Survey Research Measurement Issues in Evaluating Change: A Laboratory Investigation

Achilles A. Armenakis
Auburn University

M. Ronald Buckley
Washington State University

Arthur G. Bedeian
Louisiana State University

Efforts to operationalize the alpha/beta/gamma change typology have suffered from a notable limitation. Virtually all have been conducted in field settings, thereby limiting the degree of experimental control over outcome criteria. Recognizing this limitation, the present study employed a laboratory methodology to investigate two research questions related to scale recalibration (beta change) in temporal survey research. Application of this methodology permitted random respondent assignment, exact replication of stimuli, and systematic time interval variation for the pretest-posttest design. Furthermore, the use of these procedures permitted testing the use of the retrospective design in assessing organizational change. Implications of the findings for the measurement of change are discussed.

In considering self-report measures, substantial progress in evaluating change has been made over the last two decades (cf. Bennis, 1965; Lupton, 1965; Sofer, 1964). Of particular impact has been Golembiewski, Billingsley, and Yeager's (1976) introduction of a change typology. This typology distinguishes between three forms of change: alpha, beta, and gamma.

Gamma change refers to the reconceptualization or redefinition of a referent variable. It occurs when people change their basic understanding, from one testing period to another, of the criterion being measured. For example, peer leadership may mean something quite different at Time 1 than at Time 2, especially if a planned treatment or intervention was directed at enhancing a person's understanding of this or other related concepts.

Beta change occurs when the standard of measurement used by a person to assess a stimulus changes from one testing period to another. Such change indicates a recalibration of the person's internalized measurement scale. Thus, a person may rate a certain leader behavior a *2* (on a Likert scale) at Time 1 and the identical behavior as a *3* at Time 2.

Alpha change takes place when a change is detected on a measurement scale for which gamma and beta changes have been ruled out. That is, after determining that neither gamma change nor beta change has occurred, if a researcher observes a difference in a person's responses from Time 1 to Time 2, alpha change can be said to have been detected.

Concept redefinition (gamma change) and scale recalibration (beta change) may be legitimate goals, depending on the objectives of a change intervention. For instance, scale recalibration is a common objective of programs to train administrators in implementing performance evaluation systems. Such group training is generally aimed at enabling evaluators to recognize specific behaviors and to consistently evaluate those behaviors as a basis for equitably administering salary adjustments. It is those instances in which either (or both) gamma

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 2, June 1986, pp. 147–157
© Copyright 1986 Applied Psychological Measurement Inc.
0146-6216/86/020147-11$1.80

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

change or beta change unintentionally occurs that are especially problematic.

An analysis of the literature dealing with efforts to operationalize the alpha/beta/gamma change typology (Armenakis & Zmud, 1979; Bedeian, Armenakis, & Gibson, 1980; Golembiewski et al., 1976; Randolph, 1982; Schmitt, 1982; Terborg, Howard, & Maxwell, 1980) discloses that virtually all have been conducted in field settings. Consequently, experimental control over outcome criteria has been limited. That is, field attempts to measure either scale recalibration or concept redefinition have been limited by the conditions under which stimuli have been observed. Only through establishing highly controlled conditions can the influence of potentially confounding variables be minimized.

The purpose of the present study was to employ a laboratory methodology to deal with the above shortcoming as it relates to scale recalibration in temporal survey research. Laboratory methods are excellent when honing concepts and refining measurement techniques, because they allow the systematic variation of independent variables. As a result, conclusive answers can often be obtained and relatively precise and subtle theoretical points can be tested.

Research Questions

Two research questions were selected for investigation. Both are pertinent to the issue of evaluating change (Armenakis, Bedeian, & Pond, 1983). The first question concerns the appropriate measurement interval for temporal survey research involving pretest and posttest designs. The second relates to the use of retrospective designs.

In evaluating organizational change programs, researchers have typically used designs that prescribe collection of data prior to and subsequent to an intervention. Examples of single-group designs are the one-group pretest-posttest and the time series designs. Multiple-group designs have been used, although to a lesser extent. Examples include the pretest-posttest control group design and the nonequivalent control group design (cf. Cook & Campbell, 1979).

Regardless of which design is used, little guidance can be found for specifying suitable measurement intervals. Previous investigators engaged in evaluating change interventions have yet to establish guidelines for determining appropriate measurement intervals for temporal survey research designs. Consequently, Porras and Berg (1978) have recommended that investigators increase (1) the length of time devoted to collecting data on change and (2) the frequency with which data are collected. To contribute to a better understanding of the measurement of change, there is a need for more published research regarding the time dimension. Thus, one purpose of the present study was to explore the research question, "Does the time interval between survey administrations for pretest-posttest designs contribute to scale recalibration?"

The use of retrospective designs is by no means a new methodology in the study of organizations. Huber and Power (1985), for example, have studied the use of retrospective recollections in strategic management research. Terborg et al. (1980) have proposed the use of retrospective designs as an effective means of isolating scale recalibration in temporal survey research. Terborg et al.'s findings, however, are in general contrary to other published results on recall as a data collection method. Other researchers have concluded that recall is generally an unreliable data collection technique (cf. Bower, Black, & Turner, 1979; Green & Wright, 1979; Rippey, Geller, & King, 1978).

A subtle difference, however, exists between the work of Terborg et al. and other researchers. In their work to date, Terborg and his associates requested that respondents assess their *own* competencies after participating in skill training. By contrast, the intent of other research on the recall method has been to have respondents recall their own behaviors or those of others. It is conceivable that this subtle point accounts for the differences in Terborg et al.'s results as contrasted with those of other researchers. This possibility raises an important question for evaluation research: "What types of change interventions should be evaluated using retrospective designs?" Thus, the second

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

purpose of the present study was to investigate the appropriateness of retrospective designs in evaluating organizational change.

Method

As argued, to determine the impact of conditions affecting temporal survey responses, it is imperative that the stimulus being rated be tightly controlled. Simply stated, without knowing whether or not a target stimulus has or has not changed, a researcher cannot discount scale recalibration as a potential explanation for response variations. To achieve the necessary control, videotape vignettes were utilized in the present study. The vignettes were selected from seven scripts depicting a personnel manager interviewing a disgruntled employee (Borman, Hough, & Dunnette, 1976). Respondents were asked to evaluate seven dimensions of interviewer performance as displayed on various vignettes following a predetermined scheme. The dimensions which served as the study's dependent variables were

1. Interview structure and control (Structure),
2. Establishment and maintenance of rapport (Rapport),
3. Reaction to stress (Stress),
4. Information acquisition,
5. Conflict resolution (Conflict),
6. Subordinate development (Develop), and
7. Subordinate motivation (Motivate).

An overall dimension was formed by treating all items as a single dimension.

A 39-item survey instrument was developed from the manual accompanying the vignettes (Borman et al., 1976). This instrument, titled "Dimensions of Interviewer Performance" (DIP), incorporated a five-point response format. To minimize artificial inflation of reliability estimates, item placement was systematically varied throughout the questionnaire across all seven dimensions of interviewer performance. As indicated, respondents were requested to evaluate the extent to which a tape-specific activity was performed by the interviewer (personnel manager) in question. A sample item is provided in Figure 1.

Procedure

Undergraduate students enrolled in introductory management and psychology courses were offered an opportunity to participate in a research study for extra credit. These particular courses were deemed an appropriate source for participants since neither is concerned with interviewing, but rather with general management and psychology. All students were informed of the study at the beginning of the academic term. They were told that they would be expected to view a videotape twice during the term and respond to a series of corresponding questions. Sessions were estimated to last about 30 minutes and were scheduled at 1-hour intervals.

All participants received a research booklet containing (1) a consent form, (2) a demographics sheet, (3) a brief vignette introduction, and (4) the survey instrument. Participants were assigned individual code numbers. After signing the consent form and completing the demographics sheet, respondents were asked to read the vignette introduction silently while a laboratory coordinator read the introduction aloud:

You are about to view a videotape of the initial

Figure 1
Sample Item from DIP Questionnaire

	To a very little extent	To a little extent	To some extent	To a great extent	To a very great extent
The extent to which the manager stated the purpose of the interview.	1	2	3	4	5

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

meeting between a manager (Dave Baxter) and a subordinate (Marshall Whipker). Marshall Whipker is a 9-year employee of the company. He has demonstrated a great degree of technical knowledge but has been passed over for regular promotions, probably due to his lack of managerial skills. Dave Baxter is a newly appointed manager in the company. He has a desire to meet all of his subordinates, hence, this videotaped meeting. Please observe the tape carefully as you will be asked some questions about the meeting between Baxter and Whipker.

Vignette. For the present study, an average performance vignette was selected as the stimulus for testing. In this vignette, the interviewer's performance was designed to be neither excellent nor deficient. Judgments of interviewer performance were obtained from the manual accompanying the vignettes.

Research groups. A total of 400 students, comprising eight groups, participated in the study. Shown in Table 1 are the group sizes, vignettes viewed, time interval between viewings, and the experimental designs employed with each research group. Groups 1 through 3 participated in a "no-treatment pretest-posttest design," that is, all respondents viewed the average performance vignette at Time 1 and Time 2. The time intervals for Group 1 ($n = 36$), Group 2 ($n = 36$), and Group 3 ($n = 74$) were 4, 8, and 3 weeks, respectively (see Table 1).

Groups 4 through 8 participated in a retrospective design incorporating either a 2- or 3-week time interval. The time interval for Group 8 ($n = 42$) was 2 weeks while that for Group 4 ($n = 74$), Group 5 ($n = 66$), Group 6 ($n = 74$), and Group 7 ($n = 72$) was 3 weeks. Group 4 was formed by asking members of Group 3 to recall their Time 1 responses before viewing their scheduled second vignette (see Table 1).

Data Analysis

Due to a small subjects-to-items ratio, a factor analysis of DIP responses was not attempted. Only 146 students were included in Groups 1, 2, and 3. The remaining 254 students were included in Groups 4 through 8.

Intercorrelations of scores on the seven dimensions of interviewer performance were computed. For Time 1, the correlations ranged from .48 to .75, with a median of .60. For Time 2, the correlations ranged from .58 to .71, with a median of .62. Given the magnitude of these correlations, an Overall score was included in the analyses. Recognizing the conceptual distinctness of the seven dimension scores, however, results are presented both on the overall and the individual dimensions. Coefficient alpha estimates (Cronbach, 1951) were computed to determine the reliabilities of each dimension. Mean scores, obtained by computing across items for each dimension, were compared using dependent *t*-tests (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

Results

Results of the data analyses for the eight respondent groups are presented in Tables 2 and 3, and provide several insights into each of the two research questions being investigated. Internal consistency reliabilities for the eight dimensions of interviewer performance were quite high (see Tables 2 and 3) for all eight groups; 80% of the alphas exceeded .70. Evaluations of interviewer performance were equally consistent (reliable) among respondents in Groups 1, 2, and 3 (see Table 2). For Groups 4 through 8, coefficient alphas for the eight dimensions of interviewer performance were acceptable for both the Time 1 and the retrospective DIP administrations (see Table 3).

Time Interval Between Stimuli

Results of dependent *t*-tests for Groups 1, 2, and 3 revealed only two significant pretest-posttest differences ($p < .05$; see Table 2) in evaluated interviewer performance. These differences are probably due to chance occurrence (Feild & Armenakis, 1974). Therefore, it is arguable that measurement intervals of 3, 4, and 8 weeks are not associated with scale recalibration in the present study.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Group Sizes, Vignettes Viewed, Time Interval Between Viewings and Experimental Designs Employed with each Research Group

Group	n	Vignette		Time Interval	Design
		Time-1	Time-2		
1	36	Aver. Perf. Vignette	Aver. Perf. Vignette	4 weeks	O_1O_2
2	36	Aver. Perf. Vignette	Aver. Perf. Vignette	8 weeks	O_1O_2
3	74	Aver. Perf. Vignette	Aver. Perf. Vignette	3 weeks	O_1O_2
4[a]	—[b]	Aver. Perf. Vignette	None	3 weeks	O_1O_R
5[a]	66	Aver. Perf. Vignette	None	3 weeks	O_1O_R
6[a]	74	Aver. Perf. Vignette	None	3 weeks	O_1O_R
7[a]	72	Aver. Perf. Vignette	None	3 weeks	O_1O_R
8[a]	42	Aver. Perf. Vignette	None	2 weeks	O_1O_R

[a]Groups 4 through 8 were the retrospective design research groups. At either a two- or a three-week time interval these groups recalled, designated as O_R, their earlier response at O_1 without viewing a second vignette.

[b]Group 4 was formed by asking respondents in Group 3 to recall their time-1 responses before viewing their scheduled second vignette.

Retrospective Design

Respondents in Groups 4 through 8, utilizing retrospective designs, were required to recall their Time 1 evaluations of interviewer performance (i.e., the average performance vignette) at either a 2- or 3-week interval. Correlations between Time 1 and retrospective DIP responses ranged from .56 to .89. These findings suggest that (1) respondents were consistent in their *ability* to recall their earlier Time 1 evaluations of interviewer performance, and (2) variations between retrospective and Time 1 evaluations of interviewer performance were *systematic* (as opposed to random) for all respondents. At the same time, of 40 dependent *t*-tests on Time 1 and retrospective evaluations of interviewer performance, 24 yielded statistically significant values (all $ps < .05$). Furthermore, it can be observed that there was a general tendency for respondents to recall their earlier ratings as being less extreme

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Dimensions of Interviewer Performance, Number of Items per Dimension, Internal Consistency Reliabilities (α) Pearson-Product-Moment-Correlations (r), and Dependent t-values for Group 1, Group 2, and Group 3

Dimension	Number of Items	Group 1: Four-Week Interval (n=36)			Group 2: Eight-Week Interval (n=36)			Group 3: Three-Week Interval (n=74)		
		α	r	t-value	α	r	t-value	α	r	t-value
$Structure_1$	6	.72	.87**	-2.00*	.85	.65**	-1.70	.64	.72**	-1.44
$Structure_2$		.71			.71			.73		
$Rapport_1$	7	.84	.80**	-.10	.76	.63**	-.34	.83	.76**	-1.64
$Rapport_2$		.80			.85			.85		
$Stress_1$	3	.71	.59**	1.32	.38	.74**	-.35	.79	.59**	1.61
$Stress_2$		.63			.51			.75		
Information $Acquisition_1$	7	.84	.74**	-.63	.81	.81**	-.58	.85	.73**	-1.09
Information $Acquisition_2$		.84			.84			.86		
$Conflict_1$	5	.74	.69**	-.42	.76	.81**	-.06	.74	.79**	-2.09*
$Conflict_2$		.76			.81			.87		
$Develop_1$	5	.87	.64**	-.73	.81	.58**	-.08	.81	.63**	-.51
Develop 2		.82			.81			.90		
$Motivate_1$	6	.81	.65**	.09	.82	.71**	-1.29	.80	.78**	-1.63
$Motivate_2$		.81			.85			.87		
$Overall_1$	39	.93	.81**	-.39	.91	.82**	-1.04	.93	.80**	-1.20
$Overall_2$		.92			.88			.92		

Note. Dimension label subscripts represent pretest (1) and posttest (2) conditions.

[a] rs represent Pearson-product-moment correlations between responses for time-1 and time-2 Dimensions of Interviewer Performance Questionnaire responses.

*$p < .05$
**$p < .001$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

(i.e., more positive), as evidenced by the negative direction for all dimensions except Stress.

Of the five retrospective groups, the recall interval for four of the groups (i.e., Groups 4, 5, 6, and 7) was 3 weeks, while the recall interval for Group 8 was 2 weeks. Of the seven individual dimensions for Group 8, three were significantly different. However, when all items were collapsed

Table 3
Dimensions of Interviewer Performance, Number of Items per Dimension, Internal Consistency Reliabilities, (α), Pearson-Product-Moment Correlations (r) and Dependent t-value for Retrospective Design Groups

Retrospective Group, Time Interval, n, and Dimension[a]	α Time-1	α Retrospective	r[b]	Dependent t-value
Group 4				
Three-week interval				
(n=74)				
Structure	.64	.75	.74**	-3.20*
Rapport	.83	.86	.85**	-1.66
Stress	.79	.74	.73**	1.58
Information Acquisition	.85	.87	.76**	-1.48
Conflict	.74	.82	.85**	-5.61**
Develop	.81	.87	.71**	-1.55
Motivate	.80	.86	.76**	-3.34*
Overall	.93	.92	.86**	-3.08*
Group 5				
Three-week interval				
(n=66)				
Structure	.64	.73	.76**	-2.12*
Rapport	.79	.84	.83**	-1.76
Stress	.64	.73	.72**	2.66*
Information Aquisition	.81	.79	.75**	-3.23*
Conflict	.74	.85	.81**	-5.07**
Develop	.76	.81	.78**	-1.65
Motivate	.85	.84	.81**	-3.60**
Overall	.92	.93	.89**	-3.55**
Group 6				
Three-week interval				
(n=74)				
Structure	.59	.64	.71**	-3.70**
Rapport	.69	.81	.72**	-2.59*
Stress	.58	.61	.56**	3.14*
Information Acquisition	.77	.83	.65**	-2.02*
Conflict	.72	.76	.68**	-5.16**
Develop	.82	.86	.69**	-1.73
Motivate	.82	.86	.73**	-4.53**
Overall	.86	.89	.76**	-3.66**

-continued on the next page-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3, continued
Dimensions of Interviewer Performance, Number of Items per Dimension, Internal Consistency Reliabilities, (α), Pearson-Product-Moment Correlations (r) and Dependent t-value for Retrospective Design Groups

Retrospective Group, Time Interval, n, and Dimension[a]	α Time-1	α Retrospective	r[b]	Dependent t-value
Group 7 Three-week interval (n=72)				
Structure	.73	.68	.66**	-.31
Rapport	.79	.78	.77**	-.44
Stress	.58	.69	.56**	1.10
Information Acquisition	.77	.85	.71**	-2.34*
Conflict	.60	.67	.67**	-4.54**
Develop	.78	.76	.70**	-.66
Motivate	.75	.84	.76**	-3.43**
Overall	.89	.91	.82**	-2.54*
Group 8 Two-week interval (n=42)				
Structure	.79	.78	.72**	-.16
Rapport	.82	.86	.77**	.71
Stress	.57	.87	.65**	2.01*
Information Acquisition	.78	.88	.68**	.89
Conflict	.78	.82	.65**	-2.36*
Develop	.80	.87	.80**	.64
Motivate	.78	.84	.66**	-2.18*
Overall	.90	.91	.78**	-.32

[a]The number of items comprising each dimension is as follows: Structure = 6; Rapport = 7; Stress = 3; Information Acquisition = 7; Conflict = 5; Develop = 5; Motivate = 6; Overall = 39.

[b]rs represent Pearson-product-moment correlations between responses at time-1 and retrospective administrations of the Dimensions of Interviewer Performance questionnaire responses.

*$p < .05$

**$p < .001$

into the Overall dimension, the difference between the Time 1 and retrospective responses for Group 8 was nonsignificant. Two explanations may be offered. The first is that, because the recall interval for Group 8 was only 2 weeks, the respondents may not have experienced the memory loss of the other groups. The second and more plausible explanation is that the direction of the mean difference in the Time 1 and recall responses of Group 8 for Rapport, Stress, Information Acquisition, and

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Develop was positive while the direction of Structure, Conflict, and Motivate was negative. Unlike Groups 4–7, the mean direction for Structure, Information Acquisition, and Develop was positive. Therefore, the effect of aggregating the seven dimensions into the Overall dimension for Group 8 resulted in a nonsignificant difference. Across all five retrospective groups, these findings suggest that the respondents' ability to recall their Time 1 evaluations of interviewer performance was systematically poor.

Discussion

The present study, conducted in a laboratory setting using videotape technology, allowed random respondent assignment, exact stimuli replication, and systematic time interval variation (i.e., 3, 4, and 8 weeks) for pretest-posttest designs. Furthermore, the use of these procedures permitted testing the ability of respondents to use the retrospective design in assessing the behavior of others.

Time Interval Between Stimuli

A major conclusion drawn from the results presented is that time interval does not contribute to scale recalibration in pretest-posttest designs. This conclusion carries an important implication for the measurement of change using temporal survey techniques. Since 1976, various methods for detecting scale recalibration and concept redefinition over time have been demonstrated (e.g., Armenakis & Zmud, 1979; Randolph, 1982; Schmitt, 1982; Terborg et al., 1980; van de Vliert, Huismans, & Stok, 1985). Several theorists have proposed methods of eliminating or correcting for these concerns (Terborg et al., 1980; Bedeian et al., 1980). However, as argued by Armenakis et al. (1983), in order to establish internal and external validity, there is a need for survey researchers to investigate the causes of scale recalibration and concept redefinition before recommending further solutions.

The need for investigation of these causes is increasing due to the growing prevalence of longitudinal change research. In offering a guideline for survey researchers to follow, Arundale (1980) specified two conditions that should be met. First, there should be at least three temporal observations. Second, the time interval (i.e., frequency with which observations are made) and the measurement span (i.e., duration of time for which observations are to continue) should be predetermined. To date, survey researchers have yet to develop measurement span or time interval guidelines. Before either can be established, the relationship between time and scale recalibration must be investigated if the error component in change measurement is to be minimized.

Retrospective Design

The reported results also demonstrate an important lesson pertinent to respondent capabilities in measuring change. Studies conducted by Terborg et al. (1980) employed a retrospective design in situations where participants could accurately recall earlier states. These studies employed experimental and comparison groups. Use of the retrospective design in the present study is equivalent to a comparison group (or no-treatment control group) in a "nonequivalent control group retrospective design." Terborg et al.'s findings suggest that the comparison groups in the present study should not have evidenced change. Because the reported findings from the retrospective portion of the present study do in fact show change, an explanation is required. In this respect, Terborg et al. assert that "there is no reason to suspect Then scores except in situations where it is to the participant's advantage to give false Then responses, where participants are confused as to the instructions, or where participants in a no-treatment control group are asked to give Post and Then ratings within a few hours or days of the Pre ratings" (p. 114).

In the present investigation there was no motive for respondents to provide false evaluations, and there was no indication that respondents misunderstood the relevant instructions. Additionally, the calculated reliability estimates and correlation coefficients are too large to permit such an interpretation. The experimental conditions were such that it would be highly unlikely that either a motive

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

for false evaluations or a misunderstanding of instructions would be so consistent across respondents.

A third possible reason to suspect the retrospective ratings involves the question of whether a 2- or 3-week interval is long enough to ignore Terborg et al.'s warning of "within a few hours or days of the Pre ratings." The answer here is not so obvious. The results associated with Groups 4 through 8, however, provide a clue. First, it was concluded that the time interval between the two survey administrations was not a major concern. The logic for the pretest-posttest design was to allow respondents to observe a vignette and respond to the administered survey, wait a designated time interval (either 3, 4, or 8 weeks), and then observe the vignette a second time and respond once again. It was argued that the designated time intervals would allow for various sources of contamination (Cook & Campbell, 1979) to affect respondents' abilities to articulate their evaluations of interviewer performance at the second viewings. It was concluded that respondents could articulate their evaluations of interviewer performance equally well at both Time 1 and Time 2, regardless of time interval. Together these findings lead to the conclusion that *memory* is a plausible explanation for the reported results.

Another concern is closely related to this point. If respondents in an experimental group (i.e., a treatment group) evaluate their own competencies in a retrospective design, it may be possible for an intervention to contribute to scale recalibration. In such instances, it is arguable whether or not respondents can acceptably evaluate (recall) their previous competencies. Rather than relying on memory, they may be evaluating (recalling) from a different (recalibrated) perspective. However, if respondents are part of a control group (i.e., a no-treatment group), it would be difficult to argue that memory is *not* important in retrospective designs. Because control group respondents experience no treatment, they must recall previous questionnaire responses. It has been demonstrated repeatedly that memory cannot be relied on to produce accurate recollections (Cherry & Rodgers, 1979; Green & Wright, 1979; Heneman & Wexley, 1983; Rippey et al., 1978).

It should be apparent from these findings that the retrospective design might be applicable for randomized assignment of persons to treatment and control groups where the statistical analysis is conducted across groups rather than within groups. For example, if persons are assigned randomly to treatment and control groups, then it is expected that memory bias will be equally distributed between the groups. Then, by comparing recall of the controls to recall of the experimentals, the memory bias will be equal and differences detected will be analyzed in terms of the sources of invalidity (cf. Cook & Campbell, 1979). However, the use of the retrospective design is not recommended for use with single-group designs, for example, the one-group pretest-posttest design. Furthermore, the recall procedure is not recommended with any design in which the responses are analyzed within the group or where mathematical transformations are performed between pretests and retrospective pretests and between posttests and retrospective pretests.

If the retrospective design is to be employed in evaluating change interventions, requesting respondents to judge the behavior of other individuals who have participated in a change program is analogous to asking respondents to recall their own earlier responses (i.e., to rely on their memory). It has been demonstrated in the present study, and elsewhere, that memory produces inaccurate recollections. For this reason, the retrospective design is not recommended for evaluating organizational interventions.

Generalizability of Findings

The 3-week time interval for the retrospective design was intended to coincide with the shortest time interval selected for the pretest-posttest designs. Even though a longer time interval was not tested, it is expected that the results would be similar. The results for the 2-week time period were found to be less extreme than the 3-week interval, but still supported the overall findings that the retrospective design is not recommended for evaluating organizational change. It is not known what effects a shorter time interval might have had upon respondents' ability to recall earlier states. The rationale for the intervals selected was that the pur-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

pose of the study was to test these designs in situations appropriate for evaluating organizational change programs, which typically span weeks and months.

Because this study used the laboratory method, the generalizability of the findings to field settings might be questioned. In defense of laboratory methodology, two points should be stressed. First, it would have been difficult, if not impossible, to conduct this study in a field setting and maintain the degree of control achieved here. Second, the tasks required of the respondents were not unrealistic and were similar to those required of persons in a field setting.

References

Armenakis, A., Bedeian, A., & Pond, S. (1983). Research issues in OD evaluation: Past, present & future. *Academy of Management Review, 8,* 320–328.

Armenakis, A., & Zmud, R. (1979). Interpreting the measurement of change in organizational research. *Personnel Psychology, 32,* 709–723.

Arundale, R. (1980). Studying change over time: Criteria sampling from continuous variables. *Communication Research, 1,* 227–263.

Bedeian, A., Armenakis, A., & Gibson, R. (1980). On the measurement and control of beta change. *Academy of Management Review, 5,* 561–566.

Bennis, W. (1965). Theory and method in applying behavioral science to planned organizational change. *Journal of Applied Behavioral Science, 1,* 337–360.

Borman, W., Hough, L., & Dunnette, M. (1976). *Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error* (vol. 1). Minneapolis: Personnel Decisions, Inc.

Bower, G., Black, J., & Turner, T. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177–220.

Cherry, N., & Rodgers, B. (1979). Using a longitudinal study to assess the quality of retrospective data. In L. Moss & H. Goldstein (Eds.), *The recall method in social surveys* (pp. 31–42). London: University of London, Institute of Education.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Feild, H., & Armenakis, A. (1974). On the use of multiple tests of significance in psychological research. *Psychological Reports, 35,* 427–431.

Golembiewski, R., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133–157.

Green, G., & Wright, J. (1979). The retrospective approach to collecting baseline data. *Social Work Research & Abstracts, 15*(3), 25–30.

Heneman, R., & Wexley, K. (1983). The effects of time delay in rating and amount of information observed on performance rating accuracy. *Academy of Management Journal, 26,* 677–686.

Huber, G., & Power, D. (1985). Retrospective reports of strategic-level managers: Guidelines for increasing their accuracy. *Strategic Management Journal, 6,* 171–180.

Lupton, T. (1965). The practical analysis of change in organizations. *Journal of Management Studies, 2,* 218–227.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., & Bent, D. (1975). *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill.

Porras, J., & Berg, P. (1978). Evaluation methodology in organization development: An analysis and critique. *Journal of Applied Behavioral Science, 14,* 151–173.

Randolph, W. (1982). Planned organizational change and its measurement. *Personnel Psychology, 35,* 117–139.

Rippey, R., Geller, L., & King, D. (1978). Retrospective pretesting in the cognitive domain. *Evaluation Quarterly, 2,* 481–491.

Schmitt, N. (1982). The use and analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research, 17,* 343–358.

Sofer, C. (1964). The assessment of organizational change. *Journal of Management Studies, 1,* 128–142.

Terborg, J., Howard, G., & Maxwell, S. (1980). Evaluating planned organizational change: A method for assessing alpha, beta and gamma change. *Academy of Management Review, 5,* 109–121.

van de Vliert, E., Huismans, S., & Stok, J. (1985). The criterion approach to unraveling beta and alpha change. *Academy of Management Review, 10,* 269–275.

Author's Address

Send requests for reprints or further information to Achilles A. Armenakis, Department of Management, Auburn University, AL 36849-3501.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/