

1. Name of Presenters: Frederick R. Rohs and Christine A. Langone, Professors,
Department of Agricultural Leadership, Education and Communication
The University of Georgia, 109 Four Towers, Athens, GA 30602
706-542-8828, frrohs@uga.edu.
2. Title: The Challenges of Measuring Leadership Development in College Students.
3. Track: Research
4. Description: This study sought to determine the change in level of understanding of leadership skills by undergraduate students in a college leadership course. The findings from this study together with other studies cited suggest that when employing self-report measures, the post/then approach provides a less conservative and more accurate means of assessing a student's knowledge and understanding of the subject than would the traditional pretest/posttest approach.
5. Bio(s); Frederick R. Rohs-professor and Graduate Coordinator for the Master of Agricultural Leadership (MAL) Program. Advises graduate students and teaches core courses in the MAL Program and Coordinates the 13 state Southern Extension Leadership Development (SELD) Program.

Bio: Christine A. Langone, Professor and Program Coordinator for the Interdisciplinary Certificate in Leadership and Service at The University of Georgia. Advises graduate students and teaches Foundation of Leadership courses for both undergraduate and graduate students as well as teaching core courses in the MAL Program.

The Challenge of Measuring Leadership Development with College Students

Introduction

Measurement of change in student assessment to determine improvement in learning, behavior, or attitude has challenged educators for decades. Change scores have been used to assign grades, especially in classes with students of heterogeneous backgrounds and different previous instructional achievement levels, as an attempt to assure fairness and maximization of individual potential. Given the current strong interest in student-centered instruction, teachers may ask students to assess their own growth or change in learning and attitudes using pre- and post-instruction self-assessment measures (Lam & Bengo, 2003).

The pre-post measures design is plagued with confounding variables that can render the change scores invalid as measures of actual change in individuals (Cook & Campbell, 1979). Although recent effort has begun to address these extraneous variables or validity threats, for example, through logical argument based on careful review of the plausibility of each threat within the context of the research or evaluation (Eckert, 2000), conclusions about the internal validity of findings drawn from this approach is nonetheless speculative (Lam & Bengo, 2003). Further research is needed to support this approach in college classrooms.

Literature Review

Among the validity threats inherent in the pre-post measures design are those caused by the effects of self-reporting in the pretest. This pretest effect can also confound findings from quasi-experimental comparison groups research, if differential pretest effects between-groups are observed as an interaction between pretest and selection bias (Willson & Putnam, 1982). Since as far back as the 1950s, researchers have investigated and documented this pretest effect.

Welch & Walberg (1970) summarized findings from 10 studies on pretest effects conducted between 1949 and 1967. They found no pretest effect in four studies, treatment interference in three studies, and a pretest effect (increased posttest scores as a result of the pretest) in three studies. The authors concluded that pretest effects are less likely with cognitive tests than with affect measures, and less likely with long pretest-posttest intervals (6 months or more). Bracht & Glass (1968) obtained similar findings.

Willson & Putnam (1982) conducted a meta-analysis of results from 32 studies that investigated pretest effects. They found for cognitive tests, 81% of the groups with pretests performed better than groups without pretests. Pretest effects also occurred with attitude measures. Their meta-analysis found that for attitude measures, 62% of those pretested, on average, performed better than the non-pretested. Willson & Putnam's findings of a stronger pretest effect with cognitive measures than with affect measures, which they attributed to reactive interference of treatments in cognitive tasks, contradict the findings of Welch & Wallberg (1970) and of Bracht & Glass (1968). Willson & Putnam (1982) also found a large negative pretest effect with personality measures, reflecting a general change toward an undesirable psychological state.

The pretest effect represents a source of invalidity in research and program evaluation. Specifically, in cognitive measurements, a pretest can lower internal validity by introducing a practice or carry over effect, when participants doing the posttest recall their responses made in the pretest, thus inflating their performance on the posttest. Willson & Putnam (1982) suggested that a similar pretest effect occurs with attitude measures, because of “halo effects in which everyone feels better on second attitude testing” (p. 256). Such was the case in a study by Maltz, Gordon, McDowall, & McCleary (1980) which found that results obtained from pretest-posttest designs can lead evaluators to erroneously conclude delinquency programs as effective.

Besides the carry-over effect, pretests can introduce an opposite, interference effect, due to repetitive testing, which can also lower the degree of internal validity. In this situation, pretesting interferes with the participants’ responses in the posttest (either in cognitive or affective), by causing the participants to feel bored, fatigued or both. While carry-over effects can inflate performance on the posttest, interference effects can deflate it (Lam & Bengo, 2003).

A pretest can also create sensitization and response-shift biases, which represent further threats to construct validity. Pretest sensitization bias occurs when the pretest stimulates the participant’s curiosity and motivation, or orients the participant’s attention toward certain aspects of the treatment, or program inducing the participant’s position or attitude to be slanted in a particular direction (Hoogstraten, 1979).

Business, industry and society are telling colleges and universities that there is a grave need for leadership and human resource preparation for today’s students to succeed in the work place (Brown & Fritz, 1993). Numerous studies have shown that employers, alumni, and students nationwide recognize the need for competence beyond technical

skills. Alumni from the University of Georgia (CAES, 1998) and Texas A & M (1998) indicated that skills in people interactions, communication, problem solving, and conflict management, higher thinking skills/critical thinking, teamwork, leadership were skills related to professional success, and that their undergraduate experience was lacking in providing real life situations in the classroom to address the need for these skills.

Employers also supported the need for skills in communication, problem solving and teamwork as reported in studies by the University of Nebraska and Pennsylvania State University (Andelt , Leverne & Bosshamer 1997; Radhakrishna & Bruening 1994).

Post secondary institutions throughout the nation have responded to this need by providing leadership education in many forms. Along with the need for leadership preparation comes the need to document program results for accountability and funding purposes.

A frequently used method of tracking learning in leadership programs is the self-report assessment measure. These introspective measures may vary from a listing of skills learned (Russell & Jones, 1995) or keeping a leadership journal (Dormody, 1996) to employing standardized assessment measures (Brungardt & Crawford, 1996) or similar rating scale. Youth organizations such as FFA and 4-H also have sought to measure their impact on a youth's leadership development. When such introspective measures are employed in the classroom, the conventional pretest/posttest method of evaluation is often used. In these instances differences between pretest and posttest ratings may appear to be non-existent when actually significant differences exist. The faculty need a more accurate measurement of behavioral change than the conventional pretest/posttest method.

One consequence of most leadership development courses is changing a person's awareness or understanding of the leadership skill being measured. For example, a class participant might feel at the beginning of a course that they are an "average" leader with "average" leadership skills. The course changes their understanding of leadership skills; after the course they understand their level of functioning was below average at the beginning of a course. Whenever such a shift in understanding occurs, conventional self-report pretest/posttest designs are unable to accurately gauge the impact of instructional programs.

Several studies (Howard & Dailey, 1979; Howard et al., 1979; Pohl, 1982; Sprangers & Hoogstraten, 1988; Rockwell & Kohn, 1989; Rohs & Langone, 1997; Rohs, Langone & Coleman, 2001) have documented the "response-shift bias" phenomenon as a source of contamination of self-report measures that result in inaccurate pretest ratings and seriously compromise any assumption of internal validity.

Evidence of response shift biases have been found in college classrooms dealing with knowledge of subject matter and the learning of basic helping skills (Howard et al., 1979; Pohl, 1982). Extensive literature reviews 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.

To correct this problem, Howard et al., (1979) recommends 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/understanding as a result of the program (post). The second asks participants to report their behavior before the program (then). 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.

The purpose of this study was to determine the degree of response shift in the self report ratings of leadership skills development by undergraduate students enrolled in leadership development course.

Methods

Data were obtained from 28 students enrolled in AGR300—an undergraduate course in agricultural leadership skills at the University of Georgia for students who want to learn more about leadership and decision making skills. The 10-week course covers such topics as leadership theory, stages of group development, group maintenance skills, listening and feedback skills, conflict management and several techniques relating to group decision making and consensus building. Throughout the course students participate in various exercises that allowed them to practice skills being discussed in the class.

The Youth Leadership Life Skills Development Scale (YLLSDS), developed by Dormody et al., (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 (0=none, 3=a lot) the degree to which they possess each skill or characteristic. Total scale values can range from 0 to 90. For descriptive purposes, Dormody et al. (1993) suggested scale values of 0 to 30 as “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.

The YLLSDS was administered on the first day of class asking students how they would currently rate themselves on each of the 30 items (PRE). The YLLSDS was again administered on the last day of class asking students to respond twice for each item. First they were to report how they perceived themselves to be at the present (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 perceived themselves at the beginning of the course (THEN).

Data were summarized and analyzed using SAS 608. Statistical tests (t-test for matched groups) were employed to determine if differences existed between the sets of scores testing evidence of response shift.

Results and Discussion

Ninety six percent of the students rated themselves as "high" in leadership skills on the posttest (Table 1.) However, their pretest self-report ratings also revealed they felt they were "high" in leadership skills development. The students "then" ratings revealed a different story. Students "then" ratings indicated that only 14% rated their leadership skills as high with 75% falling into the moderate category and 11% into the low level category.

Table 1. Pretest, then, and posttest levels of leadership skill development scores.

YLLSDS Scores Posttest (z)	Pretest (x)		Then (y)		n
	n	%	n	%	
Low 0-30 % .00	0	.00	3	.12	0
Moderate 31-60 .04	5	.18	21	.75	1

x = x = 70.0, SD = 10.8

y = x = 52.0, SD = 17.5

z = x = 70.0, SD = 8.2

No significant differences were found between the pretest and posttest means (Table 2). The posttest score (X =70.0) and the “then” score (X =52.0) were significantly different. Thus, students felt their leadership skills had improved since the beginning of the course with the post/then method. To determine the response shift in students self-report rating, “pretest” means were compared with their “then” means. The difference between the two means or response shift was 18.0 points in students with perception of their leadership skills development.

Table 2. Means, standard deviations and test of significance of self-report leadership skill scores by condition.

Condition T	Pre		Then		Post	
	Mean	SD	Mean	SD	Mean	SD
Pre/Posttest NS	70	10.8			70	8.2
Then/Posttest			52	17.5	70	8.2

*** = .001 significance level

This study provides evidence of the impact of response shifts on self-report ratings of leadership skill by students enrolled in the leadership class. 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” pretest and the pretest, are treatment dependent. While the lack of a control group may limit this study, it should be noted that the danger of such an instrumentation effect cannot be eliminated by the use of a control group. The score on a given scale may have a different meaning for the “treatment” group than for those in the “control” group (Rohs & Langone, 1997). 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 “then” 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 overrepresented their posttest level rating or underrated their pre course level on the retrospective/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 pretest and the posttest/then procedure (10 weeks) would not enhance the students’ memory. Students were also asked on their posttest to record what they thought was their pretest scale score. No accurate reading occurred. The students were also told at both administrations that their responses were confidential and would not be taken into account when class grades were computed. Studies by Howard et al. (1979) also refute these alternative explanations.

Conclusions

The Then/Post analysis provided a drastically different set of conclusions regarding the effectiveness of the leadership class from the Pre/Post approach. The Then/Post data revealed the course produced major changes in the leadership skills of students versus a “no change” conclusion using Pre/Post data. Furthermore, the Pre/Then data indicate a “response shift” or change in the level of understanding of leadership skills by students took place during the course. Studies of college courses by Howard (1980), Bray & Howard (1980), Pohl (1982) have produced similar results.

Findings from this study suggest that the Then/Post approach provides a more accurate estimate of measuring change. Pre/Post methods remain popular, therefore further research is needed to assess the conditions under which a Pre/Post method would be more appropriate. Additionally, research is needed to identify and clarify the various casual determinants of the response shift. One factor may be the level of information students have at the pretest regarding the dimension, in this case leadership skills, on which they are asked to self-report.

To lessen response-shift bias “informed pretests” may be employed where a thorough description of the variable being measured is provided to the student prior to the administration of the self-report pretest. As with all research the adequacy of the measures used affects the quality of the findings. While this study employed the YLLSDS scale as a valid and reliable measure, our experience with leadership skills assessment has been that most self-report measures/studies do not. Integrating self-report, objective and behavioral measures, if possible, may help to provide a more complete assessment of change. Use of pretest, posttest and retrospective/then pretest

self-report data will provide a more sensitive assessment of a student's perspective of personal changes and skill development.

Researchers/leadership educators/evaluators' knowledge of retrospective pretest and 'response shift' is currently far from comprehensive. While the literature and this study indicate that change measurements using the retrospective pretest method are often more accurate estimates of change, other variations or alternative forms of this method have been used. These include the 'perceived change' method which asks participants or students to estimate, after the class or program, the amount and direction of change they have undergone. A further variation of the 'perceived change' method is the 'post + perceived change'. The post + perceived change method asks participants to report their status at posttest time and to also estimate the amount and/or direction of change. Lam & Bengo (2003) report that such methods have been used to measure changes in schools, teacher practices, and student learning as a result of the introduction of large-scale achievement tests. A fourth variation of the retrospective method of measuring change is called the 'post-only' method. The post-only method obtains only posttest status data and estimates the general pretest rating for all the participants. This is done by reviewing the literature and/or making logical deductions involving some group consensual process.

The post + retrospective pretest method has received considerable attention from both researchers and program evaluators regarding its effectiveness in measuring change. Such is not the case with the other non-pretest methods. Only one study (Lam & Bengo, 2003) has attempted to compare the alternative methods previously mentioned and concluded that the post + retrospective pretest method was the best way to measure change and minimize response shift bias. Further research is clearly needed in this area.

Literature Cited

- Andelt, L. L., Leverne B. & Bosshamer, B., (1997). Employer assessment of the skill preparation of students from the College of Agriculture and Natural Resources, University of Nebraska-Lincoln: implications for teaching and curriculum. *NACTA Journal*, 41 (4), 47-53.
- Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal* 5:437-474.
- 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.
- Brown, W.F. & Fritz, S.M. (1993). Determining the breath of leadership and human resource management development offerings in post secondary Departments of Agricultural Education. *NACTA Journal* 37 (3):11.
- Brungardt, C. & Crawford, C. B. (1996). A comprehensive approach to assessing leadership students and programs: preliminary findings. *Journal of Leadership Studies* 3 (1):37-48.
- College of Agricultural and Environmental Sciences (1998). *1997 Program Review: An analysis of internal and external perceptions of 96 Program areas crossing teaching, research and service*. Athens, GA: University of Georgia CAES.
- Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimental design and analysis issues for field settings*. Chicago: Rand McNally College Publishing Company.
- Dormody, T. (1996). 30 question leadership journal. *Journal of Leadership Studies* 3 (2):75-81.

- Dormody, T., Seevers, B.S. & Clason, D. L. (1993). The youth leadership life skills development scale: an evaluation and research tool for youth organizations. Los Cruces: New Mexico State University. (Agricultural Experiment Station Bulletin No. 672).
- Eckert, W. L. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank institute. *American Journal of Evaluation* 21: 185-193.
- Hoogstraten, J. (1979). Pretesting as determinant of attitude change in evaluation research. *Applied Psychological Measurement* 3:25-30.
- 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 Psychological Measurement* 3:1-23.
- Lam, T. C. & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instruction practice. *American Journal of Evaluation* 24 (1): 65-80.
- Maltz, M. D., Gordon, A. C., McDowall, D., & McCleary, R. (1980). An artifact in pretest-posttest designs: How it can mistakenly make delinquency programs look effective. *Evaluation Review* 4:225-240.
- Pohl, N.F. (1982). Using retrospective pre-ratings to counter act response-shift confounding. *Journal of Experimental Education* 50 (4):211-214.

- Radhakrishna, R. B. & Bruening, T. H., (1994). "Pennsylvania study: employee and student perceptions of skills and experiences needed for careers in agribusiness." *NACTA Journal*, 38: 15-18.
- Rockwell, S.K. & Kohn, H. (1989). Post-then pre evaluation. *Journal of Extension* 27 (2):19-21.
- Rohs, F. R. & Langone, C. A. (1997). Increased accuracy in measuring leadership impacts. *Journal of Leadership Studies* 4 (1): 150-159.
- Rohs, F. R., Langone, C. A. & Coleman, R. (2001). Response shift bias: A problem in evaluating school nutrition training using self report measures. *Journal of Nutrition Education* 33 (3): 1-14.
- Russell, M. A. & Jones, H. W. (1995). A leadership development course for animal industry careers. *NACTA Journal* 39 (4): 30-33.
- Sprangers, M. & Hoogstraten, J. (1988). Response-style effects, response-shift bias, and a bogus pipeline: a replication. *Psychological Reports* 62: 11-16.
- Texas A & M. 1998. "Results of student questionnaire." Presented to the Southern Region Ag Teaching Symposium, College Station, Texas, October 25-27.
- Welch, W. W. & Walberg, H. J. (1970). Pretest and sensitization effecting is curriculum evaluation. *American Educational Research Journal* 7: 605-614.
- Willson, V. L. & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*. 19 (2): 249-258.