 1998 and 2001. The program was conducted in various geographic locations through out the state. Participants were assigned to one of two treatment groups (received programs) based on location. Agents participating in another program served as a control group. A total of 46 agents were assigned to the pre-post group, 52 to the then-post group and 49 comprised a control group.

A paper-pencil self-report measure was used to gather information on agents competency levels. The measure sought participants reaction to 12 managerial competencies on a 4-point scale (1 = none, 5 = much).

The self-report measure was administered to Extension agents in the pre-post group at the beginning and at the conclusion of the SELD program (6 weeks later). Agents in the then-post group were given the self-report measure immediately following the last workshop session and asked to respond to each item twice. First how they perceive themselves to be at the present (post), then how they perceive themselves at the beginning of the SELD program. The self-report measure was administered to agents in the pre-post defined group in the same manner as the pre-post group. However, the measure itself defined each competency for the participants.

The SAS statistical program was used for data analysis (SAS, 1997). Paired t-test were used to compare means between pre and posttest scores for each item across the three groups (Table1). To evaluate the statistical differences in Table 2, a one-way analysis of variance (PROC GLM) was employed. To identify specific differences between item means for the three groups, the Duncan Multiple Range Test was used as a post hoc assessment. 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 test per question, each pair was tested at the probability level of .05 divided by 3 which, for 46 to 52 degrees of freedom, computes to t-values of 37 to 40.

Results

Significant differences were found between pretest and posttest scores in the three treatment groups (Table 1). The pre-post group and control group of Extension staff who completed the self-report measures at the beginning and at the conclusion of the training only reported significant differences in scores on three of the competencies. However, the then-post group reported significant changes in all twelve competencies. Differences in pre and posttest scores indicate that the then-post group, reported more increase in their competencies than those in the pre-post groups (Table 1). A closer inspection of group pretest mean scores reveals some differences (Table 2). The then-post group reported significantly lower mean pretest scores on five of the twelve competencies than did the pre-post groups.

Table 1: Mean Scores[a] and Test of Significance for Competencies

	Then/Post (N=52)				Pre/Post (N=46)				Control (N=49)			
Variable	Then	Post	T	P	Pre	Post	T	P	Pre	Post	T	P
Time Management & Prioritizing	2.63	2.94	2.47	*	2.70	2.69	-	ns	2.73	2.79	-	ns
Setting Goals & Standards	2.46	3.11	6.63	*	2.90	2.69	-	ns	2.87	2.97	-	ns
Planning & Scheduling Work	2.75	3.40	5.20	*	3.00	2.86	-	ns	2.87	3.04	-	ns
Listening & Organizing	2.55	3.40	6.23	*	3.08	3.00	-	ns	3.18	3.16	-	ns
Giving Clear Information	2.56	3.21	6.90	*	3.41	3.08	-2.84	*	2.77	3.02	2.00	*
Getting Unbiased Information	2.59	2.90	2.47	*	2.82	3.26	-	ns	2.69	2.79	-	ns
Training, Coaching & Delegating	2.65	3.03	3.04	*	2.84	2.84	-	ns	2.59	2.85	-	ns
Appraising people & Performance	2.38	2.75	2.75	*	2.86	2.84	-	ns	2.53	2.69	-	ns
Disciplining & Counseling	2.59	2.90	3.04	*	2.30	3.02	3.09	*	2.36	2.55	2.02	*
Identifying & Solving Problems	2.65	2.96	2.21	*	3.41	2.71	-7.99	*	2.89	2.91	-	ns
Making Decisions, Weighing Risk	2.75	2.96	2.28	*	2.97	2.82	-	ns	2.85	3.04	-	ns
Thinking Clearly & Analytically	2.44	2.76	2.76	*	3.28	3.15	-2.59	*	3.24	2.65	-5.64	*

[a]1 = none, 2 = little, 3 = some, 4 = much; * $p<.05$

Both participant groups experienced the same educational program, the same instructors and the same assessment items yet reported very different levels of impact. These data suggest that a response shift may have taken place in the participants who were asked to answer each assessment item twice after the training (then-post group). In doing so, these participants were evaluating themselves with the same standard of measurement or level of understanding on both their posttest responses (how they felt now) and ○ᴺʟ®⅝—○ responses (how they felt before the program). Thus, the comparisons of their ○ᴺʟ®⅝—○ ratings of managerial competencies to their post ratings reflect a more accurate assessment of their competency level than did those in the participant group who rated themselves at the beginning of the training and again at the conclusion (pre-post and control group). A comparison of the two participant groups' pretest scores (Table 2) indicates that the pre-post group rated themselves significantly higher ($p<.05$) than the then-post group in the beginning on six of the twelve competencies.

Table 2: Analysis of Variance by Group for Mean Pre Post Competency Scores

Variable	Pre-Test Scores [1,2] Then/Post	Pre/Post	Control	F Value	F Probability
Time Management & Prioritizing	2.63	2.70	2.73	.29	.7466
Setting Goals & Standards	2.46 [a]	2.90 [b]	2.87 [b]	7.16	.0001*
Planning &Scheduling Work	2.75	3.00	2.87	2.08	.1291
Listening & Organizing	2.56 [a]	3.08 [b]	3.18 [b]	10.63	.0001*
Giving Clear Information	2.56 [b]	3.41 [a]	2.77 [b]	26.00	.0001*
Getting Unbiased Information	2.59	2.82	2.69	1.56	.2127
Training, Coaching & Delegating	2.65	2.84	2.59	1.56	.2130
Appraising People & Performance	2.38 [b]	2.86 [a]	2.53 [b]	6.73	.0016
Disciplining & Counseling	2.59 [a]	2.30 [b]	2.36 [b]	2.84	.0616
Identifying & Solving Problems	2.65 [a]	3.41 [b]	2.89 [c]	19.52	.0001*
Making Decisions, Weighing Risk	2.75	2.97	2.85	1.61	.2026
Thinking Clearly & Analytically	2.49 [a]	2.76 [b]	2.65 [b]	20.10	.0001*

[1] Pre-test mean scores having a common letter within rows are non-significant at probability ($p \leq .05$)
[2] 1 = none, 2 = little, 3 = some, 4 = much

Discussion

Evidence of response shift biases has been found in educational settings dealing with knowledge of subject matter and the learning of basic helping skills (Howard & Daily, 1979). Extensive literature review indicates that when self-report measures are used, significant differences between pretest and posttest measurements with the pre-post group are not often found (Pohl, 1982).

These findings provide evidence of the impact of response shifts on self-report ratings of leadership competencies of program participants. The then-post procedure provided different results with which to evaluate these programs compared to the pre-post procedure. The response shift effects are treatment dependent, occurring in treatment groups not in control groups. While the use of control or comparison groups allow educators to more effectively document program impacts by controlling for extraneous variables, because of response shifts they are unable to accurately provide the comparison sought and may be impractical to use. The score on a given scale may have a different meaning for the participant group than for those in the control group. Response shift theory provides a plausible explanation for these findings. An increase in the participants' 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 lower than the other participants' pretest scores. However, other explanations are also possible. For example, these same results might have occurred if (1) participants remembered their pretest rating and level of functioning and consciously over represented their posttest level rating or underrated 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. Other studies (Howard et al. 1979) however, refute these alternative explanations. In a study involving undergraduate students enrolled in two sections of a communications course at a mid-western university, meeting at different times with different instructors, a self-report measure was administered to assess students' perceptions of their own helping skill level. Within each section, students were randomly divided into pre/post, then/post, and all test (pre plus then/post) groups. Students in the all test group were asked to record what they remembered their pre course skill rating to be. Memory ratings were not accurate and were significantly ($p<.05$) lower that their actual pre or then ratings (Howard et al. 1979). Similar results were found in an undergraduate leadership class where students assessed their leadership skill level using a self-report measure. Students in their class completed a pretest at the beginning of the course and then a posttest followed by a retrospective "Then" tests. After completing the "then" test, they were asked to report their pretest score from the first class period. Not one accurate pretest score was reported (Rohs & Langone, 1997).

While there may be several alternative explanations for then-pre differences, the position taken in this study is that response shifts are the result of changes in

participants' understanding or standard of measurement regarding the practices, skills or competencies being taught. This change in standard of measurement or level of understanding manifested as response shift greatly influences the level and accuracy of outcome measures for educators. In many cases, the traditional pre-post approach might indicate "no change" in understanding when improvements actually occurred (Rohs & Langone, 1996). In this study and others, the effects of response shifts served to provide a less conservative and perhaps more accurate assessment of the program, than did the more traditional pre-post design. Results from this study support earlier studies and recommendations that when self-report measures are employed to assess attitudinal consistency or behavior intentions across time and when a change in the standard of measurement or level of understanding is anticipated, it is advisable to collect ○ᴺL®⅝—○ data in place of or in addition to traditional pretest measures.

Implications

Overall implications of this study underscore the need to more accurately assess the impacts of leadership programs when employing self-report measures. While this study focused on the pre-post and then-post designs, and the resulting differences in outcomes, obtaining valid and reliable impact data requires careful thought and analysis.

When the central goal of the leadership program is to alter participant understanding of a key concept, pre/posttest designs, whether single group, quasi-experimental or experimental, may encourage response-shift bias and inaccurately assess program impact. Conversely, the then-post design can avoid these problems by employing a consistent framework for key concepts.

While many employ self-report measures to evaluate leadership programs, this study suggests that such measures be used with caution. Additional steps could be taken to collect and integrate other objective and behavioral measures along with qualitative follow-up data, such as observations documenting change. These additional data sources, coupled with the pretest and then-post data will help to provide a more complete assessment of change. A future study could be designed in which half of the then/post group be randomly selected and given the pretest, thus allowing a comparison of pre and then scores within the same sample.

Leadership educators need tools to accurately assess the impact of their training. The then- post design offers several advantages. These include ease of use (administered once at the end of the program), and under certain circumstances, increased accuracy and better documentation of results. However, a disadvantage of this design is the lack of understanding or clarity on the part of the respondent completing the self-report measure. Directions must be clear and concise to obtain valid data. Several studies have shown the value of the then-post design. Given that time and resources for evaluation are often limited, the then-post

design offers a useful and efficient tool for documenting the impact of educational programs which assess attitudinal variables or behavioral intentions over time. It is not, however, a panacea for the measurement of all education programs and should be used with caution.

References

Campbell, J.P. (1971). Personnel training and development. *American Psychological Review* 22:526-602.

Campbell, D.T, & Stanley, J.C. (1963). Experimental and quasi-experimental design for research and teaching. In: Gage, NL ed., *Handbook of research on teaching*. Chicago, Ill., Rand McNally.

Caporaso, J.A. (1973). Quasi-experimental approaches to social science: perspectives and problems. In: Caporaso, AJ & Roos, LL eds., *Quasi-experimental approaches: testing theory and evaluating policy*. Evanston, Ill., Northwestern University Press.

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

Howard, G.S. & Daily, P.R. (1979). Response-shift bias: a source of contamination in self report measures. *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 Psychological Measurement* 3:1-23.

Ladewig, H. & Rohs, F.R. (2000). Southern Extension Leadership Development: Leadership development for a learning organization. *Journal of Extension*, 38(3), http://www.joe.org/joe/2000june/az.html.

Neale, J.M. & Leibert, R.M. (1973). *Science and behavior: an introduction to research methods*. Englewood Cliffs NJ: Prentice Hall.

Patterson, T.J. (1998). Comentary II: a new paradigm for extension administration. *Journal of Extension*, 36(1), http://www.joe.org/joe/1998february/comm1.html.

Pohl, N.F. (1982). Using retrospective pre-ratings to counteract response-shift confounding. *Journal of Experimental Education*, 50(4),211-214.

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

Rohs, F.R. & Langone, C.A. (1996). Measuring leadership skills development: a comparison of methods. In: *Association of Leadership Educators Proceedings*, Burlington, Vermont 7:73-78.

Rohs, F.R. & Langone, C.A. (1997). Increased accuracy in measuring leadership impacts *Journal of Leadership Studies* 4(1):150-158.

Rohs, F.R. & Langone, C.A. (1998). Response-shift bias in student self-report assessments. *NACTA Journal* 42(1):46-49.

SAS Institute, Inc. (1997). *SAS procedures guide*, Version 608, 5th Ed. Cary, NC: SAS Institute.

Senge, P.M. (1990). *The fifth discipline. The art and practice of the learning organization*. New York: Doubleday.

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