

5-1-2005

# An Empirical Evaluation Of The Retrospective Pretest: Are There Advantages To Looking Back?

Paul A. Nakonezny

*The University of Texas Southwestern Medical Center*

Joseph Lee Rodgers

*University of Oklahoma*

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Nakonezny, Paul A. and Rodgers, Joseph Lee (2005) "An Empirical Evaluation Of The Retrospective Pretest: Are There Advantages To Looking Back?," *Journal of Modern Applied Statistical Methods*: Vol. 4 : Iss. 1 , Article 22.

DOI: 10.22237/jmasm/1114906920

## An Empirical Evaluation Of The Retrospective Pretest: Are There Advantages To Looking Back?

Paul A. Nakonezny  
Center for Biostatistics and Clinical Science  
University of Texas Southwestern Medical Center

Joseph Lee Rodgers  
Department of Psychology  
University of Oklahoma

---

This article builds on research regarding response shift effects and retrospective self-report ratings. Results suggest moderate evidence of a response shift bias in the conventional pretest-posttest treatment design in the treatment group. The use of explicitly worded anchors on response scales, as well as the measurement of knowledge ratings (a cognitive construct) in an evaluation methodology setting, helped to mitigate the magnitude of a response shift bias. The retrospective pretest-posttest design provides a measure of change that is more in accord with the objective measure of change than is the conventional pretest-posttest treatment design with the objective measure of change, for the setting and experimental conditions used in the present study.

Key words: Response shift bias, quasi-experimentation, retrospective pretest-posttest design, retrospective pretest, measuring change

---

### Introduction

More than 30 years after Cronbach and Furby (1970) posited their compelling question, “How we should measure change—or should we?,” the properties of the change score continue to attract much attention in educational and psychological measurement. Self-report evaluations are frequently used to measure change in treatment and educational training interventions. In using self-report instruments, it is assumed that a subject’s understanding of the standard of measurement for the dimension being measured will not change from pretest to posttest (Cronbach & Furby, 1970).

If the standard of measurement is not comparable between the pretest and posttest scores, however, then self-report evaluations in pretest-posttest treatment designs may be contaminated by a response shift bias (Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Maxwell & Howard, 1981). A response shift becomes a bias if the experimental intervention changes the subject's internal evaluation standard for the dimension measured and, hence, changes the subject's interpretation of the anchors of a response scale.

When a response shift is presumably a result of the treatment, a treatment-induced response shift bias should occur in the treatment group and not in the control group. However, another possible source of contamination in response shifts, for both the treatment and control groups, is exposure to the conventional pretest, which could have a priming effect and confounding influence on subsequent self-report ratings (Hoogstraten, 1982; Spranger & Hoogstraten, 1989). A response shift, nevertheless, results in different scale units (metrics) at the posttest than at the pretest, which could produce systemic errors of measurement that threaten evaluation of the basic treatment effect.

---

Paul A. Nakonezny is an Assistant Professor of Biostatistics in the Center for Biostatistics and Clinical Science at the University of Texas Southwestern Medical Center, 6363 Forest Park Rd., Suite 651, Dallas, TX 75235. Email: paul.nakonezny@utsouthwestern.edu. Joseph Lee Rodgers is a Professor of Psychology in the Department of Psychology at the University of Oklahoma, Norman, OK, 73019.

When self-report evaluations must be used to measure change, the traditional pretest-posttest treatment design can be modified to include a retrospective pretest at the time of the posttest (e.g., Howard & Dailey, 1979; Howard, Millham, Slaten, & O'Donnell, 1981; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck, & Bray, 1979). After filling out the posttest, subjects then report their memory or perception of what their score would have been prior to the treatment (this is referred to as a retrospective self-report pretest).

Because it is presumed that the self-report posttest and the retrospective self-report pretest would be filled out with respect to the same internal standard, a comparison of the traditional pretest with the retrospective pretest scores within the treatment group would provide an indication of the presence of a response shift bias (Howard et al., 1979). If a response shift bias is present, as indicated by an appreciable mean difference between scores on the conventional pretest and the retrospective pretest, then comparison of the posttest with the retrospective pretest scores would eliminate treatment-induced response shifts and, thus, provide an unconfounded and unbiased estimate of the treatment effect (Howard et al. 1979).

Thus, the retrospective self-report pretest is a method that can be used to obtain pretreatment estimates of subjects' level of functioning (on cognitive, behavioral, and attitudinal dimensions) that are measured with respect to the same internal standard (i.e., in a common metric) as the posttest rating. Retrospective self-report pretests could be used in at least three evaluation research settings: (a) to attenuate a response shift bias (as mentioned above), (b) when conventional pretest data or concurrent data are not available, or (c) when researchers want to measure change on dimensions not included in earlier-wave longitudinal data.

However, the use of retrospective self-reports in the measurement of change has not gained popular acceptance among social scientists. There seem to be at least two possible, yet related, reasons for skepticism and reservation concerning the use of retrospective ratings. First, retrospective self-reports may be perceived to be counter to the paradigm of

objective measurement that is rooted in the philosophy of logical positivism (an epistemology in the social sciences that views subjective measures as obstacles toward an objective science of measurement). Second, retrospective self-reports are susceptible to a response-style bias (e.g., memory distortion, subjects' current attitudes and moods, subject acquiescence, social desirability), which could presumably affect ratings in both the treatment and control groups.

Nonetheless, in self-report pretest-posttest treatment designs, previous psychometric research has demonstrated empirical support for the retrospective pretest-posttest difference scores over the traditional pretest-posttest change scores in providing an index of change more in agreement with objective measures of change on both cognitive and behavioral dimensions (e.g., Hoogstraten, 1982; Howard & Dailey, 1979; Howard, Millham, Slaten, & O'Donnell, 1981; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck, & Bray, 1979; Spranger & Hoogstraten, 1989).

The purpose of this article is to build on a previous line of research, by Howard and colleagues and Hoogstraten and Spranger, on response shift effects and retrospective self-report ratings. Specifically, the current study examined (a) response shift bias in the self-report pretest-posttest treatment design in an evaluation setting, (b) the validity of the retrospective pretest-posttest design in estimating treatment effects, (c) the effect of memory distortion on retrospective self-report pretests, and (d) the effect of pretesting on subsequent and retrospective self-report ratings.

### Methodology

A cross-sectional quasi-experimental pre-post treatment design (Cook & Campbell, 1979) with data from 240 participants was used to address the research objectives of this study. The design included a treatment group and a no-treatment comparison group. Participants in the treatment group were 124 students enrolled in an undergraduate epidemiology course (Class A) and participants in the no-treatment comparison group were 116 students enrolled in an

undergraduate health course (Class B). The 240 participants were undergraduate students who attended a large public University in the state of Texas during the Spring semester of 2002 and who met the following criteria for inclusion in the study:

- (a) at least 18 years of age,
- (b) must not have taken an epidemiology course or a course that addressed infectious disease epidemiology, and
- (c) must not have been concurrently enrolled in Class A and Class B.

Participants signed a consent form approved by the Institutional Review Board of the University and received bonus class points for participating. The gender composition was 29 males and 211 females, and the age range was 18 to 28 years (with an average age of 20.61 years,  $SD = 2.46$ ). The racial distribution of the study sample included 181 (75.4 %) Caucasians, 37 (15.4 %) African Americans, 13 (5.4 %) Hispanics, and 9 (3.8 %) Asians. Participant characteristics by group are reported in Table 1.

The treatment in this design was a series of lectures on infectious disease epidemiology that was part of the course content in Class A, but not in Class B. Participants' knowledge of infectious disease epidemiology—the basic construct in this study—was measured with a one-item self-report instrument and with a ten-item objective instrument, and the same item-scale instruments were used for both the treatment and no-treatment comparison groups. Each instrument was operationalized as the mean of the items measuring each scale, and was scored so that a higher score equaled more knowledge of infectious disease epidemiology.

The conventional self-report instrument, which was used in both the pretest and posttest measurement settings, consisted of one-item that asked participants to respond to the following question: "How much do you know about the principles of Infectious Disease Epidemiology?" The current study measured this one-item using a six-point Likert-type scale that ranged from 0 (not much at all) to 5 (very very much), with verbal labels for the intermediate scale points.

The retrospective self-report pretest, which was similar to the conventional self-report

pretest, consisted of one-item that asked participants to respond to the following question: "Three months ago, at the beginning of the semester, you were asked how much you knew about Infectious Disease Epidemiology. Thinking back 3 months ago, to the beginning of the semester, how much did you know about Infectious Disease Epidemiology at that time?"

The current study measured this one retrospective item using a six-point Likert-type scale like that mentioned above. The objective instrument, which was used in both the pretest and posttest measurement settings, consisted of 10 multiple choice items/questions that tapped the participants' knowledge level of infectious disease epidemiology.

Participants within each group—treatment group and no-treatment comparison group—were randomly assigned to four pretesting conditions, which represented the pretesting main effect. Participants in condition 1 completed both the self-report and objective pretests. Participants in condition 2 completed the objective pretest. Participants in condition 3 completed the self-report pretest. Participants in condition 4 completed neither the self-report pretest nor the objective pretest.

All participants, regardless of the assigned condition, completed the posttests as well as the retrospective and recalled self-report pretests. The sample size per condition by group was approximately equal, and the participants across the four conditions were not significantly different in age,  $F$ 's  $< .91$ ,  $p$ 's  $> .43$ , gender, race, and academic classification (e.g., freshman, sophomore, junior, senior), respectively,  $\chi^2$ 's  $< 1.08$ ,  $p$ 's  $> .29$ .

At the outset of the academic semester (time 1), before the treatment, all participants in the assigned condition completed the pretest(s) which measured their baseline knowledge level of infectious disease epidemiology. The pretests were collected immediately after they were completed and then the treatment was begun (for participants in the treatment group). At the conclusion of the instruction on infectious disease epidemiology (the treatment), which occurred at about the end of the 12<sup>th</sup> week of classes (time 2), participants in the treatment

Table 1. Participant Characteristics

Variable	Treatment Group (n= 124)			Comparison Group (n= 116)			p
	Mean	SD	n (%)	Mean	SD	n (%)	
Age (years)	20.5	1.9	124 (51.7)	20.6	2.9	116 (48.3)	.66 <sup>a</sup>
Gender							.68 <sup>b</sup>
Male			16 (12.9)			13 (11.2)	
Female			108 (87.1)			103 (88.8)	
Race							.58 <sup>b</sup>
White			91 (73.4)			90 (77.6)	
Black			21 (16.9)			16 (13.8)	
Hispanic			08 (06.5)			05 (04.3)	
Asian			04 (03.2)			05 (04.3)	
Classification							.65 <sup>b</sup>
Freshman			17 (13.7)			22 (19.0)	
Sophomore			42 (33.9)			41 (35.3)	
Junior			49 (39.5)			40 (34.5)	
Senior			16 (12.9)			13 (11.2)	

<sup>a</sup>F statistic was used to test for mean age differences between the treatment group and the no-treatment comparison group.

<sup>b</sup>Chi-Square statistic was used to test for differences between the treatment group and the no-treatment comparison group on gender, race and classification, respectively.

group and participants in the no-treatment comparison group (who were not exposed to the treatment) completed the objective posttest. The objective posttest was identical to the objective pretest.

One week after completion of the objective posttest (time 3), participants in both the treatment and no-treatment comparison groups completed the self-report posttest and the retrospective self-report pretest. Participants first completed the self-report posttest and, while keeping the posttest in front of them, they then filled out the retrospective self-report pretest.

The self-report posttest was identical to the conventional self-report pretest. The retrospective self-report pretest was similar to the conventional self-report pretest, but the wording of the question accounted for the retrospective time frame.

Lastly, about one month after completion of the self-report posttest and retrospective self-report pretest, at the end of the academic semester (time 4), participants in both the treatment and no-treatment comparison groups completed the recalled self-report pretest, which permitted a memory test of the initial/conventional self-report pretest completed at the outset of the academic semester (time 1)

and, thus, yielded a test for a response-style bias of the retrospective self-report pretest rating.

The recalled self-report pretest consisted of one-item that asked participants to respond to the following question: "Four months ago, at the beginning of the semester, you were asked how much you knew about Infectious Disease Epidemiology. Please recall, remember, and be as accurate as possible, how you responded at that time regarding your knowledge level of Infectious Disease Epidemiology (i.e., how did you respond at that time?)." The current study measured this one-item using a six-point Likert-type scale similar to that described above.

The research objectives of this study were addressed by analyzing the series of pretest and posttest ratings using the dependent *t* test, the Pearson product-moment correlation (*r*), and analysis of variance (ANOVA). Estimates of the magnitude of the effect size were also computed (Rosenthal, Rosnow, & Rubin, 2000). The effect size estimators that accompanied the dependent *t* test and the ANOVA were Cohen's (1988) *d* and eta-square ( $\eta^2$ ), respectively.

The Pearson product-moment correlation (*r*) was also used as the effect size estimator in the specific regression analyses. To test the response shift hypothesis, the dependent *t* test was carried out comparing the retrospective self-report pretest to the conventional self-report pretest within the treatment and no-treatment comparison groups. The dependent *t* test also was used to compare the recalled self-report pretest to the conventional self-report pretest, which tested for the effect of memory distortion in the retrospective pretest-posttest design.

The Pearson correlation between the recalled self-report pretest and the conventional self-report pretest and between the recalled self-report pretest and the retrospective self-report pretest also was used to test for memory distortion. To examine the relative validity of the retrospective pretest-posttest design in estimating treatment effects, a simple correlation analysis was further used to assess the relationship between the self-reported measures of change and the objective measure of change in both the conventional and retrospective

pretest-posttest designs for the treatment and no-treatment comparison groups.

One-way ANOVA was used to assess the pretesting main effect (the four pretesting conditions) on the conventional self-report posttest, the retrospective self-report pretest, and the recalled self-report pretest. The Ryan-Einot-Gabriel-Welsch multiple-range test was used to carry out the cell means tests for the pretesting main effect for the ANOVA. A separate ANOVA was performed for the treatment group and the no-treatment comparison group.

## Results

### Response Shift

Using the conventional pre/post self-report change score and the objective pre/post change score, effects were found in the treatment group,  $t$ 's  $> 8.60$ ,  $p$ 's  $< .0001$ , but not in the no-treatment group,  $t$ 's  $< .84$ ,  $p$ 's  $> .40$ . The dependent *t* test, averaged across conditions 1 and 3, revealed a marginally significant mean difference between the retrospective self-report pretest and the conventional self-report pretest in the treatment group,  $t(61) = -1.56$ ,  $p < .10$ ,  $M = -0.16$ ,  $SD = .81$ ,  $d = -0.20$ , and, unexpectedly, a significant mean difference in the no-treatment comparison group,  $t(54) = -2.99$ ,  $p < .004$ ,  $M = -0.30$ ,  $SD = .76$ ,  $d = -0.39$ . These findings provide moderate support for the response shift hypothesis. Means and standard deviations for the pretests and posttests by condition and group are reported in Table 2.

### Treatment Effects

To assess the relative validity of the retrospective pretest-posttest design in estimating treatment effects, the self-reported measures of change were compared with the objective measure of change in both the conventional and retrospective pretest-posttest designs for the treatment and no-treatment comparison groups. For the treatment group, averaged across conditions 1 and 2, the Pearson correlation results indicated that the retrospective pre/post self-report change score was somewhat more in accord with the objective pre/post measure of change ( $r = .32$ ,  $p < .01$ ) than was the conventional pre/post self-report

Table 2. Means and Standard Deviations for the Pretests and Posttests by Condition and Group.

		Treatment Group (n = 124)					
		Self-Report				Objective	
Pretest Condition		Pretest	Posttest	Retro	Recalled	Pretest	Posttest
<b>Condition 1</b>							
M		1.10	2.34	0.86	1.13	1.82	4.06
SD		1.01	0.81	0.87	0.91	0.76	0.98
<b>Condition 2</b>							
M			2.81	1.18	1.65	1.89	3.78
SD			0.82	0.93	1.09	0.82	0.73
<b>Condition 3</b>							
M		0.99	2.21	0.91	1.30		3.48
SD		1.14	0.92	0.80	0.95		1.01
<b>Condition 4</b>							
M			2.43	1.03	1.26		3.66
SD			0.77	0.93	0.86		0.93
		No-Treatment Comparison Group (n = 116)					
		Self-Report				Objective	
Pretest Condition		Pretest	Posttest	Retro	Recalled	Pretest	Posttest
<b>Condition 1</b>							
M		0.79	0.86	0.52	0.83	1.67	1.50
SD		0.82	0.87	0.78	0.85	0.66	0.79
<b>Condition 2</b>							
M			1.09	0.71	0.99	1.68	1.67
SD			0.98	0.86	0.89	0.56	0.67
<b>Condition 3</b>							
M		1.07	1.19	0.73	1.03		1.82
SD		0.84	0.75	0.87	0.91		0.61
<b>Condition 4</b>							
M			1.43	0.67	0.83		1.55
SD			0.89	0.71	0.79		0.72

Note. Retro = retrospective self-report pretest; Recalled = recalled self-report pretest (used to test for the threat of memory distortion). Participants in condition 1 completed both the self-report and objective pretests; Participants in condition 2 completed the objective pretest; Participants in condition 3 completed the self-report pretest; Participants in condition 4 completed neither the self-report pretest nor the objective pretest. All participants, regardless of the assigned condition, completed the posttests as well as the retrospective and recalled self-report pretests. The sample size per condition by group was approximately equal.

change score with the objective pre/post measure of change ( $r = .26, p < .18$ ).

Conversely, as anticipated, for the no-treatment comparison group averaged across conditions 1 and 2, the magnitude of the correlation between the conventional pre/post self-report change score and the objective pre/post measure of change,  $r = .27, p < .16$ , was greater than the correlation between the retrospective pre/post self-report change score and the objective pre/post change score,  $r = .04, p < .75$ , albeit neither was significant.

#### Memory Distortion

The effect of memory distortion within the retrospective pretest-posttest design was also examined. For the treatment group, averaged across conditions 1 and 3, the results of the dependent  $t$  test revealed no significant mean difference between the recalled self-report pretest ( $M = 1.22, SD = .93$ ) and the conventional self-report pretest ( $M = 1.05, SD = 1.07$ ),  $t(61) = 1.56, p < .12, M = .17, SD = .89, d = 0.19$  (Table 2). Further, the no-treatment comparison group had nearly identical average scores on the recalled self-report pretest ( $M = .933, SD = .882$ ) and the conventional self-report pretest ( $M = .935, SD = .832$ ), averaged across conditions 1 and 3, suggesting no significant mean difference,  $t(54) = -0.01, p < .99, M = -0.002, SD = .85, d = -0.002$  (Table 2). The dependent  $t$  test results suggest no significant presence of memory distortion in the retrospective pretest-posttest treatment design.

A simple correlation analysis also was used to test for memory distortion. The Pearson correlations between the recalled pre/post self-report change score and the conventional pre/post self-report change score, averaged across conditions 1 and 3, and between the recalled pre/post self-report change score and the retrospective pre/post self-report change score, averaged across all four conditions, were significant and reasonably high in the treatment group ( $r = .64$  and  $r = .63$ , respectively,  $p$ 's  $< .0001$ ) and in the no-treatment comparison group ( $r = .54$  and  $r = .56$ , respectively,  $p$ 's  $< .0001$ ).

Further, the Pearson correlations between the recalled self-report pretest and the conventional self-report pretest, averaged across

conditions 1 and 3, and between the recalled self-report pretest and the retrospective self-report pretest, averaged across all four conditions, were significant and fairly high in the treatment group ( $r = .61$  and  $r = .62$ , respectively,  $p$ 's  $< .0001$ ) and in the no-treatment comparison group ( $r = .60$  and  $r = .68$ , respectively,  $p$ 's  $< .0001$ ).

#### Pretesting Effects

The ANOVA revealed a significant pretesting effect on the conventional self-report posttest in the treatment group,  $F(3, 120) = 3.04, p < .03, \eta^2 = .07$ , but not in the no-treatment comparison group,  $F(3, 112) = 2.11, p < .10, \eta^2 = .05$ . The cell means tests, however, indicated no significant difference between the conventional self-report pretest condition and the no-pretest condition on the conventional self-report posttest score in the treatment and no-treatment comparison groups,  $t$ 's  $< 1.05, p$ 's  $> .30$ . Further, the ANOVA revealed no significant pretesting effect on the retrospective self-report pretest and on the recalled self-report pretest in the treatment group,  $F$ 's(3, 120)  $< 1.64, p$ 's  $> .18, \eta^2$ 's  $< .04$ , and in the no-treatment comparison group,  $F$ 's(3, 112)  $< 0.46, p$ 's  $> .70, \eta^2 = .01$ . The ANOVA results suggest that pretesting had little effect on the subsequent and retrospective self-report ratings. Means and standard deviations for the pretests and posttests by condition and group are reported in Table 2.

#### Response Shift

Do treatments in evaluation research alter participants' perceptions in a manner which contaminates self-report assessment of the treatment? The findings of the current study indicate moderate evidence of a response shift bias in the conventional pretest-posttest treatment design in the treatment group, suggesting that the knowledge ratings from self-report pretest to posttest were partially a result of respondents recalibrating their internal evaluation standard for the dimension measured. A plausible interpretation of this moderate response shift bias in the treatment group is that the use of explicitly worded anchors on response scales in measuring the participant's self-



reported knowledge of infectious disease epidemiology—a cognitive construct—in a classroom setting helped to mitigate the magnitude of a response shift effect.

The degree of a response shift bias is, in part, conditional upon the experimental setting, the type of constructs measured, and the response scale anchors. Previous research (e.g., Collins et al., 1985; Finney, 1981; Howard, Schmeck, & Bray, 1979; Maisto et al., 1982) suggests that the magnitude of a response shift bias seems to be smaller when cognitive constructs are measured (such as knowledge ratings) and when questions and anchors on response scales are explicit.

Although no treatment effects were found in the no-treatment comparison group, as expected, a significant mean difference between the retrospective self-report pretest and the conventional self-report pretest was found, suggesting a non-treatment-related response shift. Typically, a response shift is a result of respondents changing their internal evaluation standard for the dimension measured between pretest and posttest because of exposure to the treatment. There are, however, alternative sources of bias in response shifts—such as a pretesting effect, memory distortion, and subject acquiescence—which could presumably affect ratings in both the treatment and no-treatment comparison groups (Collins et al., 1985; Howard & Dailey, 1979; Sprangers & Hoogstraten, 1989).

Because the results of the current study suggest that memory distortion and pretesting had little effect on subsequent self-report ratings, a plausible explanation for the response shift bias in the no-treatment comparison group is subject acquiescence. In the case of subject acquiescence, participants in the no-treatment comparison group might have realized that their knowledge level had not changed since their initial pretest rating, but their desire to provide the experimenter with a favorable set of results (given that bonus grade points were given for participation in the study) led them to lower their retrospective self-report rating. The retrospective rating was administered at the same time as the self-report posttest, allowing participants in the no-treatment comparison group the opportunity to adjust their

retrospective preratings in a downward direction.

#### Treatment Effects in the Retrospective Pre/Post Design

The principal focus of the current study was to evaluate the validity of the retrospective pretest-posttest design in estimating treatment effects. The findings of the present study favor the retrospective pre/post self-report measure of change in providing a measure of self-reported change that better reflects the objective index of change on a construct of knowledge rating. This finding is in line with previous psychometric research (e.g., Hoogstraten, 1982; Howard & Dailey, 1979; Howard et al., 1979; Howard, Schmeck, & Bray, 1979; Spranger & Hoogstraten, 1989), and is most likely a result of the self-report posttest and the retrospective self-report pretest being filled out with respect to the same internal standard, the same metric. This, therefore, mitigates the treatment-induced response shift bias, minimizes errors of measurement, and provides an unconfounded and unbiased estimate of the treatment effect (Howard et al., 1979).

Although there is empirical support for the retrospective pretest-posttest difference scores over the conventional pretest-posttest change scores in providing an index of change more in agreement with objective measures of change, this is not to suggest that the conventional self-report pretest should be substituted by the retrospective self-report rating. Rather, in light of the findings of this study as well as those from previous studies, the suggestion put forward is that retrospective self-report pretests could be used in at least three evaluation research settings: (a) to test for and attenuate a response shift bias in the conventional pretest-posttest treatment design, (b) when conventional pretest data or concurrent data are not available, or (c) when researchers want to measure change on dimensions not included in earlier-wave longitudinal data.

#### Testing for Threats to Validity

Also evaluated were the potential threats of memory distortion and pretesting effect to the internal validity of the retrospective pretest-posttest treatment design in the current study.

Retrospective self-report ratings could be limited by memory lapses and pretests could exert a confounding influence on subsequent self-report ratings, including retrospective ratings, which could threaten evaluation of the treatment effect (Collins et al., 1985; Howard & Dailey, 1979; Sprangers & Hoogstraten, 1989). In general, the present study found no significant presence of memory distortion or a pretesting effect in the retrospective pretest-posttest treatment design used in the current study.

This is not to suggest that memory distortion or a pretesting effect should not be accounted for as potential threats to the basic retrospective pretest-posttest design. Rather, what this finding suggests is that memory distortion and pretesting are not influencing the interpretation of the treatment effect in the type of retrospective pretest-posttest design used in the present study. The conventional self-report pretest and the recalled self-report pretest were only separated by four months, which may have in part mitigated the effect of memory distortion. Previous research (e.g., Finney, 1981; Howard, Dailey, & Gulanick, 1979; Howard, Schmeck, & Bray, 1979; Maisto et al., 1982), nonetheless, suggests that a pretesting effect can be mitigated and moderate-to-high recall accuracy is possible when cognitive constructs are measured (such as knowledge ratings) and when retrospective questions are specific and anchors on response scales are explicit (these conditions are consistent with those used in this study).

#### An Application of the Retrospective Pre/Post Design

In this section, a study by Nakonezny, Rodgers, and Nussbaum (2003) which applied the retrospective pretest-posttest treatment design to a unique research setting is briefly described.

Nakonezny et al. (2003) examined the effect of later life parental divorce on solidarity in the relationship between the adult child and older parent. This examination was achieved by testing the buffering hypothesis that greater levels of predivorce solidarity in the adult child/older parent relationship buffers damage to postdivorce solidarity. The unique and uncommon nature of the phenomenon of later life parental divorce, however, precluded access

to these atypical divorcees prior to their divorce, which led to the necessity to use a retrospective pretest-posttest treatment design by Nakonezny et al. (2003).

As mentioned earlier, one research scenario under which retrospective self-report pretests could be used is when conventional pretest data are not available, which was the case in the Nakonezny et al. (2003) study.

In the retrospective design used in Nakonezny et al. (2003), predivorce/pretest solidarity included retrospective measures of the same scale-item instruments that were used to measure postdivorce/posttest solidarity. The wording of the questions, however, was changed to account for the retrospective time frame. Parents in the divorced group were asked to remember the period before their divorce and to provide a retrospective self-report account of solidarity in the relationship with their oldest living adult child during the predivorce period. The average number of years from the divorce decree to the date of data collection was about 8 years.

Also, parents in the intact two-parent family group (the no-treatment comparison group) were asked to remember back approximately five years from the date of participation in the study and to provide a retrospective self-report account of solidarity in the relationship with their oldest living adult child during that period, which represented the pretest period for the intact group. The basic findings of Nakonezny et al. (2003), using a retrospective pretest-posttest treatment design, were in the hypothesized directions for both groups. Nakonezny et al. (2003) can be consulted for a complete explanation of this application of the retrospective pretest-posttest treatment design in a social science evaluation research setting.

#### Future Research

The current study and previous research suggest that, under certain conditions, the retrospective pretest-posttest treatment design provides a more accurate assessment of change than that of the conventional pretest-posttest treatment design. However, the retrospective pretest-posttest treatment design still remains something of an enigma, and future research

concerning the validity of the retrospective pretest-posttest design is still needed. Further research is needed to address the effect of subject acquiescence and other extraneous sources of invalidity on self-report ratings in the retrospective pretest-posttest treatment design.

Further research also is needed to determine the different types of retrospective pretest-posttest designs, experimental conditions, treatment interventions, constructs, and time lapses that are most susceptible to a response shift bias and that most affect recall accuracy of retrospective self-report ratings. Most importantly, a next step in this line of evaluation research is to continue to explore the research settings and applications in both the social and behavioral sciences under which retrospective self-report ratings are appropriate and under which the retrospective pretest-posttest design produces unbiased estimates of treatment effects.

#### Conclusion

The empirical findings support that a moderate response shift bias occurred in the conventional pretest-posttest treatment design in the treatment group, and are highly suggestive that the knowledge ratings from self-report pretest to posttest were partially a result of respondents recalibrating their internal evaluation standard for the dimension measured (presumably because of exposure to the treatment). The results further suggest that the use of explicitly worded anchors on response scales as well as the measurement of knowledge ratings (a cognitive construct) in an evaluation methodology setting mitigated the magnitude of a response shift bias. Subject acquiescence is a likely explanation of the unexpected non-treatment-related response shift bias that occurred in the no-treatment comparison group.

Further, the current study suggests that the retrospective pretest-posttest treatment design provides a more accurate assessment of change than that of the conventional pretest-posttest treatment design for the setting and experimental conditions used in the present study. Based on these results, it is suggested that researchers collect both a conventional self-report pretest and a retrospective self-report pretest when

using a conventional pretest-posttest treatment design in evaluation research settings. Retrospective self-report pretests could be used, however, when conventional self-report pretest data are not available. In support of this scenario, we present an example of an innovative application of the retrospective pretest-posttest treatment design in a social science research setting. Finally, the ultimate value of this work may lie in its ability to renew interest in the retrospective pretest-posttest treatment design, to motivate future research, and to sharpen the empirical focus of that research.

#### References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Collins, L. M., Graham, J. W., Hansen, W. B., & Johnson, C. A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. *Applied Psychological Measurement, 9*, 301-309.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.
- Cronbach, L. J., & Furby, L. (1970). How we should measure change--or should we? *Psychological Bulletin, 74*, 68-80.
- Finney, H. C. (1981). Improving the reliability of retrospective survey measures: Results of a longitudinal field survey. *Evaluation Review, 5*, 207-229.
- Hoogstraten, J. (1982). The retrospective pretest in an educational training context. *Journal of Experimental Education, 50*, 200-204.
- Howard, G. S., & Dailey, P. R. (1979). Response shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*, 144-150.
- Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subjective response style effects on retrospective measures. *Applied Psychological Measurement, 5*, 89-100.

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

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.

Maisto, S. A., Sobell, L. C., Cooper, A. M., & Sobell, M. B. (1982). Comparison of two techniques to obtain retrospective reports of drinking behavior from alcohol abusers. *Addictive Behaviors, 7*, 33-38.

Maxwell, S. E. & Howard, G. S. (1981). Change scores: Necessarily anathema? *Educational and Psychological Measurement, 41*, 747-756.

Nakonezny, P. A., Rodgers, J. L., & Nussbaum, J. F. (2003). The effect of later life parental divorce on adult-child/older-parent solidarity: A test of the buffering hypothesis. *Journal of Applied Social Psychology, 33*, 1153-1178.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

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