A Review of the Retrospective Pretest: Implications for Performance Improvement Evaluation and Research

Kim Nimon
Jeff Allen

University of North Texas

Return to the Table of Contents

Abstract

The retrospective pretest has been used in evaluating program outcomes for over 50 years as a moderator for the threat of response-shift bias. This paper reviews its origins, describes methodology that encompasses its use, identifies strengths and weaknesses, and concludes with a research agenda to enhance contemporary research designs concerned with determining program impact. The paper cites over 40 key references, provides sample SPSS code, and contains an index of representative studies that have employed retrospective pretest methodology.

Introduction

The retrospective pretest is experiencing a resurgence in behavioral science research. Building on psychometric studies conducted in the early 1950s, Howard (1980) began prescribing this research tool as a remedy for response-shift bias. In essence, a retrospective pretest distinguishes itself from the traditional pretest by its relationship to the intervention (Howard, Ralph, Gulanick, Maxwell, & Gerber, 1979). That is, the retrospective pretest is administered post-intervention, asking subjects to recall their behavior prior to the intervention (or treatment). For some types of self-report measures, the retrospective pretest is a more accurate measure of pre-intervention behavior because it removes the confounding effect of the standard of measurement changing due to the subjects' understanding of their level of functioning as a consequence of the intervention (Howard, Schmeck, & Bray, 1979).

Over 25 years have passed since the seminal work of Howard, Ralph, et al. (1979). However, the retrospective pretest is making its way toward the forefront as evidenced by its appearance in recent trade and academic literature (e.g., Hill & Betz, 2005; Lamb, 2005). This paper reviews the origins of the retrospective pretest, describes methodology that encompasses its use, identifies strengths and weaknesses, and concludes with a proposed research agenda to enhance contemporary research

designs incorporating the retrospective pretest.

The Origins

In its first implementation, retrospective pretesting was used across areas of psychology to obtain refined psychometrics, such as effects of racially mixed housing on prejudice (Deutsch & Collins, 1951), measurements of fear (Walk, 1956), and patterns of child rearing (Sears, Maccoby, & Levin, 1957). In these cases, traditional pretest measurements were impossible to obtain. However, by administering a retrospective pretest, researchers were able to ascertain pre-intervention differences between experimental and control groups and curb threats to validity that would have been associated with a posttest-only design.

Similarly, Campbell and Stanley (1963) offered retrospective pretesting as a tool to enhance experimental designs. Their assertion was that while the tool had certain limitations, it not only supplemented posttest-only designs, but also served to partially curb threats to validity including history, selective mortality, and shifts in subject perceptions.

Observing problems with detecting treatment effect, Howard and his colleagues referred back to earlier work employing the retrospective pretest (e.g., Deutsch & Collins, 1951; Sears et al., 1957; Walk, 1956) and found the tool to be an effective means to determine program outcomes in the presence of response-shift bias (Howard, Ralph, et al., 1979). The researchers' initial experimentation and experience with the pretest is documented through an article (i.e. Howard, Ralph, et al., 1979) that records a series of five empirical studies. Because literature (e. g., Hill & Betz, 2005; Lamb & Tschillard, 2005) often refers to these five seminal studies, they are briefly summarized in the following sections. Based on the findings of these and related studies (Howard & Dailey, 1979; Howard, Dailey, & Gulanick, 1979; Howard, Schmeck, et al., 1979), Howard (1980) argued that retrospective pretesting should be incorporated into traditional pretest-posttest designs (i.e., Campbell & Stanley's [1963] Designs 2 and 4) at least until an alternative method for removing the confounding effect that experimental treatment can have on instrumentation could be developed. As of late, such an alternative has not been recommended (Lam & Bengo, 2003).

Study I

Howard, Ralph, et al. (1979) set the stage for retrospective pretest when

they observed paradoxical findings in a self-report evaluation of an Air Force communication skills training program aimed at reducing dogmatism. After employing a traditional pretest-posttest design and finding an apparent increase in dogmatism following the workshop, Howard and his colleagues interviewed workshop participants and found that as a result of attending the workshop, participants changed their perceptions of their initial levels of dogmatism. Howard, Ralph, et al. recognized the change in subjects' basis for determining their level of functioning as response-shift and initiated a second study to examine its impact on determining treatment effect.

Study II

Howard, Ralph, et al. (1979) reasoned that analyzing self-report measures not contaminated by response-shift bias might yield different conclusions regarding the effectiveness of treatment. Therefore, they sought a method by which pre- and post-intervention ratings could be measured with respect to the same internal standard. It was hypothesized that substituting a traditional pretest with a retrospective pretest would eliminate the effects of treatment-produced response-shifts.

Armed with their theories as to how to account for the effects of response-shift bias, Howard, Ralph, et al. (1979) replicated Study I. In this instance, they divided participants into two groups. Group 1 employed a traditional pretest-posttest (post-pre) design. Group 2 employed a retrospective pretest-posttest (post-then) design. A statistically and practically significant number of Group 2 members reported becoming less dogmatic than group members using the post-pre procedure ($\bullet 2(1) = 11.17$, $p < .001$). While Group 1 reported no difference in subjects' levels of dogmatism as a result of the workshop, 71% of Group 2 participants reported becoming less dogmatic following the workshop. Additionally, the second group's results were in close agreement with the participants' written comments regarding the effectiveness of the workshop. As further measure of the study's validity, posttest scores between the two groups were compared and found not to be statistically significant.

The results of Study II suggested that for self-report measures, a retrospective pretest-posttest procedure might yield more accurate change scores than a conventional pretest-posttest design. This conclusion led the research team to consider other studies that could be conducted to validate their claim. In particular, the researchers refocused their efforts to examine variables that could be measured objectively so

that the data produced from the two different types of pretest could be analyzed in concert with independent outcome measures, thereby comparing the nonsubjective accuracy of the post-then change ratings to the post-pre change ratings. Studies III, IV, and V reflected this refocus.

Study III

In the third study, Howard, Ralph, et al. (1979) randomly assigned women who scored “highly feminine” on the Bem Sex-Role Inventory to control or experimental groups. The experimental groups were designed to promote androgyny by fostering the development of skills typically stereotyped as masculine. In order to monitor the effectiveness of these groups, self-report measures of assertiveness, sex-role orientation, and attainment of individual goals were administered as well as objective measures of change. Objective measures of change consisted of analyzing and coding subject’s verbal responses to eight taped stimulus situations pre- and post-treatment. Both groups also completed pretests, posttests, and retrospective pretests on the self-report measures, thus allowing generation of post-pre and post-then change scores to be compared to objective measures of change. Follow-up assessments were also made 2 months and 1 year after the treatment.

The effects of response-shift bias (defined as evidence of a statistically significant difference between pretest and retrospective pretest ratings (pre-then)) were evident for treatment subjects but not for control groups. Similar to Studies I and II, post-pre analyses demonstrated minimal treatment effects (4 out of 12 analyses resulted in statistically significant differences between groups), while post-then analyses produced statistically significant treatment effects in 8 out of the 12 analysis performed. Substituting 2-month follow-up scores for posttest scores produced similar results. Retrospective ratings obtained at the 1-year follow-up remained more similar to the original then-ratings than the original pre-ratings on the majority of scales employed, confirming the stability of the retrospective ratings over an extended period of time. Most important to the aim of the study, objective measures of change correlated more highly with post-then self-report measures of change than with the post-pre self-report index, adding further validity to the claim that retrospective pretests yielded more accurate change scores than conventional pretests for some types of self-report measures.

Study IV

The design of the fourth study (Howard, Ralph, et al., 1979) introduced

additional self-report measures of assertiveness and sex-role orientation to ascertain whether they were also subject to response-shift bias. Post-then analyses in Study IV replicated results from prior studies involving the retrospective pretest design methodology and produced more evidence for the effectiveness of the treatment than the post-pre self-report approach.

Study V

In the last study, Howard, Ralph, et al. (1979) analyzed changes in levels of helping skills for students taking a semester-long course. Subjects completed pretests, posttests, and retrospective pretests reporting their level of helping skills. Additionally, participants conducted half-hour interviews with volunteer clients before and after the course. Judges' ratings of the interviews and post-then comparisons found statistically significant treatment effects, whereas post-pre comparisons failed to show overall treatment effects. After completing posttests and retrospective pretests, subjects recalled their pretest ratings. Mean memory ratings were almost identical to pretest ratings, but statistically significantly different from retrospective pretest ratings, suggesting that the response-shift bias reflected something more than mere systematic memory distortions.

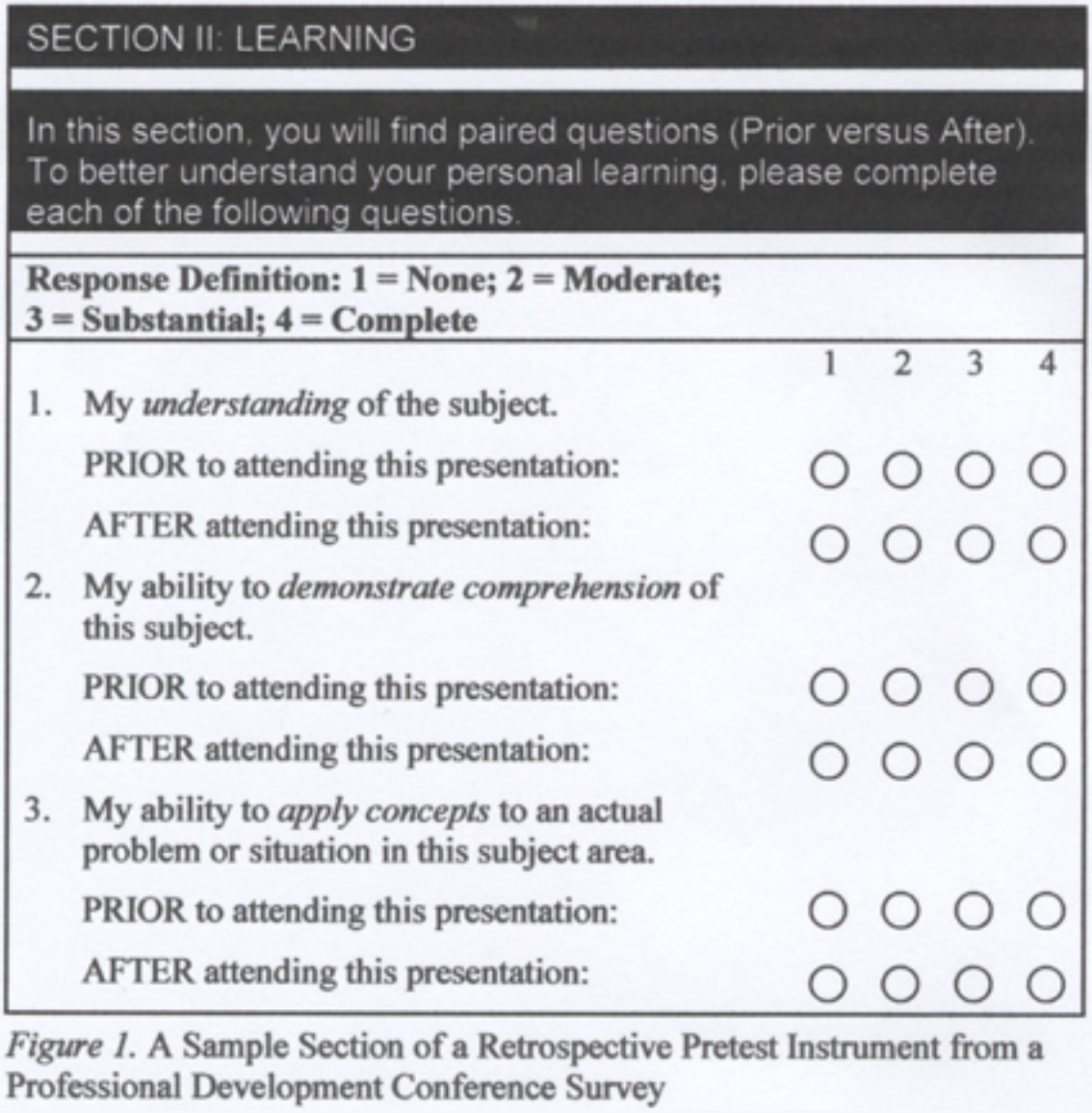
Interviews with subjects indicated that participants were typically aware that their retrospective ratings provided a different picture of their pretreatment level of functioning and volunteered explanations of why they believed their pretests to be inaccurate. Many individuals reported that as a consequence of taking the course, they changed their perceptions of their initial level of functioning.

This study culminated the seminal work of Howard, Ralph, et al. (1979) and lent strong support to the contention that when self-report measures are used in traditional pretest-posttest designs, the results might well be confounded by response shift. The retrospective pretest was thus proposed as a moderator to the confounding effect that an intervention can have on instrumentation and identified as adding a valuable dimension to the effort involved in determining treatment effect (Howard, Ralph, et al., 1979).

Methodology

Incorporating the retrospective pretest into a conventional pretest-

posttest design (pre-post-then design) requires changing posttest procedures so that subjects rate themselves twice: (a) as they perceive themselves post intervention and (b) as they perceived themselves prior to the treatment. See Figure 1 for sample instrumentation that simultaneously captures posttest and retrospective pretest ratings. The pre-post-then design also requires that researchers choose a method to analyze the data that results from the pretest, posttest, and retrospective pretest information (see Table 1 for hypothetical data on 20 subjects created for illustrative purposes). In particular, researchers must consider how to use the said constructs to detect the presence of a response-shift, measure the effect of treatment when such bias occurs, and control for error.



SECTION II: LEARNING

In this section, you will find paired questions (Prior versus After). To better understand your personal learning, please complete each of the following questions.

Response Definition: 1 = None; 2 = Moderate; 3 = Substantial; 4 = Complete

	1	2	3	4
1. My *understanding* of the subject.				
PRIOR to attending this presentation:	○	○	○	○
AFTER attending this presentation:	○	○	○	○
2. My ability to *demonstrate comprehension* of this subject.				
PRIOR to attending this presentation:	○	○	○	○
AFTER attending this presentation:	○	○	○	○
3. My ability to *apply concepts* to an actual problem or situation in this subject area.				
PRIOR to attending this presentation:	○	○	○	○
AFTER attending this presentation:	○	○	○	○

Figure 1. A Sample Section of a Retrospective Pretest Instrument from a Professional Development Conference Survey

Table 1

Hypothetical Data Set of Posttest, Pretest, and Retrospective Pretest Ratings From Self-Report Instrument With Two Subscales (A and B) and an Overall Scale (C) and Posttest and Pretest Ratings for Hypothetical Performance Instrument With One Scale (D)

Scale A			Scale B			Scale C			Scale D	
Post	Pre	Retro	Post	Pre	Retro	Post	Pre	Retro	Post	Pre
49.72	49.49	48.51	48.72	48.49	47.51	49.22	48.99	48.01	49.55	48.55
49.64	49.93	48.39	48.64	48.93	47.39	49.14	49.43	47.89	50.09	48.84
49.66	49.89	48.87	48.16	48.89	47.87	48.91	49.39	48.37	50.80	50.05
50.30	51.40	49.16	49.30	50.40	48.16	49.80	50.90	48.66	50.78	49.78
49.92	49.73	48.60	48.92	48.73	47.60	49.42	49.23	48.01	50.30	49.05
50.70	50.30	49.65	49.70	49.30	48.15	50.20	49.80	48.90	51.42	50.42
49.35	48.55	50.60	48.35	47.55	49.60	48.85	48.05	50.10	49.58	50.83
51.98	49.57	50.39	50.98	48.17	49.39	51.48	48.87	49.89	50.35	48.85
48.92	49.15	47.51	47.92	48.15	46.51	48.42	48.65	47.01	49.99	48.74
50.07	49.48	49.45	49.07	48.48	48.15	49.57	48.98	48.80	50.86	50.36
51.38	49.33	50.09	50.38	48.33	49.09	50.88	48.83	49.59	49.75	48.25
49.03	50.36	48.42	48.03	49.36	47.42	48.53	49.86	47.92	50.49	49.49
49.66	50.29	48.51	48.66	49.29	47.51	49.16	49.79	48.01	48.42	46.92
48.95	51.68	47.55	47.95	50.68	46.55	48.45	51.18	47.05	49.47	47.72
50.47	50.51	49.47	49.47	49.51	48.47	49.97	50.01	48.97	49.51	48.26
49.06	47.37	48.42	48.06	46.37	47.42	48.56	46.87	47.92	47.17	46.17
48.22	50.56	47.48	47.22	49.56	46.48	47.72	50.06	46.98	50.48	51.73
50.54	50.34	49.55	49.54	49.34	48.55	50.04	49.84	49.05	51.16	49.91
51.02	50.26	50.80	50.02	49.26	49.80	50.52	49.76	50.30	49.07	48.57
51.90	51.07	50.78	50.90	50.07	49.78	51.40	50.57	50.28	50.78	52.28

Detecting Response-shift Bias

While there are no standard procedures for detecting response-shift bias, the technique most often employed and promoted is the paired-samples t test (Craig, Palus, & Rogolsky, 2000). Alternative methods used to test for response-shift bias include item response theory (IRT)-based techniques (Craig et al., 2000) and response-shift bias model fitting procedures (Koele & Hoogstraten, 1988).

While alternative methods may arguably be superior to the paired-

samples t test, there are noteworthy limitations. Craig et al. (2000) supported item-response analysis and maintained that, in the presence of response-shift bias, interpreting the magnitude of differences between retrospective pretests and conventional pretests is just as flawed as comparing results between conventional pretests and posttests. However, preliminary work examining IRT-based techniques as a method for detecting response-shift bias employed small samples sizes and should be interpreted with caution (Craig et al., 2000). Koele and Hoogstraten's (1988) model fitting procedure is limited to designs that randomly assign participants to control or experimental groups and does not address the more common practice of testing treatments with intact groups (R. Henson, personal communication, October 13, 2005).

Because of these limitations and the preponderance of published research that examines response-shift bias using the paired-samples t test (e.g., Mezoff, 1981; Pratt, McGuigan, & Katzev, 2000; Sprangers & Hoogstraten, 1989; Umble, Upshaw, Orton, & Matthews, 2000), the SPSS syntax identified in Figure 2 depicts it use. As the paired-samples t test results in Table 1 illustrate, there is a statistical and practical difference between the retrospective pretests and traditional pretests for scales A, B, and C ($t = 2.59, 2.61, 2.60$ and $d = .82, .82, .83$, respectively), suggesting that a response shift bias has occurred. When armed with this information, the researcher should interview subjects to see whether and why they believe a response shift has occurred (Umble et al., 2000). If a substantial response-shift bias has occurred between pretest and posttest, the researcher must then select the appropriate pretest measure and technique for subsequent data analysis to measure treatment effect (Howard, 1980).

Measuring Treatment Effect

As with response-shift bias, a medley of methods exist to measure for treatment effects in the presence of response-shift bias (Bray, Maxwell, & Howard, 1984; Howard, Ralph, et al., 1979). Among the possible methods of analysis (comparison of posttest only scores, comparison of posttest to pretest scores, comparison of posttest means adjusted either by pretest means or retrospective pretest means through analysis of covariance [ANCOVA]), the technique that leads to an unbiased estimate of the treatment effect in the presence of a negative response shift is the comparison of posttest to retrospective pretest means (Howard, Ralph, et al., 1979). The other four approaches, on average, underestimate the true treatment effect when the difference between treatment effects and the response shift parameters are opposite in sign.

Not only does the comparison of posttest to retrospective pretest means

```
COMMENT Perform Paired-Sample T-Test (Post-Pre, Post-
RPT, Pre-RPT).

T-TEST
  PAIRS = APost APost APre BPost BPost BPre CPost CPost
CPre DPost  WITH APre ARetro ARetro BPre BRetro BRetro
CPre CRetro
  CRetro DPre (PAIRED)
  /CRITERIA = CI(.95)
  /MISSING = ANALYSIS.

COMMENT Compute Gain Scores

COMPUTE APost_Pre = APost - APre.
EXECUTE.

COMPUTE APost_Retro = APost - ARetro.
EXECUTE.

COMPUTE BPost_Pre = BPost - BPre.
EXECUTE.

COMPUTE BPost_Retro = BPost - BRetro.
EXECUTE.

COMPUTE CPost_Pre = CPost - CPre.
EXECUTE.

COMPUTE CPost_Retro = CPost - CRetro.
EXECUTE.
```

Figure 2. Sample SPSS Syntax for Analyzing Retrospective Pretest, Pretest, and Posttest Data

generally provide a more accurate estimate of the treatment effect, it is overall the most powerful method of analysis when there is a response shift (Bray et al., 1984). The SPSS syntax identified in Figure 2 illustrates this use. In the heuristic example, the analyses of scores for scales A, B, and C depicted in Table 2 show a statistical and practical treatment effect when comparing posttest to retrospective means ($t = 6.70, 6.50, 6.66$ and $d = .88, .89, .89$, respectively), whereas the comparison of posttest to pretest means yields little practical or statistical significance ($t = .21, .19, .20$ and $d = .06, .06, .06$, respectively). While analyzing posttest to retrospective means often yields a more accurate estimate of treatment effect and is the most powerful method of analysis

in the condition of response shift, researchers (e.g., Pratt et al., 2000; Umble et al., 2000) also recommend analyzing the traditional pretest to posttest means and reporting both results, similar to Table 2.

Table 2
Self-Report (Scales A-C) and Performance (Scale D) Pretest (Pre), Retrospective Pretest (RPT), and Post-test (Post) Item Means, Standard Deviations, t-values, Effect Sizes, and Correlations

	Means (Standard Deviations)			*t* values (Cohen's *d*)			Pearson's *r* between Self-Report and Performance Gain	
Scale	Pre	RPT	Post	Post-Pre	Post-RPT	Pre-RPT	Post-Pre	Post-RPT
A	49.96 (.97)	49.11 (1.06)	50.02 (1.02)	.21 (.06)	6.70*** (.88)	2.59* (.84)	.03	.59**
B	48.94 (.98)	48.07 (1.05)	49.00 (1.03)	.19 (.05)	6.50*** (.89)	2.61* (.86)	.05	.57**
C	49.45 (.98)	48.59 (1.05)	49.51 (1.03)	.20 (.06)	6.66*** (.89)	2.60* (.85)	.04	.59**
D	49.23 (1.50)	n/a	50.00 (1.00)	3.56** (1.27)	n/a	n/a	n/a	n/a

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. All are two-tailed paired-samples *t* tests, $df = 19$.

Because using multiple methods to assess change in behavioral science research is recommended (Gall, Gall, & Borg, 2003; Howard, Ralph, et al., 1979), objective measures of change provide another data point to consider when analyzing a design that incorporates the retrospective pretest. In these designs, researchers (e.g., Hoogstraten, 1985; Martineau, 2004; Umble et al., 2000) typically correlate and report self-report gains to performance gains as another indication of the treatment effect and the appropriate pretest measure to use. For example, in the heuristic data depicted in Table 2, the retrospective pretest appears to be a better measure of treatment effect because the gains in self-report measures (scales A, B, C) correlate higher with performance gains (scale D) than difference scores using the traditional pretest (r = .59, .57, .59 versus r = .03, .05, .04). Despite this evidence of concurrent validity, Howard, Ralph, et al. (1979) recommended viewing these comparisons with caution because there are fundamental problems with the reliability of change scores. Because simplified formulas, assuming parallelism in

pretest and posttest measures, present the reliability of gain scores in their most unfavorable light (Williams & Zimmerman, 1996), gain scores are perceived by some psychometric traditionalists as being inherently unreliable (Zumbo, 1999). Researchers interested in a review of alternative methodologies to assess the reliability and correlates of change are directed to Zumbo (1999).

Controlling for Error

As Table 2 illustrates a typical set of analyses involving retrospective pretests, it is important to note the implications of conducting multiple tests on the same data set (i.e., total of 10 paired sample t tests). While controlling for Type I error beyond comparing the results of each test to the ubiquitous .05 alpha level is not often depicted in studies employing the retrospective pretest (e.g., Hoogstraten, 1982; Lamb & Tschillard, 2005), there are exceptions. Lam and Bengo (2003) and Manthei (1997) employed the Scheffé method to maintain experimentwise error rate (αEW) at .05 for the comparisons performed on the retrospective pretest, pretest, and posttest information. Pratt et al. (2000) maintained αEW at .05 for the pairwise t tests performed by applying a strict Bonferroni adjustment. Readers interested in strategies for controlling for inflation of Type I error are directed to Maxwell and Delany (2004).

Post-then Design

Not all recent uses of the retrospective pretest follow the pre-post-then methodology described in this paper (e.g. Lam & Bengo, 2003; Raidl et al., 2005). In fact, some researchers (e.g., Lamb & Tschillard, 2005; Martineau, 2004) promoted the use of the retrospective pretest in lieu of the traditional pretest. Citing data that suggest traditional pretests underestimate the impact of intervention, Lamb and Tschillard (2005) asserted that the retrospective pretest is just as useful as the traditional pretest in determining program impact in the absence of response-shift bias and is even more useful in the case where subjects' understanding of their level of functioning changes as a consequence of the intervention. Similarly, Martineau (2004) opined that the retrospective pretest accounts for response-shift bias and correlates more highly with objective measures of change than self-report gains based on traditional pretest ratings.

Replacing the traditional pretest with the retrospective pretest is commonly referred to as a post-then design (Umble et al., 2000). In this design, after the intervention is complete, participants are asked two sets

of questions for each behavior measured. The first set are posttest questions because participants provide information about their behavioras it exists after the program. The second are then-test questions because they ask participants about how they felt then (i.e., before the program). Data analyses for this design follow the traditional pretest-posttest analyses as the traditional pretest information is simply replaced by the retrospective pretest data.

Strengths and Weaknesses

Amidst debates in behavioral science research, the use of the retrospective pretest solicits polarized perspectives and emotional exchanges (Howard, 1980; Martineau, 2004). As such, identifying the strengths and weaknesses of the retrospective pretests is not a simple binary process. Not only do researchers disagree concerning the robustness of the instrument (Martineau, 2004), but, in some cases, findings (Mann, 1997) have been promoted without full disclosure of their limitations. As such, the strengths and weaknesses summarized in Figure 3 are represented along a continuum in lieu of being depicted in a strict binary format. The placement along the continuum is based on a synthesis of the literature and is relative to the use of the tool in its encompassing methodology.

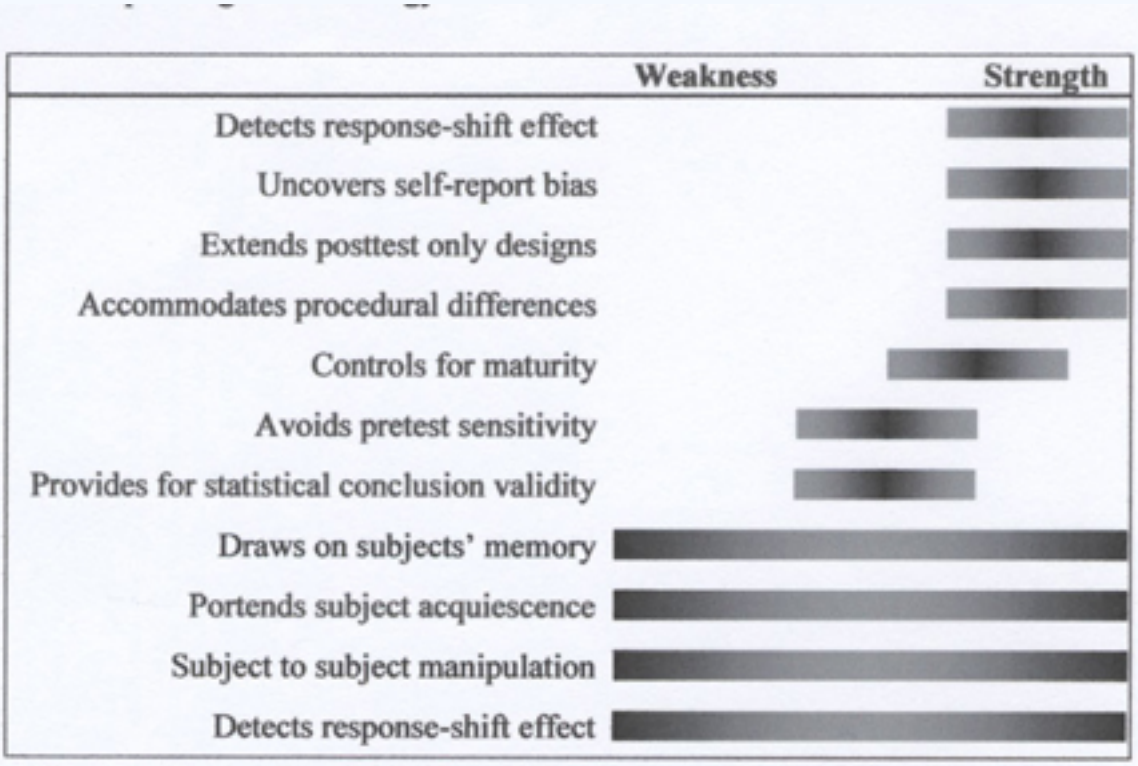


Figure 3. Strength and Weakness Continuum of the Retrospective Test

Perhaps the most noncontentious consideration of the retrospective pretest is that it provides another data point to post-only designs and helps researchers detect response-shift bias (Campbell & Stanley, 1963; Howard, 1980; Pratt et al., 2000). Notwithstanding the issues surrounding the methodology of its use in detecting response-shift bias, the retrospective pretest stands alone as a potential remedy for the confounding effect that an intervention can have on instrumentation (Lamb, 2005). Along the lines of response-shift bias, research also shows the retrospective pretest as an important tool in detecting other sources of bias in pretest self-report measures, including experience limitation, condition justification, altered states, and self-presentation (Aiken & West, 1990). To date, the retrospective pretest has been used to measure bias in self-report measures involving a broad range of cognitive, attitudinal, and skill-based variables. See Table 3 for a representative set of studies employing the retrospective pretest.

Table 3
Representative Studies Incorporating the Retrospective Pretest

Type	Variable	Study
Attitude	Career planning	Francis-Smythe and Philip (1997)
	Dogmatism	Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979)
	Internal motivation	Terborg and Davis (1982)
	Job involvement	Francis-Smythe and Philip (1997)
	Job motivation	Terborg and Davis (1982)
	Job satisfaction	Terborg and Davis (1982)
	Organizational commitment	Francis-Smythe and Philip (1997)
	Prejudice	Zwiebel (1987)
	Self-efficacy	Mann (1997); Townsend, Lai, Lavery, Sutherland, and Wilton (1999)
	Peer relationships	Aiken and West (1990)
Behavior	Abusiveness	Aiken and West (1990); Rhodes and Jason (1987)
	Anxiety	Townsend, Lai, Lavery, Sutherland, and Wilton (1999)
	Assertiveness	Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979)
	Exercise	Raidl, Johnson, Gardiner, Denham, Spain, and Lanting (2004)
	Food management	Raidl et al.(2004)
	Food safety	Raidl et al. (2004)

	Food management	Raidl et al.(2004)
	Food safety	Raidl et al. (2004)
	Nutrition	Raidl, Johnson, Gardiner, Denham, Spain, and Lanting (2004)
	Self-expression	Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979)
Knowledge	Instructional design	Lamb and Tschillard (2005)
	Learning theory	Howard, Schmeck, and Bray (1979)
	Conditioning principles	Howard, Schmeck, et al. (1979)
	Statistics	Pohl (1982)
Skill	Communication	Skeff, Stratos, and Bergen (1992); Levinson, Gordon, and Skeff (1990)
	Estimation	Sprangers and Hoogstraten (1988); Hoogstraten (1985)
	Evaluation	Skeff, Stratos, and Bergen (1992)
	Feedback	Skeff et al. (1992)
	Innovation	Hoogstraten (1982)
	Interpersonal	Levinson, Gordon, and Skeff (1990)
	Interviewing	Howard and Dailey (1979)
	Leadership	Craig, Palus, and Rogolsky (2000)
	Teaching	Skeff, Stratos, and Bergen (1992)
	Motivating	Levinson, Gordon, and Skeff (1990)
	Management	Levinson et al. (1990)
	Counseling	Manthei (1997)

As a consequence of collecting data at one point in time, researchers (Levinson, Gordon, & Skeff, 1990; Pratt et al., 2000; Raidl et al., 2004) agreed that the retrospective pretest controls for maturity and avoids pretest sensitivity, when used in a post-then design. However, it is important to note that totally replacing the traditional pretest with the retrospective pretest (i.e., post-then design) is not recommended by all proponents of the tool (e.g., Hill & Betz, 2005; Terborg & Davis, 1982; Umble et al., 2000). Instead, these researchers strongly argued that the retrospective pretest be added to the pretest-posttest design. Without including the traditional pretest information, researchers have no means to determine whether and why a response-shift effect has occurred and are limited to reporting on pretest information collected post-intervention

(Hill & Betz, 2005; Terborg & Davis, 1982; Umble et al., 2000).

When incorporating the retrospective pretest into pretest-posttest designs, the retrospective test has been found to be robust with respect to procedural differences (Sprangers & Hoogstraten, 1989). Results indicate that within an educational training context, a 2-week time interval does not exert an influence on readministered nor on delayed retrospective preratings.

Another commonly promoted attribute of the retrospective pretest is that its gain scores correlate more highly with objective measures than traditional pretests (Howard, 1980), thereby providing a means for statistical conclusion validity. However, some studies (i.e., Mann, 1997) fail to consider the inherent assumptions in examining correlations of simple change scores. In particular, these comparisons require cautious review due to problems associated with measuring reliability (Zumbo, 1999).

The most common criticism of retrospective pretest is memory distortion and subject acquiescence. While studies (Howard, Ralph, et al., 1979; Sprangers & Hoogstraten, 1988) have investigated these threats and found the tool to be robust across a wide span of time frames and subject-style responses, the research community continually raises these as global weaknesses of the tool (Pratt et al., 2000). Further empirical work is needed to delineate the conditions under which retrospective self-report measures are inappropriate (Howard, 1980; Lam & Bengo, 2003).

Implications and Conclusions

The methodologies surrounding the retrospective pretest tool described in this paper provide important implications for evaluation and research. Notwithstanding the statistical debates concerning the detection of response-shift bias and correlates of change, the retrospective pretest extends Campbell and Stanley's (1963) experimental designs. In particular, it provides additional information for both the posttest only and traditional pretest-posttest designs (Pratt et al., 2000).

In the case of the posttest only design, the retrospective pretest provides information about the intervention not available in posttest only measures (Campbell & Stanley, 1963). In interventions where accountability is a necessity and the opportunity to benchmark pre-intervention variables has been missed, the retrospective pretest provides

the evaluator a unique opportunity to measure the impact of an intervention (Benjamin, 1982). For the traditional pretest-posttest design, the retrospective pretest provides the evaluator with data to detect response-shift bias (Pratt et al., 2000). This effect in turn can help the evaluator discover the most powerful pretest measure to use when determining program impact (Bray et al., 1984).

As with any evaluation measurement tool, the retrospective pretest can be used in an inappropriate manner (Campbell & Stanley, 1963; Hill & Betz, 2005; Howard, Ralph, et al., 1979). Therefore, research should be designed to delineate the conditions under which retrospective self-report measures should be undertaken (Howard, 1980; Lam & Bengo, 2003). For example, continuing to identify the types of knowledge, skills, and attitudes subject to response-shift bias is important to understanding the full applicability of the instrument (Umble et al., 2000). Additionally, more work is needed to test the saliency of the instrument when the time between before and after measures represents periods beyond a year (Howard, Ralph, et al., 1979). As well, further research is needed to test how variables such as social desirability, hindsight bias, and content sensitivity impact results (Lam & Bengo, 2003). Consistent with the issue of construct validity, it is important to conduct exploratory factor analyses in determining the latent structure of assessment instruments employed in post-then or pre-post-then designs (Henson, Capraro, & Capraro, 2004; Reise, Ventura, Nuechterlein, & Kim, 2005). Finally, the effect of post-then only designs should be compared to pre-post-then designs to determine whether losing the ability to detect response-shift effect outweighs the opportunity to overcome threats to validity such as pretest sensitization (Terborg & Davis, 1982).

The statistical issues in using the instrument also require further analyses. Formulas for making the decision for response-shift effect should continued to be offered, assessed, and refined (Howard, 1980). Techniques to control for experiment Type I error in pre-post-then designs should be identified and considered in the disclosure of statistical results. The assumptions and strategies for correlating change scores between self-report and performance measures should be examined in lieu of reporting change scores with inappropriate statistical techniques or without underlying statistical assumptions (Zumbo, 1999).

When considering the longevity of the tool, it seems surprising that there are so many issues regarding its application. Because its use in behavioral science is now experiencing a resurgence, today’s researchers

have a unique opportunity to expand this area of research and initiate studies that strengthen its validity and broaden its domain. Future studies should strive to model exemplary research designs surrounding its use and refine statistical procedures to maximize its power and validity. As well, future programs concerned with the measurement of behavioral variables should go beyond the aforementioned applications of the tool, recognize the retrospective pretest as a viable but imperfect instrument, and accurately report on new variables measured by the test. It is only by recognizing and improving on its weaknesses and strengths that the tool can help practitioners accurately account for treatment effect in behavioral science interventions.

References

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. Evaluation Review, 14, 374-390.

Benjamin, E. R. (1982). Using the post-then method of evaluation. Training, 19(11), 72.

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

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. Chicago: RandMcNally.

Craig, S. B., Palus, C. J., & Rogolsky, S. (2000). Measuring change retrospectively: An examination based on item response theory. In J. Martineau (Chair) Measuring behavioral change: Methodological considerations. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Deutsch, M., & Collins, M. E., (1951). Interracial housing: A psychological evaluation of a social experiment. Minneapolis: University of Minnesota Press.

Francis-Smythe, J., & Smith, P. M. (1997). The psychological impact of assessment in a development center. Human Relations, 50, 149-168.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). Educational research: An introduction (7th ed.). New York: Allyn and Bacon.

Henson, R. K., Capraro, R. M., & Capraro, M. M. (2004). Reporting practices and use of exploratory factor analyses in educational research journals: Errors and explanation. Research in the Schools, 11(2), 61-72.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. American Journal of Evaluation, 26, 501-507.

Hoogstraten, J. (1982). The retrospective pretest in an educational training context. Journal of Experimental Education, 50, 200-204.

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

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. Evaluation Review, 4, 93-106.

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., 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., & Gerber, S. K. (1979). Internal validity 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 validity in studies employing self-report instruments: A suggested remedy. Journal of Educational Measurement, 16, 129-135.

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

Lamb, T. (2005). The retrospective pretest: An imperfect but useful tool. The Evaluation Exchange11I(2). Retrieved October, 19, 2005, http://www.gse.harvard.edu/hfrp/eval/issue30/spotlight.html

Lamb, T. A., & Tschillard, R. (2005, Spring). Evaluating learning in professional development workshops: Using the retrospective pretest. Journal of Research in Professional Learning. Retrieved October 19, 20005, from http://www.nsdc.org/library/ publications/research/lamb.pdf

Levinson, W., Gordon, G., & Skeff, K. (1990). Retrospective versus actual pre-course self-assessments. Evaluation & the Health Professions, 13, 445-452.

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

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

Martineau, J. (2004). Evaluating leadership development programs: A professional guide. Greensboro, NC: Center for Creative Leadership.

Maxwell, S. E., & Delany, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

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

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

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. American Journal of Evaluation, 21, 341-349.

Raidl, M., Johnson, S., Gardiner, K., Denham, M, Spain, K., & Lanting, R. (2004). Use retrospective surveys to obtain complete data sets and measure impact in extension programs. Journal of Extension, 42(2). Retrieved October 19, 2005, from http://www.joe.org/joe/2004april/rb2.shtml

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. Journal of Personality Assessment, 84, 126-136.

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

Sears, R. R., Maccoby, E. E., & Levin, H. (1957). Patterns of child rearing. Evanston, IL: Row, Peterson.

Skeff, K. M., Stratos, G. A., & Bergen, M. R. (1992). Evaluation of a medical faculty development program: A comparison of traditional pre/ post and retrospective pre/post self-assessment ratings. Evaluation & the Health Professions, 15, 350-366.

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

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

Terborg, J. R., & Davis, G. A. (1982). Evaluation of a new method for assessing change to planned job redesign as applied to Hackman and Oldham's job characteristic model. Organizational Behavior and Human Performance, 29, 112-128.

Townsend, M., Lai, M. K., Lavery, L., Sutherland, C., & Wilton, K. (1999, December). Mathematic anxiety and self-concept: Evaluation change using the "then-now" procedure. Paper presented at the combined meeting of the Australian Association for Research in Education, Melbourne, Australia.

Umble, K., Upshaw, V., Orton, S., & Matthews, K. (2000, June). Using the post-then method to access learner change. Presentation at the AAHE Assessment Conference, Charlotte, NC.

Walk, R. D. (1956). Self-ratings of fear in a fear-invoking situation. Journal of Abnormal and Social Psychology 52, 171-178.

Zumbo, B. (1999). The simple difference score as an inherently poor

measure of change: Some reality, much mythology. In B. Thompson (Ed.), Advances in social science methodology (pp. 269-304). Stanford, CT: JAI Press Inc.

Zwiebel, A. (1987). Changing educational counselors' attitudes toward mental retardation: Comparison of two measurement techniques. International Journal of Rehabilitation Research, 10, 383-389.

Kim Nimon, Graduate Assistant, and Jeff Allen, Associate Professor, are in the Technology and Cognition Department at University of North Texas in Denton, Texas.