©Academy of Management Review, 1983, Vol. 8, No. 2, 320-328.

Research Issues in OD Evaluation: Past, Present, and Future

ACHILLES A. ARMENAKIS
ARTHUR G. BEDEIAN
Auburn University
SAMUEL B. POND, III
North Carolina State University

Early OD evaluation concerns at the macrolevel dealt principally with research design complexities. Current microlevel research is concerned largely with technical issues related to the accurate measurement of change. It is suggested that future evaluation research should continue this macro- to micro- evolution by (1) reconciling certain current differences and testing new methodologies; (2) drawing on studies dealing with time-order error; and (3) investigating questions of "time interval" and "measurement span" associated with longitudinal research designs.

It has been some 20 years since attention first began to focus on both implementation *and* evaluation of organization development (OD) programs (Blake, Mouton, Barnes, & Greiner, 1964; Harrison, 1962; Shephard, 1960). An early contributor to this area, Bennis (1965) emphasized the importance of evaluating the efforts of change agents. He especially stressed that change agents should devote as much effort in the evaluation as in the implementation of change programs. In retrospect, research on OD evaluation has moved through three distinct but overlapping phases: (1) identification of general evaluation problems and development of evaluation guidelines; (2) demonstrations of methods to deal with commonly encountered evaluation problems; and (3) resolution of specific methodological problems common to evaluation efforts.

The purpose of the present manuscript is to review briefly these phases to highlight progress of the last two decades and suggest several advancements that will contribute to the continued maturation of OD as a discipline.

Phase 1: General Problems and Guidelines

Perhaps the first attempt to identify problems associated with evaluation and to develop guidelines for their resolution in an OD context was that of Harrison (1971). Drawing on personal experience, Harrison identified eight problem areas and suggested accompanying guidelines to follow in evaluating OD programs. The problems identified were: (1) difficulties in using control groups; (2) insufficient longitudinal research *after* an intervention; (3) limitations of research designs that restrict the measurement of change; (4) inadequate schema for classifying training outcomes; (5) lack of standardization in training experiences; (6) improper timing for the collection of pretest data; (7) difficulties in eliminating the influence of experimenter-participant relationships in laboratory settings; and (8) statistical difficulties associated with measuring change.

A similar attempt to identify problems facing OD practitioners capitalized on the experiences of a large number of change agents. In two papers, Armenakis and his colleagues (Armenakis, Feild, & Holley, 1976; Armenakis, Feild, & Mosley, 1975) identified the evaluation practices and problems of organization development consultants through the use of a mail questionnaire. Guidelines then were developed for conducting evaluations. Subject areas covered were: (1) selection and measurement of "soft" criteria; (2) use of comparison groups; (3) control of extraneous influences; (4) development and use of "hard"

criteria; (5) coping with time lags (i.e., time that elapses between changes in soft criteria and concomitant changes in hard criteria); and (6) commitment of resources to OD evaluation efforts.

A third study indicative of this phase was conducted by Nicholas (1979). He specifically identified problems arising from failure to plan evaluation adequately. Among the problems discussed, together with guidelines for their resolution, were: (1) vaguely defined objectives; (2) inadequately developed theoretical models; (3) omission of key decision makers in the design of an evaluation effort; (4) failure to utilize multiple methods in measurement of criteria; (5) reliance on unreliable criteria; (6) inability to rule out rival hypotheses; and (7) failure to distinguish between statistical and practical significance in detection of criteria differences.

The above studies are similar in two respects and different in a third. One similarity lies in the types of problems presented. In each case, the problems identified are at a macrolevel—problems that consultants would encounter during design and execution of an evaluation. Their second similarity is that the authors of each study attempted to develop guidelines that change agents could follow, within limitations imposed by field settings, in evaluating the success of their efforts.

The studies are noticeably different, however, in that their methodologies were quite disparate. Harrison relied predominantly on his own experience as a consultant. The studies by Armenakis and his colleagues empirically identified evaluation problems through surveying practicing change agents. Finally, Nicholas principally surveyed the literature on evaluation research. It is significant, however, that despite differing methodologies, there is a surprising convergence among the findings of these studies.

Phase 2: Demonstration of Methods

Studies that are characteristic of Phase 2 demonstrate specific methods used in evaluating OD efforts. The value of these studies is that they explain an aspect of research design (e.g., statistical technique) or a quasi-experimental design using data collected in a field setting. In learning from these researchers, scientific rigor of future investigations can be enhanced. Scientific rigor in evaluation research is determined largely by four factors: (1) type of experimental/quasi-experimental design; (2) selection and operationalization of criteria; (3) statistics employed; and (4) manner in which extraneous variables can be systematically discounted. The importance of increasing scientific rigor is to be able to discount rival hypotheses that may influence findings of an OD effort. Change agents then may more confidently refine or discard interventions that do not produce desired results, employing only those interventions that are successful.

Numerous demonstration studies have appeared over the last 15 years. However, for present purposes, six investigations, representative of those appearing in the literature, were selected and have been summarized in Table 1. These are: (1) Miles (1965), (2) Friedlander (1967), (3) Golembiewski and Carrigan (1970), (4) Harvey and Boettger (1971), (5) Armenakis and Feild (1975), and (6) Evans (1975).

From the standpoint of advancing the state of OD knowledge, each of these studies made a unique methodological contribution. For example, several demonstrated the use of various experimental and quasi-experimental designs. Miles (1965) employed a Solomon four-group design and explained the possibility of test-treatment interaction as a plausible rival hypothesis threatening external validity. Friedlander (1967) used a nonequivalent control group design and explained the importance of establishing group equivalence when subjects are not randomly assigned to experimental treatments. Golembiewski and Carrigan (1970), who utilitzed a modified time series design, were the first OD researchers to explain, systematically, rival hypotheses affecting the internal validity of an OD intervention. Evans (1975) explained how a researcher could patch up a weak design as an investigation progressed.

Use of criteria to evaluate effects of OD interventions is equally noteworthy. Miles (1965) relied on a number of prefabricated instruments for evaluating change. Friedlander (1967) explained how to tailor questionnaires for use with a specific organization, determine empirical dimensions of organizational behavior using factor analysis, and compute measures of test-retest reliability for a tailored instrument.

Utilization of existing hard criteria in a time series design was explained by Armenakis and Feild (1975). The major contribution of this study was the recognition that certain criteria (e.g., productivity) may reflect an increasing or decreasing trend indicative of autocorrelation. Hence, if statistical tests are not

Table
Selected OD

Authors/ Date	*Type of Design*	*Configuration of Design*[a]	*Sample Size*[b]
Miles (1965)	Solomon four-group design	1. $O_1\ X\ O_2$ 2. $O_3\quad O_4$ 3. $\quad X\ O_5$ 4. $\quad\quad O_6$	1. $n_X = 17$ 2. $n_C = 17$ 3. $n_X = 17$ 4. $n_C = 17$
Friedlander (1967)	Nonequivalent control group design	$O_1\ X\ O_2$ $O_3 \quad O_4$	1. $n_X = 31$ 2. $n_C = 60$
Golembiewski & Carrigan (1970)	Modified time series design	$O_1\ X\ O_2 O_3$	$n_X = 16$
Harvey & Boettger (1971)	1. Modified time series design 2. One group pretest posttest design	1. $O_1\ X\ O_2 O_3$ 2. $O_1\ X\ O_2$	1. $n_X = 15$ 2. $n_X = 15$
Armenakis & Feild (1975)	1. Time series design	1. $O_1 O_2 O_3 \ldots O_7$ $\quad\quad X$ $O_8 O_9 O_{10} \ldots O_{13}$	Not reported in Huse & Beer (1971)
Evans (1975)	1. Static comparison group design 2. After-only longitudinal design with a comparison group 3. Abbreviated time series design	1. $X\ O_1$ $\quad O_2$ 2. $X\ O_1\ O_2 O_3 O_4 O_5 O_6$ 3. $O_1 O_2\ X\ O_3 O_4$	1. Not reported 2. Not reported 3. Not reported

[a] O represents the time period of data collection; X represents the OD intervention.
[b] n_X and n_C represent the notations for the size of the experimental and comparison groups, respectively.
[c] LBDQ and GPS are the acronyms for the Leader Behavior Description Questionnaire and the Group Participation Scale, respectively.

modified to compensate for such trends, conventional inferential statistics (e.g., ANOVA) will not detect autocorrelation and consequently will render inappropriate conclusions. Armenakis and Feild provided a statistical procedure to deal with this situation.

For those OD programs for which unobtrusive criteria (e.g., absenteeism) are not readily available from archival records, a change agent may be forced to improvise. As an example, Friedlander (1967) showed

1
Evaluation Efforts

Criteria[c]	*Statistics*	*Methodological Contribution(s)*
1. LBDQ 2. GPS 3. Open-ended perceived change measure 4. Performance test 5. Anchored trainer ratings 6. Self-perceived learning measure 7. Organizational measures 8. Participation measures 9. Personality measures	1. Correlations 2. ANOVA 3. Cluster analysis	1. Demonstrated the use of a sophisticated experimental design 2. Assessed possible test treatment interaction 3. Used large number of independent and dependent measures
1. Self-report measures developed as part of the study 2. Quantifiable criteria developed as part of study (e.g., number of meetings)	1. ANCOVA 2. ANOVA 3. Factor analysis 4. Test-retest reliabilities 5. Correlation coefficient & test of significance	1. Demonstrated the use of an experimental design group employing a control group 2. Tailored self-report measures to the organization 3. Developed quantifiable criteria for a managerial work group 4. Demonstrated a method of matching questionnaires at T_1 and T_2 and maintaining anonymity 5. Established equivalence of groups
1. Likert (1967) Profile of Organization Characteristics		1. Explained sources of internal invalidity in an OD context
1. Number of memos 2. Potential dollar savings	1. Comparisons of number of memos 2. Comparisons of the cost per memo	1. Provided a means for developing hard criteria for managerial work groups
1. Productivity measured in number of units from Huse & Beer (1971)	1. Modified ANOVA	1. Provided a statistical procedure for evaluating organizational change with data that do not meet independence assumption required for statistical tests 2. Could be used to determine time lags in hard criteria
1. Self-report measures tailored to the organization	1. Not specified	1. Demonstrated how a researcher can patch up a weak design and discount rival hypotheses

how the frequency of meetings could be used to evaluate an OD intervention. Harvey and Boettger (1971) demonstrated how to use number of memoranda and their comparative cost to evaluate an intervention designed to improve work group communications.

The purpose for summarizing these studies is twofold. First, they are representative of the studies published prior to 1976. Second, they illustrate types of issues that were of importance to OD practitioners.

The common thread in these studies is that each demonstrated a method of dealing with a general evaluation problem.

Phase 3: Specific Methodological Issues

As mentioned, pre-1976 OD research was concerned largely with evaluation issues that dealt primarily with relatively macrolevel issues. This orientation changed markedly in 1976 with Golembiewski, Billingsley, and Yeager's (1976) operationalization of three types of change: (1) alpha change or real change, (2) beta change or scale recalibration, and (3) gamma change or concept redefinition. Although their paper can be regarded as a classic (indeed, it received the 1975 Douglas McGregor Memorial Award) that has stimulated much-needed research, one aspect of this change typology, the concept of scale recalibration, was actually introduced to OD readers by Walker, Shack, Egan, Sheridan, and Sheridan (1972). These researchers, in turn, referred to work by Hurley and Hurley (1969), which revealed that, during the course of a training session, several participants realized that their pretest Jourard Self-Disclosure Questionnaire (JSDQ) scores reflected an unrealistically high assessment of their actual level of self-disclosure. Consequently, on the following posttest, approximately half of the 50 participants showed a decrement in their JSDQ responses.

Analyzing this result, Walker et al. (1972) found significant decrements in JSDQ posttest responses in two experimental groups ($n_1 = n_2 = 12$), but not for a comparison group ($n_3 = 12$). Anecdotal and nonquantifiable reactions offered by participants revealed that during the posttest they realized that they had overrated their self-disclosure. Walker et al. subsequently concluded that participants had undergone a learning process during the session, which prompted them to reevaluate their pretest scores and thus recalibrate the JSDQ response scale.

Following the work of Walker et al., several investigations were published that pursued the study of change by employing statistical procedures at the group level. That is to say, questionnaires were analyzed by comparing responses of one group with those of another. An implicit assumption of this level of analysis is that errors that are responsible for scale recalibration and/or concept redefinition are relatively systematic. Consequently they can be associated with a specific group. Studies by Golembiewski et al. (1976), Armenakis and Smith (1978), Armenakis and Zmud (1979), and Koch and Rhodes (1979) are characteristic of this methodology.

In order to refine these procedures, Terborg, Howard, and Maxwell (1980) and Bedeian, Armenakis, and Gibson (1980) have proposed methodologies to detect change for *each* individual within a group. An implicit assumption of the proposed methodologies is that errors responsible for scale recalibration and/or concept redefinition are *not* relatively systematic and may vary by individual. Stated differently, moderating variables may be so numerous (or perhaps unknown) that the researcher cannot objectively group respondents. Therefore, the most feasible strategy is to analyze responses individually.

Two points should be made regarding the above methods. First, differences exist in the two procedures. Advantages and disadvantages are associated with each. Second, the methods in question identify the existence but not the cause of scale recalibration and concept redefinition. It appears, therefore, that future OD research is needed in at least three areas. One is to reconcile differences in available procedures for detecting scale recalibration and concept redefinition and to propose and test new methodologies. Such methodologies should not be restricted to using only perceptual measures. Researchers should investigate the possibility of using unobtrusive measures to corroborate the existence of scale recalibration and concept redefinition (Sechrest, 1979). A second is to determine causes of the scale recalibration and concept redefinition phenomena. In addition, because evaluation research invariably is conducted over time, a third focus of needed future research is the issue of properly executed longitudinal studies.

Future Research

Scale Recalibration and Concept Redefinition

The preceding review of the chronological development of OD research suggests that the type of evaluation questions addressed has progressed from a macrolevel to a microlevel. Indeed, this may be the natural evolution of a science. In this regard, OD is a new and rapidly developing discipline. OD practitioners have become more aggressive and have become more specific about the questions addressed in their research. Perhaps it is time to begin question-

ing some of the implicit methodological assumptions that have been ignored with the wide emphasis on macroissues. For example, much of the research data that form the foundation of current OD knowledge has been acquired via survey research using a Likert scale or its equivalent. Yet, in the very manuscript that introduced this scale, Likert, quoting Rice (1930), issued the following warning:

> The difficulties of building scales similar to Thurstone's and of applying them to the measurement of the attitudes of social groups, become increasingly difficult once we leave the classroom, the discussion club and the other small, comparatively infrequent and highly selected groups that enjoy having experiments tried upon them. Such groups already have developed ways of making their attitudes articulate. It is the more numerous work-a-day groupings of society, which are inaccessible to his controlled measurements, about whose attitudes the social scientist is in the most need of information. Students may be required, good natured academicians may be cajoled, and sundry needy persons may be paid to sort cards containing propositions into eleven piles. But it is difficult to imagine securing comparable judgments of satisfactory measurements in the final application, from bricklayers, businessmen, Italian-Americans, nuns, stevedores, or seamstresses. And, unless the scale itself is based upon equal-seeming differences to a random sample of the group which is to be measured, its validity—the degree to which it purports to measure—becomes open to question (Likert, 1932, p. 24).

This warning points to the foundation of the scale recalibration issue. For instance, one may hypothesize that a respondent's inability to articulate an opinion, for whatever reason, may be responsible for observed beta change. This issue was being investigated some 50 years ago in psychology and appears analogous to what is referred to in psychophysics (Guilford & Park, 1931; Pratt, 1933) as time-order error (TOE).

According to Guilford (1954), TOE exists when stimuli are presented for comparative purposes and the second of a pair is judged to be greater or less than would be expected. Conditions affecting TOE include: (1) general level of stimuli; (2) range of stimuli applied; (3) time interval between stimuli; (4) experience of an observer in an experiment; (5) background stimuli; and (6) other incidential conditions. The examples offered by Guilford are not couched in terms of OD research, but illustrations can be formulated that are readily relevant (see Table 2).

For example, in typical psychophysics experiments subjects are requested to indicate the similarity of line lengths or similarity of sound tones. In OD research a change agent may be concerned with leader behavior as measured through a self-report instrument. An experiment to determine if the general level of a

Table 2
Listing of Time-Order Errors in Survey Research[a]

Time-Order Error (TOE)	*Description*	*Example*
General level of stimuli	TOE may vary by level of stimuli depending on the framework in which it is presented.	A certain level of leader behavior (e.g., consultative) may be associated with TOE, and another (e.g., authoritative behavior) may not.
Range of stimuli	TOE may vary as a function of the observed extremes, i.e., the difference between the lower and upper limits, of a group of stimuli.	If at a given time an observed leader behavior is judged to be authoritarian and at a subsequent time participative, the difference in assessment may be due to the contrasts of the two behaviors.
Time interval between stimuli	TOE may vary as a function of the time interval between the pairs of stimuli.	A longer time interval (e.g., 12 weeks) between two administrations of a survey research questionnaire may be associated with more TOE than a shorter time interval (e.g., 2 weeks).
Observer experience	TOE may vary as a function of a respondent's experience in performing an experimental task.	Respondents who are inexperienced in articulating their perceptions of leader behavior on a survey research instrument may evidence more TOE than those who are experienced.
Background stimuli	TOE may vary as a function of background stimuli that impinge on a respondent simultaneously with a comparison stimuli or those interpolated between or those extropolated before or after.	If conditions of test administrations are different at Time 1 from Time 2, any divergence in response patterns may be due to TOE.
Other incidental conditions	Different methods of measurement may be associated with different magnitudes of TOE.	A behaviorally anchored rating scale may be associated with a different magnitude of TOE than a Likert-type scale.

[a]Adapted from Guilford (1954).

stimulus is responsible for scale recalibration could involve showing subjects, via videotape, various levels of leader behavior (e.g., varying from authoritative through participative) to determine if one "magnitude" of behavior is more associated with TOE than another. An example of an experiment to determine the impact of the time interval between pairs of stimuli (e.g., questionnaire administrations) might discern if lengthy, as opposed to short, time intervals are more closely associated with the TOE phenomena.

Admittedly, the conditions described in Table 2 have not been tested in an evaluation setting. However, from laboratory studies (Needham, 1934) it seems logical that such phenomena are relevant to survey research. A significant contribution could be made to the understanding of scale recalibration if the conditions identified in Table 2 were developed as testable hypotheses.

Longitudinal Studies

Only a cursory review of the social science literature is necessary to reveal that an increasing number of researchers are calling for investigations employing longitudinal designs (Brightman, 1971; Cummings, Molloy, & Glen, 1977; Ivancevich & Matteson, 1978; Kimberly, 1976). After a critique of 35 published OD evaluations (screened from a total of 160), Porras and Berg (1978) recommended that investigators (1) increase the length of time devoted to collecting data on change and (2) increase the frequency with which data are collected.

Although the need for more longitudinal research is understood, a more concise meaning of the term *longitudinal* is necessary, especially with respect to research involving self-report measures. Numerous researchers have concluded logically that a longitudinal design should contain at least three observations, but Arundale (1980) persuasively argues that two additional conditions should be met. The first is the time interval or frequency with which observations are to be taken (e.g., every 10 days). The second is the span of measurement or duration of time for which observations are to continue (e.g., six weeks).

In order to satisfy these conditions, Arundale has provided a general guideline of importance to organization researchers. For a discrete state variable (i.e., one that would be measured with a Likert scale or equivalent), "the sampling interval must be equal to (or shorter than) the shortest time interval for which the variable under study can remain in any one of its states" (1977, p. 261). Obviously, this guideline implies that organization researchers must be aware of the distribution characteristics of a variable or dimension being investigated and must be capable of matching the nature of the measurement strategy employed to the specific kind of intervention in question. Both requirements suggest that experiments should be designed to increase basic knowledge of variables commonly measured. In order to gain this understanding, it would seem necessary to incorporate both emerging research on the evolution of organizational dimensions over time, e.g., organizational life cycles (Kimberly, 1980) and established research on the sampling theorem (Cherry, 1957). With regard to the latter, as Arundale (1977) points out, the sampling theorem provides guidance for sampling across time in order to obtain representative data. Basically, it indicates that the time interval for ascertaining measurements should equal roughly one-half of the cycle time necessary for a variable to progress from one state to the next. In other words, if a variable attains a value of X at T_1, and next assume the value of Y at T_5, then the time interval for measurement should be $(T_5 - T_1)/2$. Measurements should be ascertained, therefore, at T_1, T_3, and T_5. To illustrate, if one organization dimension, say System 2 in the Likert (1967) framework, exists at T_1 and another (System 3) exists at T_5, then the measurement of the dimension should be taken at every $(T+2)$ units of time.

At present, very little is known about the transient nature of organizational dimensions. For example, it is not known whether these dimensions evolve through something similar to an organizational life cycle or whether their evolution follows another pattern. However, there appears to be an implied assumption in most OD research that organizational dimensions are stable (at least over relatively short time periods) between interventions. If one rejects this assumption and accepts the plausibility of a phenomenon similar to an organizational life cycle, then the necessity for investigating the relevance of the sampling theorem to OD becomes more relevant. Admittedly, the current level of understanding relating both to matching of measurement strategies to specific kinds of interventions and to distribution characteristics of organizational dimensions is limited. With respect to the latter, as Kimberly points

out, it is not known whether "there are laws that govern the development of organizations, analogous to those that apparently govern the development of [biological] organisms" (1980, p. 7). One thing, however, is clear. The determination of whether such laws exists is imperative. The magnitude of needed research may be uncertain, but the direction is obvious.

Summary and Conclusion

Published research on OD evaluation can be classed into three categories: (1) identification of general problems and development of guidelines; (2) demonstrations of methods for evaluating change programs; and (3) identification and resolution of specific methodological issues.

Initially, OD research was concerned with issues at a macrolevel (e.g., experimental design and use of statistical methods). However, since 1976 researchers have concentrated on microlevel issues (e.g., measurement of types of change). For the immediate future it appears that there are at least three microissues that should be addressed: (1) reconciliation of advantages and disadvantages in dealing with measurement of types of change at the individual and group levels of analysis as well as testing new methodologies; (2) identification of reasons for scale recalibration and concept redefinition; and (3) investigation of the time interval and measurement span issues as related to longitudinal research designs.

This review has summarized past and present research on evaluation and has offered directions for future research. Similar reviews are needed for other aspects of the OD process. For example, developments in diagnosis have increased significantly the comprehensiveness of diagnostic methodologies (Jenkins, Nadler, Lawler, & Cammann, 1975). Similarly, several differing conceptual diagnostic frameworks have been proposed (Nadler & Tushman, 1977; Tichy, Hornstein, & Nisberg, 1976; Weisbord, 1976). Evaluation as an aspect of OD comprises but a single element of a much larger and integrated whole.

References

Armenakis, A., & Feild, H. Evaluation of organizational change using nonindependent criterion measures. *Personnel Psychology,* 1975, 28, 39-44.

Armenakis, A., & Smith, L. A practical alternative to comparison group designs in OD evaluations: The abbreviated time series design. *Academy of Management Journal,* 1978, 21, 499-507.

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

Armenakis, A., Feild, H. & Holley, W. Guidelines for overcoming empirically identified evaluation problems of organization development change agents. *Human Relations,* 1976, 29, 1147-1161.

Armenakis, A., Feild, H. & Mosley, D. Evaluation guidelines for the OD practitioner. *Personnel Journal,* 1975, 54 (2), 99-103.

Arundale, R. Sampling across time for communication research: A simulation. In P. Hirsch, P. Miller, & F. Kline (Eds.), *Strategies for communication research.* Beverly Hills, Cal.: Sage, 1977, 257-285.

Arundale, R. Studying change over time: Criteria for sampling from continuous variables. *Communication Research,* 1980, 7, 227-263.

Bedeian, A., Armenakis, A., & Gibson, R. The measurement and control of beta change. *Academy of Management Review,* 1980, 5, 561-566.

Bennis, W. Theory and method in applying behavioral science to planned organization change. *Journal of Applied Behavioral Science,* 1965, 1, 411-428.

Blake, R., Mouton, J., Barnes, L. & Greiner, L. Breakthrough in organization development. *Harvard Business Review,* 1964, 42 (6), 133-155.

Brightman, H. The need for repeated measurement designs in organizational research. *Academy of Management Journal,* 1971, 14, 398-402.

Cherry, C. *On human communication.* New York: Wiley, 1957.

Cummings, T., Molloy, E., & Glen, R. A methodological critique of fifty-eight selected work experiments. *Human Relations,* 1977, 30, 675-708.

Evans, M. Opportunistic organizational research: The role of patch-up designs. *Academy of Mangement Journal,* 1975, 18, 98-108.

Friedlander, F. The impact of organizational training laboratories upon the effectiveness and interaction of ongoing work groups. *Personnel Psychology,* 1967, 20, 289-308.

Golembiewski, R., & Carrigan, S. The persistence of laboratory induced changes in organization styles. *Administrative Science Quarterly,* 1970, 15, 330-340.

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

Guilford, J. *Psychometric methods.* 2nd ed. New York: McGraw-Hill, 1954.

Guilford, J., & Park, D. The effect of interpolated weights upon comparative judgments. *American Journal of Psychology,* 1931, 43, 589-599.

Harrison, R. Impact of the laboratory method on perceptions of others by the experimental group. In C. Argyris (Ed.), *Interpersonal competence and organizational effectiveness.* Homewood, Ill.: Irwin, 1962, 261-271.

Harrison, R. Research on human relations training: Design and interpretation. *Journal of Applied Behavioral Science,* 1971, 7, 71-85.

Harvey, J., & Boettger, C. Improving communication within a managerial work group. *Journal of Applied Behavioral Science,* 1971, 7, 164-179.

Hurley, J., & Hurley, S. Toward authenticity in measuring self-disclosure. *Journal of Counseling Psychology,* 1969, 16, 271-274.

Huse, E., & Beer, M. Eclectic approach to organizational development. *Harvard Business Review,* 1971, 49 (5), 103-112.

Ivancevich, J., & Matteson, M. Longitudinal organizational research in field settings. *Journal of Business Research,* 1978, 6, 181-201.

Jenkins, C., Nadler, D., Lawler, E., & Cammann, C. Standardized observations: An approach to measuring the nature of jobs. *Journal of Applied Psychology,* 1975, 60, 171-181.

Kimberly, J. Issues in the design of longitudinal organizational research. *Sociological Methods and Research,* 1976, 4, 321-347.

Kimberly, J. The life cycle analogy and the study of organizations: Introduction. In J. Kimberly, R. Miles, & Associates, *The organizational life cycle.* San Francisco: Jossey-Bass, 1980, 1-17.

Koch, J., & Rhodes, S. Problems with reactive instruments in field research. *Journal of Applied Behavioral Science,* 1979, 15, 485-506.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology,* 1932, 22, 5-55.

Likert, R., *The human organization.* New York: McGraw-Hill, 1967.

Miles, M. Changes during and following laboratory training: A clinical-experimental study. *Journal of Applied Behavioral Science,* 1965, 1, 215-242.

Nadler, D., & Tushman, M. A diagnostic model for organizational behavior. In J. Hackman, E. Lawler, & L. Porter (Eds.), *Perspectives on behavior in organizations.* New York: McGraw-Hill, 1977, 85-100.

Needham, J. The time error in comparison judgements. *Psychological Bulletin,* 1934, 31, 229-243.

Nicholas, J. Evaluation research in organizational change interventions: Considerations and some suggestions. *Journal of Applied Behavioral Science,* 1979, 15, 23-40.

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

Pratt, C. The time error in psychophysical judgments. *American Journal of Psychology,* 1933, 45, 292-297.

Rice, S. Statistical studies of social attitudes and public opinion. In S. Rice (Ed.). *Statistics in social studies.* Philadelphia: University of Pennsylvania Press, 1930, 171-196.

Sechrest, L. (Ed.). *New directions for methodology of behavioral science: Unobtrusive measurement today.* San Francisco: Jossey-Bass, 1979.

Shephard, H. Three management programs and the theory behind them. In *An action research program for organization improvement.* Ann Arbor: Foundation for Research on Human Behavior, University of Michigan, 1960, 1-5.

Terborg, J., Howard, G., & Maxwell, S. Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. *Academy of Management Review,* 1980, 5, 109-122.

Tichy, N., Hornstein, H., & Nisberg, J. Participative organization diagnosis and intervention strategies: Developing emergent pragmatic theories of change. *Academy of Management Review,* 1976, 1 (2), 109-120.

Walker, R., Shack, J., Egan, G., Sheridan, J., & Sheridan, E. Change in self-judgments of self-disclosure after group experience. *Journal of Applied Behavioral Science,* 1972, 8, 248-251.

Weisbord, M. Organizational diagnosis: Six places to look for trouble with or without a theory. *Group and Organization Studies,* 1976, 1, 430-447.

Achilles A. Armenakis is Director of the Auburn Technical Assistance Center, Auburn University.

Arthur G. Bedeian is E. L. Lowder Professor of Management, Auburn University.

Samuel B. Pond is Assistant Professor of Psychology, North Carolina State University, Raleigh.