Journal of Leisure Research
2007, Vol. 39, No. 2, pp. 295-315

Addressing Response-shift Bias: Retrospective Pretests in Recreation Research and Evaluation

Jim Sibthorp, Ph.D.
Department of Parks, Recreation, and Tourism
University of Utah

Karen Paisley, Ph.D.
Department of Parks, Recreation, and Tourism
University of Utah

John Gookin, BA
The National Outdoor Leadership School

Peter Ward, MBA
Department of Recreation Management and Youth Leadership
Brigham Young University

The self-reported pretest/posttest has been commonly used to assess change in recreation research and evaluation efforts. The viability of comparing pre and post measures relies on the assumption that the scale of measurement, or metric, is the same before and after an intervention. With self-report measures, the metric resides within the study participants and, thus, can be directly affected by the intervention. If participants' levels of self-knowledge change as the result of a recreation program, then this metric may also shift, making comparisons between measures from before and after the program problematic. This article aims to both synthesize the theory and literature surrounding this problem and to offer a mixed-methods, data-based example, which illustrates the problem in a recreation context and posits possible reasons for differences in reported pre-course attribute levels by reporting time. Results generally supported using a retrospective pretest as a way to address changing metrics with self-report measures. This article further discusses when and how it is appropriate to use retrospective pretests in recreation research and evaluation.

KEYWORDS: *Response-shift bias, program evaluation, outcome measurement.*

Introduction

Few topics are more central to professionals in parks and recreation, outdoor education, or organized camping than the efficacy of their programs in bringing about positive change and facilitating development of service recipients. Methodological issues, however, have consistently compromised researchers' efforts to assess the impact of recreation programs on participants (e.g., Christensen, 1995; Witt, 2000). Despite its inherent limitations,

Address correspondence to: Jim Sibthorp, University of Utah, Department of Parks, Recreation, and Tourism, 250 South 1850, Room 200, Salt Lake City, UT 84112-0920; e-mail: jim.sibthorp@health.utah.edu.

Author note: The authors wish to thank Dr. Gary Ellis and three anonymous reviewers for constructive comments on earlier versions of this manuscript.

one of the more common ways to evaluate a program's impact is to use a pretest/posttest design. In leisure research and evaluation, this approach commonly involves administering a self-report pretest of participants' status on variables that are targeted by the program (e.g., resiliency, self-esteem, self-efficacy, identity development), administering the program, and then having the participants complete a posttest of the same self-report measures to determine if change has occurred (e.g., Caldwell & Baldwin, 2004; Green, Kleiber, & Tarrant, 2000; Hurtes, Allen, Stevens, & Lee, 2000; Searle, Mahon, Iso-Ahola, Sodrolias, & Van Dyck, 1995). This very intuitively compelling approach, however, is fraught with important methodological issues (cf. Cronbach & Furby, 1970; Rogosa, Brandt, & Zimowski, 1982). At the heart of these issues is the challenge of accomplishing measurement of change that yields the potential to make appropriate inferences about the extent of participants' growth as a result of program participation.

The viability of comparing pre and post measures of any type relies on the assumption that the scale of measurement, or metric, is the same before and after the treatment, intervention, or program. With objective measures (e.g., cognitive tests) or behavioral measures (e.g., the use of observers/raters), the metric is not directly affected by the program as it lies *outside* of the program participants. However, with self-report measures, the metric resides *within* the study participants and, thus, can be directly affected by the intervention. If participants' levels of self-knowledge and awareness change as the result of the recreation program, then this metric, or internalized standard of measurement, may also shift, making comparisons between measures from before the program and those after the program problematic. Consider, for example, a pretest/posttest scenario, which entails the administration of identical questionnaires before and after a program for the purpose of assessing change. At pretest, a participant circles a number at the midpoint of a seven-point scale, in response to an item reading, "I am competent in my wilderness navigational skills." The number indicated by the participant corresponds to the descriptor "about average." Imagine also that the program proves to be incredibly enlightening and, as a result, the participant comes to understand that "average" wilderness navigation skills are much greater than she or he imagined at the pre-test occasion of measurement. At the end of the course (posttest), "about average" would carry a very different meaning, but the participant could potentially choose the same response in the context of that new meaning. The pretest/posttest change score (posttest score less pretest score) on that item would then be zero, despite the fact that substantial change may have occurred. The change is not accurately measured because "about average" carries such radically different meanings on posttest versus pretest occasions. The failure to detect change is a result of recalibration of the instrument; the metric has changed for this (and other) participants. This recalibration of participants' internal metric from the beginning to the end of the program is commonly referred to as a "response-shift bias" in measurement. Thus, the primary aims of this article are threefold: 1) to synthesize the literature and theory regarding response-shift bias as it is most applicable to recreation research and evalu-

ation, 2) to determine if a response-shift bias may be present in some recreation related outcomes through a mixed-methods, data-based example, and 3) to offer possible insight into both why this bias might occur and how it might be addressed.

Response-shift Bias and the Retrospective Pretest

The idea of response-shift bias can largely be traced to the work of Howard and his colleagues (e.g., Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber; 1979; Howard, Schmeck & Bray, 1979). Through a series of studies, they consistently found that, under certain circumstances, the program or intervention directly affected the self-report metric between the pre-program administration of the instrument and the post-program administration. These recalibrations of the underlying metric often dramatically understated the effect size attributable to the program and robbed the analysis of the research or evaluation data of statistical power. While the specific impact will vary with the degree of response-shift bias present, through a series of analytic and Monte Carlo techniques, Bray, Maxwell, and Howard (1984) found that the presence of response-shift bias resulted in the substantial loss of statistical power; in some cases as much as 90 percent. The predominant solution to this problem has been to suggest that, in certain situations, researchers use a retrospective pretest, or "then-test", to more accurately gauge the pre-program level of the attribute of interest.

At the conclusion of a program or intervention, a retrospective pretest essentially asks the questionnaire respondent to reflect back to a previous time (usually pre-program) and indicate his or her current perception of the level of an attribute he or she possessed at that previous time. This approach holds that, after a program, participants will be better able to define and understand the construct being measured and will be applying the same metric as they assess both their pre- and post-program levels of the attribute. As the retrospective pretest and the posttest are completed at the same time and on the same metric, any response-shift bias caused by a changing metric will be absent from the data. For example, Toupence and Townsend (2000) used the retrospective pretest approach to investigate whether perceived leadership skills are gained though organized camping. After participating in a week-long camp, they asked campers, camp counselors, and counselors-in-training to assess their perceived leadership skills *before* camp on a retrospective pretest version of the Leadership Skills Inventory (LSI), indicating their pre-camp levels of leadership. After completing this questionnaire, the study participants were then asked to complete a traditional posttest version of the LSI that indicated their current perceived levels of leadership. Comparing the retrospective pretest score to the posttest score then assessed change.

The retrospective pretest has been successfully used to address the issue of response-shift bias for a variety of outcomes relevant to recreation program research and evaluation, including leadership (Goedhart & Hoogstra-

ten, 1992; Rohs, 1999; Rohs & Langone, 1997; Toupence & Townsend, 2000); quality of life in cancer patients (Breetvelt & van Dan, 1991); management training (Mann, 1997; Mezoff, 1981); reduced substance abuse (Rhodes & Jason, 1987); and attitudes towards persons with disabilities (Lee, Paterson, & Chan, 1994). Research has also found evidence that retrospective pretests provided higher correlations with objective and performance measures than did pre-program pretests (Hoogstraton, 1985; Howard & Dailey, 1979; Pohl, 1982; Pratt, McGuigan, & Katzer, 2000). In addition, interviews and qualitative approaches have consistently verified that respondents' lack of pre-program self-knowledge and understanding was the reason behind the response-shift bias in a variety of contexts (Cantrell, 2003; Manthei, 1997; Mezoff, 1981).

The response-shift bias is most pronounced when it is likely that the program can change the underlying metric for the participants (Howard et al., 1979; Howard, 1980). In some cases, this may even be the intent of the program; for example, to help corporate managers in leadership training to redefine leadership. However, there is little evidence of response-shift bias on either well-known topics or for populations with specific expertise in the topical area being assessed (Sprangers & Hoogstraten, 1988b). For example, if participants' definitions of the variables of interest are well established and stable, the metric will not be changed by the program. Thus, when participants have a solid grasp of the concepts being investigated, or if the intervention is unable to change the underlying metric/self-knowledge, then the use of a retrospective pretest is not only unnecessary, but also undesirable, as the limitations of this approach (which are discussed later) are realized without the benefits (Koele & Hoogstraten, 1988; Townsend & Wilton, 2003).

While the retrospective pretest is the most widely-used technique to address a response-shift bias, several alternatives have been studied with varying degrees of success. Howard, Dailey, and Gulanick (1979) tried to better stabilize subjects' internal metric before a program intervention by better defining the construct of interest through an informational pretest. This approach essentially uses a pretest to better define the attributes of interest, thus giving the respondent a better idea of what to expect and how to self-assess. However, they found this approach to be ineffective as response-shift biases were still present in the data, and Howard (1980) questions the effectiveness of short-term efforts to change a metric that will likely evolve over the course of the training or education program. Sprangers and Hoogstraten (1989) did find that a behavioral pretest, one that requires actual performance by the study participants, can reduce response-shift bias in traditional pretest/posttest assessments. They speculate that this approach is effective as it vividly demonstrates to the participants the inadequacy of their existing internal metrics. In addition, Spangers and Hoogstraton (1987; 1988b) employed what they termed a "bogus pipeline" where they indicated to the study participants that their actual levels of the attribute would be assessed with an objective measure and compared with their self-reports in an effort to verify the accuracy of their self-perceptions. This approach has shown

some promise in cases where the idea of an objective measure appeared credible to the program participants. One viable recommendation remains the use of behavioral or more objective measures to address the response-shift bias issue, essentially isolating the metric from the intervention.

However, alternatives to addressing response-shift bias via a retrospective pretest are often inappropriate, impractical, and plagued with their own weaknesses. It is not always feasible to have participants complete a lengthy behavioral pretest where they might, for example, be asked to assume a leadership position in a group of their peers. Likewise, more objective measures, for example physiological responses, are often costly, intrusive, and logistically challenging to utilize. Thus, self-reports will likely remain common in recreation research and evaluation efforts, and retrospective pretests may enhance the utility of these in some situations.

Potential Issues with a Retrospective Pretest Approach

While the retrospective pretest can potentially mitigate the response-shift bias problem due to a lack of pre-program knowledge, it remains subject to other traditional limitations of self-report measures and to some additional limitations. One of the most common concerns with all self-report measures remains the truthfulness of the participants' responses. While this bias is possible in all self-reports, the proximity of the pre- and post-program scores makes the retrospective design especially susceptible to biased reporting. Participants either seeking to show increases or decreases in the training content can easily do this by artificially increasing or decreasing either the retrospective pretest or posttest scores. While an artificial increase might be offered in an effort to please the investigator or program (i.e., acquiescence) or to justify the effort expended in the program, a false negative result might be presented in the hope of gaining access to additional training or treatment (e.g., Howard & Dailey, 1979; Spangers & Hoogstraten, 1988). This issue can sometimes be addressed through the inclusion of a "lie scale" or "social desirability scale" which seeks to discriminate those who are responding candidly from those who are not. This technique has been widely employed and advocated in social scientific research where respondents may have a compelling reason to misrepresent their levels of an attribute (Hopkins, 1986).

Retrospective pretests face some additional limitations beyond those of a traditional pretest/posttest. One of the most notable of these is the issue of memory: Can participants accurately gauge their levels of an attribute at a moment days or even weeks in the past? Pearson, Ross, and Dawes (1991) offer several challenges to questioning approaches that rely substantially on memory recall. Basically, they posit that recall can be influenced by personal theories of either "stability" or "change." When a recall type question is asked, the most common cognitive process involves determining the current status of the attribute (for example, today's attitude toward the environment) and then deciding if the past (recalled) status of this attribute was the same

(stability) or different (change). Pearson and colleagues believe that these inherent theories can lead to either under or over estimation of change. Despite these limitations, Pearson et al. do acknowledge that changing standards of measurement (metrics) are especially problematic to measuring some types of variables via traditional pretest/posttest approaches, as respondents "often fail to make adjustment to initial standards" (p. 84).

In examining study participants' memories of their pretest ratings (i.e., the score they believe they reported at pretest), their actual pretest ratings (i.e., the score they actually reported at pretest), and their retrospective pretest ratings (i.e., the score they believe accurately represents their pretest level after the program), Howard et al. (1979) found that the *remembered* pretest ratings were closely related to the *actual* pretest ratings, but that the *retrospective* pretest ratings remained the lowest. They posit that this indicates that the response-shift bias present is more than simply a systematic memory distortion, but rather is related to the program's impact on the participant's internal metric. Pratt et al. (2000) posit that memory is a greater problem when the questions are more general in focus. That is, specific memories are more easily and accurately recalled. However, while memory is a clear limitation, it does appear that retrospective pretests offer one viable and reasonable approach to addressing response-shift bias.

While the use of a retrospective pretest does not offer a viable solution for those seeking measures of actual knowledge, skills, or behavior, it does offer an alternative approach for measuring self-perceptions such as affective states, attitudes, or behavioral intentions. Such outcomes are common in recreation research and evaluation efforts. Therefore, an effort was made to determine the role that response-shift bias may play in recreation program research and evaluation. The purpose of the data-based portion of this paper is to compare the traditional pretest/posttest format to a retrospective pretest/posttest format, and, if differences in these formats exist to posit a possible reason for the change in pre-program (a course) levels by reporting time (before the course or ~30 days later). This purpose was investigated through a mixed methods approach, which combined quantitative and qualitative analyses.

Methods

To conduct this study, two complementary approaches were employed. In the quantitative approach, participants in five National Outdoor Leadership School (NOLS) courses were asked to complete both pretest and retrospective versions of the NOLS Outcome Instrument (NOI) as well as a posttest version for the calculation of change scores (a pretest, retrospective pretest, posttest repeated measures design). This design allowed for the control of within-subject variation while offering the potential to detect response shift biases in the data. A qualitative component was also included. Interviews were conducted with students on two of the five courses that participated in the quantitative portion of the study in efforts to ascertain why differences in the reported pre-course levels of attributes were reported.

Participants and Settings

Participants in the quantitative portion of the study were enrolled in five NOLS courses in the fall of 2004 and the winter of 2005. Two of these courses were NOLS semester courses (Fall Semester in the Rockies), while the other three courses were sailing or kayaking courses offered in Mexico (Baja Coastal Sailing and Baja Sea Kayaking). For the qualitative portion of the study, a convenience sample of the two Rocky Mountain semester courses was used, as these courses were geographically available to the two lead researchers.

Established in 1965, NOLS strives to be the leader in wilderness education by combining the development of leadership and technical outdoor skills with education regarding biology and natural history in their naturally occurring environments. Courses are tailored to various specific populations including youth, college-age students, individuals 25 years of age and older, individuals either currently working as or seeking to become outdoor educators, and individuals seeking to become NOLS instructors. Course offerings range from eight days to a full academic semester in length, and students can elect to earn college credit at the undergraduate or graduate level for their studies with NOLS. These courses all target the six generic NOLS learning objectives as well as course- and location-specific objectives.

Measures

In order to measure the targeted NOLS outcomes, the objectives needed to be operationalized and converted to a measurement instrument. The NOLS Outcome Instrument (NOI) was developed over several years using both classical test theory and congeneric measurement theory to assess the viability of the proposed instrument.[1] All questions were positively worded and were scored on a 10-point Likert-type scale ranging from 0 (not like me) to 9 (like me). The version of the instrument used in this study has 29 items and is thought to measure six distinct constructs: Communication, Leadership, Group Behavior, Judgment in the Outdoors, Outdoor Skills, and Environmental Awareness. The final version of this instrument was given to 517 NOLS students in 2004, and the following statistics on the instrument were calculated from this sample (see Sibthorp, Paisley, Gookin, & Ward, 2005). Communication was measured by a four-item subscale (e.g., "I express my ideas clearly"). Leadership was measured by a five-item subscale (e.g., "I take initiative in completing group tasks"). A five-item subscale measures Group Behavior (e.g., "I am patient with others"). Judgment in the Outdoors was measured by a four-item subscale (e.g., "I can identify potentially dangerous areas in wilderness settings"). Outdoor Skills were measured by a five-item subscale (e.g., "I am competent in my wilderness navigational skills"). Envi-

[1] A congeneric measurement model, in this case, was employed using confirmatory factor analysis and allows less restrictive assumptions than classical test theory while allowing for individual items to be differentially weighted as a function of the latent variables.

ronmental Awareness was measured by a four-item subscale (e.g., "I understand the purpose of Leave No Trace with respect to wilderness travel"). Internal consistencies (Cronbach's alphas) were acceptable for all subscales and ranged from a low of .76 to a high of .86. A "lie scale" is imbedded in the NOI in an effort to detect artificially elevated change scores and remove them from subsequent analyses. The lie scale consists of two items that address outdoor skills that are mutually exclusive on most NOLS courses: assessing avalanche slope stability and predicting tides and currents. All subscale scores were calculated by summing the associated items and dividing the score by the number of items in the subscale, as is consistent with classical test theory.

To assess the NOI's ability to measure six distinct constructs, a correlation matrix was calculated for the six subscales (n = 517). The between subscale correlations, with one exception, were moderate to low and fell below the a-priori cut-off of .70. The correlation between outdoor skills and judgment was higher than desired ($r = .77$, $p < .05$) and does not support the premise that these two subscales are measuring distinct constructs. However, as this study is primarily concerned with the response-shift biases present in the data, this notable correlation should not affect the study findings.

Procedures

After arriving at the program location and prior to beginning their course, participants in these five NOLS courses were invited to participate in the study. All 57 enrolled students agreed to participate and then completed a pretest version of the NOI before departure. Immediately after completing their courses, or the first section of their semester courses (the first section is approximately 30 days long), the participants completed a second version of the NOI formatted in a post-retrospective pretest format. They first indicated their perceived levels of each attribute at the current time (posttest) and then, subsequently, before their courses began (retrospective pretest). This retrospective approach is advocated by Lam and Bengo (2003). Thus, for each participant, three scores on the NOI were collected: pretest, posttest, and retrospective pretest.

As a convenience sample, all of the 29 participants in the two semester courses in the Rocky Mountains were invited to participate in the qualitative portion of the study. All of the students on these two courses agreed to be interviewed in order to obtain a more complete picture of why scores may vary from pretest to retrospective-pretest. Immediately after these participants completed their retrospective pretests and posttests, their questionnaires were matched to their pretests by birth date and their initial pretest scores were recorded on the retrospective form so that the two perceptions of entry-level knowledge were readily available for comparison. In one-on-one interviews with one of the two lead researchers, participants were asked whether they remembered their pretest scores from the beginning of the course. Then, all participants were asked to explain the discrepancies be-

tween their pretest and retrospective pretest scores for attributes exhibiting more than a 20% (2 points on a 10-point scale) change in magnitude (either plus or minus). Attributes with smaller changes in magnitude were not investigated with the participants as they seemed too small to be distinct. These questions stem from the work of Pearson et al. (1991), described previously, regarding stability and change.

Data Analysis

After initial data entry, cleaning, and screening, the five groups were compared to see if differences between them would limit aggregating the data for subsequent analysis. The lie items were inspected and compared to the a priori criterion (a gain greater than 2 points on a 10 point scale warranted subject deletion from the analysis). Then, a repeated measures multivariate analysis of variance (MANOVA) technique was used to investigate the hypothesized relationships between the pretest, the retrospective pretest, and the posttest for the six dependent variables. Significant MANOVA results were to be further explored through simple a-priori contrasts which would compare both the pretest and the retrospective pretest to the posttest. All analyses significant at $p < .05$ were interpreted.

The qualitative data from the interviews were analyzed through enumeration, for the question of whether participants remembered their pretest scores, and through constant comparison, for the discrepancy data. For the latter, two researchers coded the data thematically and common codes were interpreted.

Results

For the quantitative dimension of the study, a total of fifty-seven participants on five different NOLS courses completed pretests, retrospective pretests, and posttests of the NOI during the fall 2004 and the winter of 2005. The participants ranged in age from 16 to 46 (M = 21.0) and 45% were female. Fifty-five percent reported having done something similar to a NOLS expedition before participating in their current course. Using SPSS 13.0, descriptives and frequencies were inspected for univariate outliers, normality, and illegal scores. Missing data were inspected and, when it was viable, the missing value was replaced with the mean of the other items measuring the same construct (e.g., the missing value for a 5 item measure was replaced by the mean of the remaining 4 items designed to measure the same construct). None of the dependent variables were missing more than two cases (3.6%). To determine if the groups were different, a MANOVA was run with the pretest scores as the dependent variables and the course number as the independent variable. Differences were statistically significant, thus the groups were not "equivalent" on reported pre-course levels of the attributes. While the researchers were primarily interested in the evidence of response-shift bias and the qualitative follow-up data, this must be noted as a limitation of

the study. The lie scale, which was included to detect artificially elevated program gains, did not exhibit sufficient changes to warrant deletion of any subjects in this sample. One participant was removed from the analysis as his pretest scores exhibited an unlikely response pattern (all were at the extreme level at the pretest, but returned to more believable levels during the post and retrospective data collections). Removal of this participant left a usable sample size of 56 for the remainder of this study. However, given the inclusion of six dependent variables measures on three NOI versions, listwise deletion further reduced the sample entered into the MANOVA analysis to 50 participants.

The GLM technique in SPSS 13.0 was used for the doubly multivariate analysis of variance.[2] In addition to assessing normality through the descriptives, Box's M test was not significant, supporting the underlying assumption of homogeneity of the variance-covariance matrix necessary for the use of MANOVA. With the use of Wilk's criterion, the combined dependent variables were found to differ significantly by test type ($\Lambda = .16$, $F = 16.33$, $p < .05$). The marginal means for the MANOVA analysis are reported in Figure 1 and Table 1.

Following the significant multivariate tests, simple contrasts were used to determine if significant differences were present between the posttest and the two versions of the pretest (true pretest and retrospective pretest). While all of these contrasts are statistically significant (see Table 2), the greater statistical power available from the analysis of the retrospective pretest is readily evident from the universally higher values for partial η^2. The increases in partial η^2 values for the six dependent variables ranged from a high of +.532 for Communication to a low of +.075 for Environmental Awareness.

Analysis of Difference Scores

While the main intent of this study was to test the mean difference in scores, it was possible that, if scores moved both up and down on a given variable depending upon the participant, the true effect of the response-shift bias could be masked. Therefore, analyses were carried out to examine the bivariate correlations between the difference scores calculated by subtracting the pretest from the posttest and those calculated by subtracting the retrospective pretest from the posttest. These ranged from a low of .30 for Leadership to a high of .72 for Outdoor Skills. The lowest correlations supported the premise that the Communication, Leadership, and Group Behavior skills were least consistent when pretest and retrospective pretest scores were used and compared, while the Judgment in the Outdoors, Outdoor Skills, and Environmental Awareness change scores were the most consistent, regardless of type of pretest score used (change scores calculated using the pretest score versus those calculated with the retrospective pretest score.) This finding

[2]A doubly multivariate analysis of variance refers to multiple dependent variables (i.e., the six outcomes in this study) measured over multiple times (i.e., the three versions of the instrument).

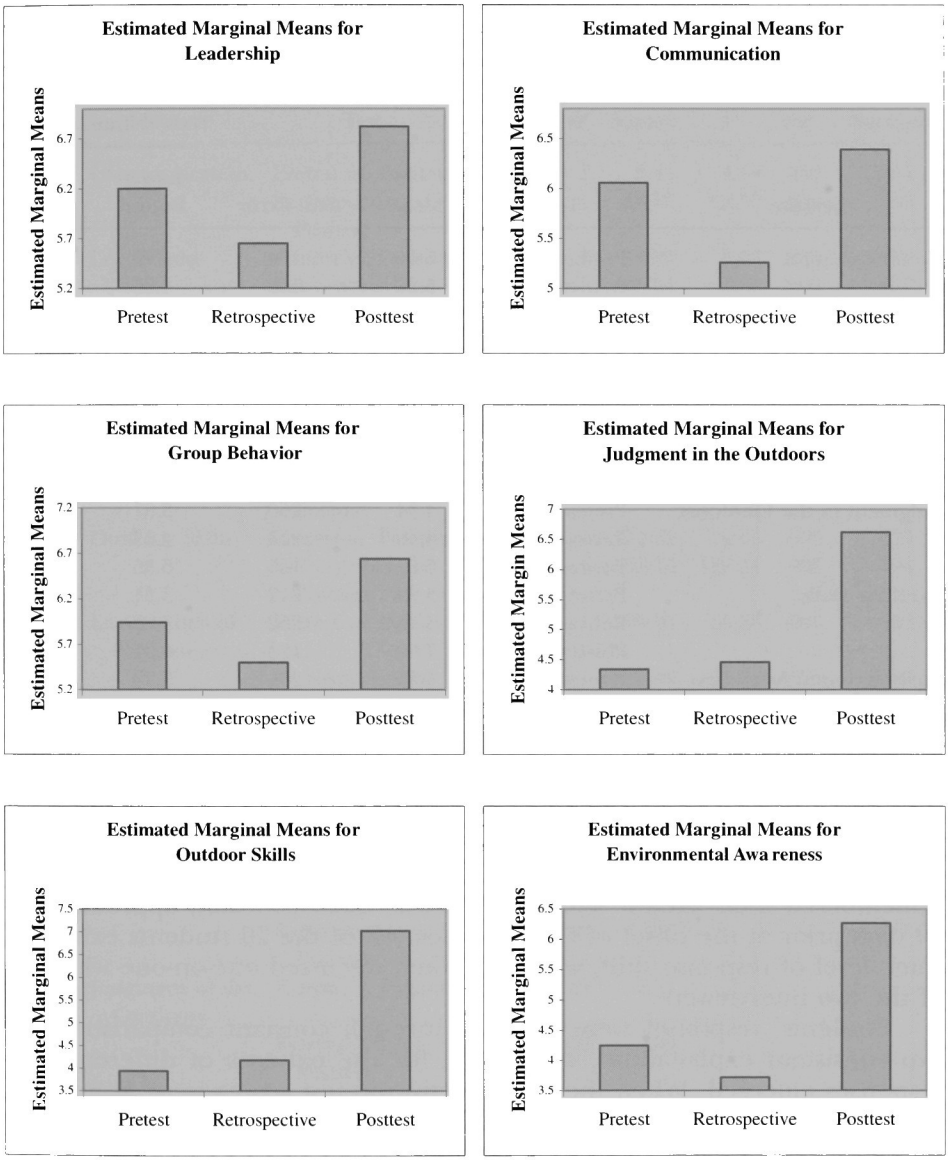

Figure 1. Marginal means for the six dependent variables by test.

largely supports the conclusions from both the inspection of means (see Table 1) and the examination of the partial η^2 values (see Table 2).

Qualitative Analysis

All 29 participants in the two semester-long courses were interviewed after completing their retrospective pretests in an effort to determine the

TABLE 1
Descriptive Statistics for the Six Dependent Variables at the Three Levels of Test

Measure	Test	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Communication	Pretest	6.06	.164	5.73	6.39
	Retrospective	5.26	.198	4.86	5.66
	Posttest	6.39	.171	6.05	6.73
Leadership	Pretest	6.20	.194	5.81	6.59
	Retrospective	5.65	.179	5.29	6.01
	Posttest	6.82	.149	6.52	7.12
Group Behavior	Pretest	5.94	.183	5.58	6.31
	Retrospective	5.50	.182	5.14	5.87
	Posttest	6.64	.189	6.26	7.02
Judgment in the Outdoors	Pretest	4.34	.260	3.81	4.86
	Retrospective	4.46	.214	4.03	4.89
	Posttest	6.62	.135	6.35	6.89
Outdoor Skills	Pretest	3.94	.312	3.31	4.57
	Retrospective	4.18	.252	3.68	4.69
	Posttest	7.02	.123	6.77	7.27
Environmental Awareness	Pretest	4.25	.234	3.78	4.72
	Retrospective	3.72	.260	3.19	4.24
	Posttest	6.27	.181	5.91	6.63

reasons for any variation between the pretest and retrospective responses. Enumeration of students' responses indicated that none of the 29 students remembered their pretest scores, which had been collected approximately 30 days prior at the onset of their courses. All of the 29 students exhibited some level of response shift, which was then discussed one-on-one with one of the two interviewers.

Students' responses were analyzed through constant comparison, and two consistent explanations, or themes, for the patterns of differences in responses emerged. When the retrospective pretest scores were *higher* than the initial pretest scores, it seemed to be due to a sense of underestimation of ability. For example, regarding the item, "I can identify potentially dangerous areas in wilderness settings" one student "thought there would be more to it" but then recognized "it was pretty common sense." Her retrospective pretest score, then, was higher than her initial pretest score as she reevaluated her skill level. This propensity to underestimate ability may be exacerbated by the instructors, as well. One student commented that an instructor had warned the group that they would be camping in some "serious stuff," which led the student to doubt his abilities—even though he had taken a difficult NOLS course previously. Being "out there," however, "reminded [him] about what [he] knew."

TABLE 2
Simple Contrasts for Six Dependent Variables across the Three Levels of Test

Source	Measure	Test	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test	Communication	Pretest vs. Posttest	5.45	1	5.45	4.50	.039	.084
		Retrospective vs. Posttest	63.85	1	63.85	93.30	.000	.656
	Leadership	Pretest vs. Posttest	19.22	1	19.22	8.68	.005	.150
		Retrospective vs. Posttest	68.21	1	68.21	86.08	.000	.637
	Group Behavior	Pretest vs. Posttest	24.50	1	24.50	10.06	.003	.170
		Retrospective vs. Posttest	64.98	1	64.98	83.57	.000	.630
	Judgment in the Outdoors	Pretest vs. Posttest	261.06	1	261.06	70.43	.000	.590
		Retrospective vs. Posttest	234.36	1	234.36	119.68	.000	.710
	Outdoor Skills	Pretest vs. Posttest	474.32	1	474.32	79.62	.000	.619
		Retrospective vs. Posttest	402.15	1	402.15	128.31	.000	.724
	Environmental Awareness	Pretest vs. Posttest	205.03	1	205.03	55.80	.000	.532
		Retrospective vs. Posttest	326.40	1	326.40	75.56	.000	.607
Error	Communication	Pretest vs. Posttest	59.31	49	1.21			
		Retrospective vs. Posttest	33.53	49	.68			
	Leadership	Pretest vs. Posttest	108.50	49	2.21			
		Retrospective vs. Posttest	38.83	49	.79			
	Group Behavior	Pretest vs. Posttest	119.30	49	2.44			
		Retrospective vs. Posttest	38.10	49	.78			
	Judgment in the Outdoors	Pretest vs. Posttest	181.63	49	3.71			
		Retrospective vs. Posttest	95.95	49	1.96			
	Outdoor Skills	Pretest vs. Posttest	291.92	49	5.96			
		Retrospective vs. Posttest	153.58	49	3.13			
	Environmental Awareness	Pretest vs. Posttest	180.03	49	3.67			
		Retrospective vs. Posttest	211.66	49	4.32			

When retrospective pretest scores were *lower* than the initial pretest scores, it was due to an overestimation of ability—students "didn't know what they didn't know." According to one student, she "thought [she] knew it but found out [she] had some things to learn." Another student said that he did not know, before the course, "how stupid some of the things I do are." One of the most notable examples of this pattern occurred on the item, "I am patient with others." Most students overestimated their skills in this area because they had not spent substantial time living in a small group and, as such, had not had their skills "tested" before. However, due to interpersonal challenges on the trip, many students lowered their scores on this item in retrospect.

Regardless of which score was higher, students explained that the differences were due to a lack of initial understanding of what the item actually entailed. Further, when asked if they could remember their initial pretest scores, participants, universally, could not. Thus, it appears that, despite some inconsistencies in the direction of the shift, students made more informed responses *after* the experience. One student said, "It's like a whole new set of numbers now. I need a whole new approach."

Discussion

The purpose of the databased portion of this study was to compare the traditional pretest/posttest format to a retrospective pretest/posttest format and, if differences existed between these formats, to further examine why differences may occur. While this study focused on participants on courses run by the National Outdoor Leadership School, the nature of the outcome variables in question are considered important to recreation program research and evaluation in general. Both the quantitative and qualitative findings from this study largely support the existence of a response-shift bias. For four of the outcome variables, the means were significantly different from the pretest to the retrospective pretest, and the effect sizes as measured by the partial η^2 were universally larger, yielding greater statistical power. Based on data from the qualitative interviews, the reason for this response bias largely supported previous studies: Participants became more fully aware of the variables as the program progressed (Cantrell, 2003; Manthei, 1997; Mezoff, 1981).

The use of a retrospective pretest as a way to address this bias was also supported. Four of the six mean scores for the dependent variables showed a significant downward shift from pretest to retrospective pretest as the participants, presumably, recalibrated their internal metrics as they became more informed over the duration of the course. However, the qualitative data seemed to indicate that the pattern of metric movement (up vs. down) remained somewhat individual in nature and depended on whether the participants had viewed the "skill set" as either easier or more difficult than they actually found it to be. In general, however, participants seemed better prepared to respond realistically to the items once they had experienced the constructs being measured.

While a response-shift bias was not present in all the variables, these data provide no compelling support for the use of a pretest over the use of a retrospective pretest. In contrast, as four of the six variables did exhibit the hypothesized response-shift bias, there seems to be a compelling design reason to utilize a retrospective pretest in recreation programs targeting outcomes similar in nature to Communication, Leadership, Group Behavior, and Environmental Awareness skill sets. The socially-oriented nature of most recreation programs may provide substantial opportunities for participants to reevaluate their skill levels in socially oriented attributes (e.g., communication, leadership, or group behavior). For outcome measures that may be more concrete in nature, however, participants may be more able to grasp the meaning of the items and adhere to a more consistent internal metric.

Limitations

While the outcomes of this study are relevant to recreation programs in general, data were collected from a small group of participants on five courses offered by a very specific program: the National Outdoor Leadership School. The intensive nature of small group living may have exacerbated the extent of the reflection on socially-oriented outcomes which, in turn, could have accentuated the magnitude of the shift in these metrics. As such, the data may not be representative of shorter format recreation programs or recreation programs targeting fundamentally different outcomes.

The quantitative portion of this study is vulnerable to the previously mentioned limitations of using a retrospective pretest. These include the challenges of being able to accurately recall pre-program levels of an attribute and ease of faking positive growth or change. In addition, the measures for this portion were self-reports and are subject to all the traditional limitations associated with using self-report measures.

The groups aggregated together for this study came in with self-reported differentials in the targeted outcomes, had different group dynamics, different experiences, and different instructors. While this is not centrally related to the aim and purpose of this study, it is notable in that relative importance of response-shift bias remains contextual, and is likely to be more dramatic in some contexts and courses than in others. Thus, it must be noted as a limitation of the empirical portions of this manuscript.

Lack of a control group may also be considered a limitation. While it seems unlikely that metric reconceptualization of targeted course outcomes would occur in a control group, this remains a possibility. Thus, changing metrics could simply be a function of time rather than of program intervention. However, it is notable that there was no support for this alternate hypothesis in the qualitative data, as the participants attributed their changing metrics to their course experiences.

Another possible limitation is the disconnect between how the quantitative and qualitative data were analyzed. The qualitative data were collected and examined at the item level; the interviewers explored discrepancies in how the participants responded to individual items. In contrast, the quanti-

tative data were examined as composite scores (summed items) thought to represent constructs. While it was logistically easier to examine item-level differences during the qualitative portion, this approach is not entirely consistent with how researchers use summative scales (such as the ones in this study), where the items are thought to represent the presence (or absence) of an unobservable construct (e.g., communication skills). While response-shift bias has been examined at both the item (e.g., Rohs, 1999) and variable levels (e.g., Howard et al., 1979) in the past, this disconnect is a limitation of this study.

Implications for Practice

The retrospective approach, as a self-report measure of affect and attitudes, while supported in this case, should not preclude other approaches to data collection depending on the nature of the target outcome variables. This is especially true when the outcome variables have a stable metric. Qualitative designs still offer tremendous insight into program effectiveness, and operate with a different set of concerns than those addressed here. Similarly, behavior anchored rating scales, observational rating, and other designs that allow the metric of measurement to exist outside of the program participants, are still viable options depending on nature of the targeted outcome variables. In addition, growth curve modeling offers another possible alternative to the assessment of change (cf. Raudenbush & Bryk, 2002; Rogosa et al., 1982). For example, Bialeschki, Sibthorp, and Ellis (2006) used this approach to examine anger reduction in campers over a four-year longitudinal study.

Despite these alternatives, many and most of the common research designs can be negatively impacted by a response-shift bias if they employ self-reports. For example, consider how this bias might impact findings from a Solomon four-group design: group one gets a pretest, a posttest, and the treatment; group two gets a treatment and a posttest only; group three gets a pretest and a posttest, but not treatment; and the final group gets only a posttest. If the treatment intervention truly leads to a changing internal metric of measurement, then groups one and two should both be evaluating their posttest levels on this "new" metric. Groups three and four would be evaluating their levels on the less stable, but unchanged, metric that has not been exposed to the program. While the Solomon four-group is often considered one of the most robust study designs, it is clear from this example that a) change from pretest to posttest attributable to program could be masked by a changing metric, and b) that comparisons between posttest scores for participants exposed to the program (groups 1 and 2) and posttest scores for participants that were not exposed (groups 3 and 4) could be without merit, as the two groups could be using different metrics (one reconceptualized through the program and the other not).

While acknowledging the merits of other approaches, sometimes alternatives to self-reports are simply not viable; they are too difficult, require specialized training, or are too expensive to administer. Under these circumstances, using a retrospective pretest does offer some practical benefits. First there is the comparative ease of administration. Rather than administering a questionnaire twice and matching responses, participants take the questionnaire only once, which substantially reduces administrative effort. This might even be required in some evaluation settings where programs, sensitive to taxing participants upon arrival, mandate that pretest not be used as they adversely impact the participants' experiences. It is important, however, that the directions provided to participants are clear, as the potential for confusion with a retrospective pretest format may be increased.

A second potential advantage of the retrospective pretest approach is that the retrospective approach cannot, itself, become part of the intervention, compared to a pretest, which could frontload the expectations of the program. For example, a pretest measuring environmental attitudes would likely provide program participants an explicit understanding of what the program hopes to accomplish. While this is not necessarily undesirable, it does leave one wondering if the same program impacts would be evident without such an understanding. This problem is more generically referred to as a pretest by treatment interaction and is considered a threat to external validity (cf. Campbell & Stanley, 1963).

Deciding when it is appropriate or inappropriate to use a retrospective pretest is not an easy task, as response-shift bias occurs because of a combination of program and participant factors. For example, if a program is targeting attributes that are not self-perceptions or ones that are not dependent upon an internal metric (e.g., caloric intake), then response-shift bias cannot occur. However, even if the program is targeting attributes that are potentially susceptible to response-shift bias, the program itself may not be able to change the internal metric. The metric could remain stable either because the program does not assist with the reconceptualization of the attribute, or because the participants' conceptualizations are stable (this is likely if they have expertise regarding the attribute of interest). Thus, there is no good general rule for when response-shift bias may be a problem, although it is only potentially a problem when measuring attributes whose definition is likely to shift over the duration of a program. Thus, response-shift bias will not be a problem when measuring variables that are not commonly operationalized as self-perceptions and where the metric can reside outside of the participant, such as knowledge, skill, or behavior.

In general, Koele and Hoogstraten (1988) advocate the use of both a pretest and a retrospective pretest as a way to assess the presence or absence of a response-shift bias. If the bias is present, then the retrospective score should be used to assess change. If the bias is absent, then the traditional pretest should be used. While this approach makes sense for more substantial research and evaluation efforts, it is likely that smaller evaluation projects

may well determine which approach to use based on the relative potential of response-shift bias in their targeted outcomes for their populations.[3]

Included in the potential caveats of using a retrospective pretest approach are the problems of satisficing and social desirability. Satisficing is the tendency to exert minimal effort in responding (Lam & Bengo, 2003); an individual will respond in a manner that requires the least amount of physical or cognitive exertion. For example, when a task or skill is difficult for an individual to perform, he/she will choose the response that is the easiest to complete. The same satisficing notion can be applied to a participant who completes a questionnaire in a socially desirable manner, where that participant bases his/her responses on what is expected or preferred by a particular "society" rather than exerting the effort to respond according to how he/she really feels. For example, participants are often asked about the effectiveness of their instructor. The easy, socially desirable response, which requires little cognitive effort, would be "highly effective" or toward that end of a scale. Responding honestly would require participants to evaluate the instructor's complete performance, expending substantial cognitive energy and effort.

The additional cognitive effort required to complete a retrospective pretest, which may be interpreted as a difficult task, may lead some respondents to engage in satisficing to reduce their amounts of effort. However, this limitation is not supported by the patterns of students' responses in these data: The different levels and directions of and reasons for change in students' scores suggests that they did actually evaluate the items individually and respond thoughtfully. Further, the fact that students' mean pre-program responses shifted both up and down suggests that social desirability was not a driving force behind their scores. If so, one would expect to see all the students' pre-program scores shift downward in efforts to please NOLS or their instructors with the volume they had learned (i.e., lower pre-course scores represent greater learning).

Conclusion

This paper contributes to the literature and to the profession in several ways. First, and most simply, it introduces a technique that is, at present, not widely used within our profession. This introduction is performed as objectively as possible, identifying the strengths and weaknesses of the approach, providing a specific example, and identifying circumstances under which such an approach might be appropriate.

Second, by combining quantitative and qualitative approaches, this study both identifies and suggests an explanation for a phenomenon—albeit among a specific sample. The quantitative results suggest that response-shift

[3]While this study specifically addressed the use of the retrospective pretest + posttest format, interested readers might see Lam and Bengo (2003) for an informative comparison of alternate formats including perceived change and post + perceived change.

bias exists among these participants on these outcomes, and the qualitative data provide insight as to *why* it is happening. The opportunity for a study to address both issues is somewhat unique.

Finally, our ability to provide meaningful and effective recreation programs for our participants depends upon our ability to measure the impacts of our efforts. As we are often dependent on self-report measures, which carry both strengths and weaknesses, we must seek to ensure that these measures are providing the most accurate information possible. To do this, and to advance the state of knowledge in the profession, we must be willing to critically examine our evaluation efforts. This paper represents steps toward such an examination. While it does not suggest the use of retrospective formats as a panacea, it does advocate for intentional selection of evaluation strategies and appropriate use of alternatives to familiar approaches. As we become better able to measure our impacts, we can move beyond the documentation of change to focusing our efforts on identifying the specific mechanisms responsible for that growth and change among participants. When we can isolate the mechanisms of change, we can fully capitalize on the potential of intentional programming.

References

Bialeschki, D., Sibthorp, J., & Ellis, G. (2006) Intentionality and targeted youth development outcomes: Four years at Morry's Camp. *2006 NRPA Leisure Research Symposium, Seattle, WA, 59.*

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

Breetvelt, I. S., & Van Dam, F. S. (1991). Underreporting by cancer patients: The case of response-shift. *Social Science Medicine, 32,* 981-987.

Caldwell, L. L., & Baldwin, C. K. (2004). Preliminary effects of a leisure education program to promote healthy use of free time among middle school adolescents. *Journal of Leisure Research, 36,* 310-335.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research* Chicago: Rand McNally.

Cantrell, P. (2003). Traditional vs. retrospective pretests for measuring science teaching efficacy beliefs in preservice teachers. *School Science and Mathematics, 103*(4), 177-185.

Christensen, J. (1995). Statistical and methodological issues in leisure research. In L. Barnett (Ed.), *Research about leisure: Past, present, and future* (*2nd ed.*) (pp. 231-251) Champaign, IL: Sagamore.

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

Goedhart, H., & Hoogstraten, J. (1992). The retrospective pretest and the role of pretest information in valuative studies. *Psychological Reports, 70,* 699-705.

Green, G., Kleiber, D., & Tarrant, M. (2000). The effect of an adventure-based recreation program on the development of resiliency in low income minority youth. *Journal of Park and Recreation Administration, 18*(3). 76-97.

Hoogstraten, J. (1985). Influence of objective measures on self-reports in a retrospective pretest-posttest design. *Journal of Experiment Education, 53*(4), 207-210.

Hopkins, K. D. (1986). Affective dependent measures: The use of a response integrity scale to enhance the validity of experimental and quasi-experimental research. *Journal of Special Education, 20,* 43-47.

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, 66*(2), 144-151.

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

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 evaluation and a re-evaluation 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.

Hurtes, K. P., Allen, L. R., Stevens, B. W., & Lee, C. (2000). Benefits-Based Programming: Making an impact on youth at-risk. *Journal of Park and Recreation Administration, 18*(1), 34-49.

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

Lam, T. & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation, 24*(1), 65-80

Lee, T. M., Paterson, J. G., & Chan, C. C. (1994). The effect of occupational therapy education on students' perceived attitudes towards persons with disabilities. *American Journal of Occupational Therapy, 48,* 633-638.

Mann, S. (1997). Implications of the response-shift bias for management. *Journal of Management Development, 16* 328-337.

Manthei, R. J. (1997). The response-shift bias in a counsellors education programme. *British Journal of Guidance & Counselling, 25*(2), 229-238.

Mezoff, B. (1981). Pre- then posttesting: a tool to improve the accuracy of management training program evaluation. *Performance and Instruction. 20*(8), 10-11, 16.

Pearson, R. W., Ross, M., & Dawes, R. M. (1991). Personal recall and the limits of retrospective questions in surveys. In J. Tanur (Ed.), Questions about questions (pp. 65-100). New York: Russell Sage Foundation.

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

Pratt, C. C., Mcgulgan, W. M., & Katzer, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation, 21,* 341-350.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rhodes, J. & Jason, L. (1987). The retrospective pretest: An alternative approach in evaluating drug prevention programs. *Journal of Drug Education, 17,* 345-356.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726-748.

Rohs, F. R. (1999). Response-shift bias: A problem in evaluating leadership development with self-report pretest-posttest measures. *Journal of Agricultural Education, 40*(4), 28-37.

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

Searle, M. S., Mahon, M. J., Iso-Ahola, S. E., Sodrolias, H. A., & Van Dyck, J. (1995). Enhancing a sense of independence and psychological well-being among the elderly: A field experiment. *Journal of Leisure Research, 27,* 107-124.

Sibthorp, J., Paisley, K., Gookin, J., & Ward, P. (2005). Participant Development through Adventure-Based Recreation Programming: A Model from The National Outdoor Leader-

ship School [Abstract]. *Proceedings of the 2005 NRPA Leisure Research Symposium, San Antonio, TX,* 11.

Sprangers, M., & Hoogstraten, J. (1988a). On delay and reassessments of retrospective ratings. *Journal of Experimental Education, 56*(3), 148-153.

Sprangers, M., & Hoogstraten, J. (1988b). Response-style effects, response-shift bias, and a bogus pipeline: A replication. *Psychological Reports, 62*(1), 11-16.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*(2), 265-273.

Townsend, M. & Wilton, K. (2003). Evaluating change in attitude towards mathematics using the 'then-now' procedure in a cooperative learning programme. *British Journal of Educational Psychology, 73,* 473-488.

Toupence, R. & Townsend, C. (2000). *Leadership development and youth camping: Determining a relationship.* In A. Stringer et. al. (Eds.). Coalition for Education in the Outdoors Fifth Research Symposium Proceedings (pp. 82-88). Cortland, NY: CEO.

Witt, P. (2000). If leisure research is to matter II. *Journal of Leisure Research, 32,* 186-189.