AD-A104 550 NEBRASKA UNIV LINCOLN DEPT OF MANAGEMENT F/G 5/10
A RELIABILITY ASSESSMENT OF PARTICIPANT OBSERVATIONAL MEASURES --ETC(U)
JUL 81 F LUTHANS, D LOCKWOOD, M CONTI N00014-80-C-0554

UNCLASSIFIED NL

1 OF 1
AD
A104550

END
DATE
FILMED
08
DTIC

LEVEL II (11)

A RELIABILITY ASSESSMENT OF PARTICIPANT OBSERVATIONAL MEASURES OF LEADER BEHAVIOR IN NATURAL SETTINGS

Fred Luthans
Diane Lockwood
Mary Conti

University of Nebraska-Lincoln

Send correspondence to:

Fred Luthans
Department of Management
University of Nebraska
Lincoln, Nebraska 68588
Phone: 402-472-2324/3915

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 3

2. GOVT ACCESSION NO.: AD-A104550

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): A Reliability Assessment of Participant Observational Measures of Leader Behavior in Natural Settings

5. TYPE OF REPORT & PERIOD COVERED: Interim rept.

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Fred Luthans, Diane Lockwood, [redacted] Mary Conti

8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0554

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Department of Management
University of Nebraska
Lincoln, NE 68588-0400

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR170-913

11. CONTROLLING OFFICE NAME AND ADDRESS:
Organizational Effectiveness Research Program
Office of Naval Research (Code 452)
Arlington, VA 22217

12. REPORT DATE: Jul[redacted]81

13. NUMBER OF PAGES: 30

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 38

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Leadership Measures
Reliability of Observation
Observational Measures
Interrater Reliability
Measuring Leader Behavior
Leadership in Natural Settings

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study makes a reliability assessment of 88 trained participant observers who measured the behavior of 120 target leaders in 5 diverse organizational settings. Eight trained outside observers were used as agreement checks. Drawing from three methods of calculation, the interrater agreement was quite impressive. Other analysis techniques employed in the study support the value of the training given to the observers. The overall conclusion of the study is that, especially in light of the current dissatisfaction, observation may be an effective measurement alternative.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

395368

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A RELIABILITY ASSESSMENT
OF PARTICIPANT OBSERVATIONAL MEASURES
OF LEADER BEHAVIOR IN NATURAL SETTINGS

Abstract

This study makes a reliability assessment of 88 trained participant observers who measured the behavior of 120 target leaders in 5 diverse organizational settings. Eight trained outside observers were used as agreement checks. Drawing from three methods of calculation, the inter-rater agreement was quite impressive. Other analysis techniques employed in the study support the value of the training given to the observers. The overall conclusion of the study is that, especially in light of the current dissatisfaction surrounding leadership theory and research methods, observation may provide a reliable alternative measurement technique to widely used standardized leadership questionnaires.

Dist
A

A RELIABILITY ASSESSMENT OF PARTICIPANT OBSERVATIONAL MEASURES OF LEADER BEHAVIOR IN NATURAL SETTINGS

Some areas of psychology are depending more on observational research methods. For example, about a third of recently published research articles in the developmental area used observational techniques (Mitchell, 1979). On the other hand, the use of such direct measures has been almost totally ignored in field research in organizational behavior. Popular research topics such as leadership have depended almost solely upon indirect questionnaire techniques. These measures may be a major contributing factor to the dreary state of leadership theory and research. Miner (1975), for example, has called for the demise of leadership altogether and Kerr (1976) has suggested that substitutes for leadership be identified and used in explaining influence processes.

Comprehensive analyses indicate that the widely used leadership questionnaires have not been demonstrated to have reliability nor validity (Schriesheim, Bannister, & Money, 1979; Schriesheim & Kerr, 1977). In a reflective analysis of the leadership field, Campbell (1977) noted that at best we may have a science of questionnaire behavior rather than leader behavior. In other words, the time seems ripe to explore alternatives to questionnaire measures of leader behavior. Until better measures are developed, there seems little hope for advancing our knowledge of leadership. As Korman (1974) forcefully pointed out: "Measurement and theory go hand-in-hand and the development of one without the other is a waste of time for all concerned. . . The point is not that adequate measurement is 'nice.' It is <u>necessary</u>, <u>crucial</u>, etc. Without it, we have nothing" (p. 199).

Questionnaires will obviously always have a place in leadership research, but they should not be used to the exclusion of all other measures. As Graen and Cashman (1975) noted: "Until we develop several different measures of

our leadership construct and establish that these different measures form a consistent network of relationships, we can have little confidence in whatever our measure of leadership is really tapping" (pp. 150-151). Observation techniques, as successfully developed and used in other areas of psychology, would seem to be a viable alternative measurement approach in the field of leadership.

There are a number of reasons why observational measures have been largely ignored in field research in organizational behavior in general and leadership research in particular, but perhaps the best explanation is simply that it is much easier (in terms of time, money, and effort) to ask than to observe. Because it is much harder to use observational measures, the potential problems are often highlighted and emphasized by those depending upon questionnaires to gather data. For example, several organizational scholars have stressed that observational techniques are incapable of measuring important dimensions such as planning or thinking (Carroll & Taylor, 1968; Hemphill, 1959; Kelly, 1969; Penfield, 1974). The logic is that there are as many or even more problems associated with observations as with questionnaires, so why not stick with questionnaires.

There is no doubt that there are potential problems associated with observational measures (see Weick, 1968, for a comprehensive analysis of the pros and cons of observational measures). By the same token, there are procedures that can be employed (e.g., careful training of the observers) to help overcome potential reliability problems. It also must be remembered that just as observation cannot measure private thoughts, questionnaires are incapable of measuring the important interactive dimensions of leadership. As Kerlinger (1973) points out: "observations must be used when the variables of research studies are interactive and interpersonal in nature" (p. 554). With the increasing recognition of the need for an interactionist perspective for organizational behavior in general (e.g. Campbell, Dunnette, Lawler, &

Weick, 1970; Davis & Luthans, 1980; Terborg, Richardson & Pritchard, 1980) and leadership in particular (Barrow, 1976; Davis & Luthans, 1979; Greene, 1975; Luthans, 1979), observational methods take on new importance.

This study takes the important first step in assessing whether or not participant observers can reliably measure leader behavior in natural settings. An operational definition of measurement reliability is difficult to state because it is determined somewhat differently depending upon the type of methodology being assessed. Traditionally, reliability is considered to be an assessment of the True Score component contained in an Observed Score (i.e., True Score = Observed Score + Error) and is generally considered to be the extent to which a measurement is repeatable (Nunnally, 1978).

In the case of observational research, reliability has not received as much attention as it has in the more traditional questionnaire methods (Johnson) & Bolstad, 1973). Mitchell (1979) identifies three ways in which reliability of observational data can be assessed: first, the extent to which two observers, working independently, agree on what behaviors are occurring; second, the observational measure could be treated as a special case of a standardized psychological test and classic psychometric techniques (i.e. test-retest) could be employed; finally, Cronbach's theory of generalizability (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Cronbach, Nageswari & Gleser, 1963) that recognizes the influence of a number of different aspects of the observation situation could be used. In choosing which of these approaches to use in assessing the reliability of observationally gathered leadership data, the potential sources of error should be explored.

Sources of unreliability (i.e., disagreement or error) in observational research of leadership behavior would seem to stem mainly from two sources: (1) the categorizaton and recording scheme of the observational instrument itself; and (2) the problems in information-processing (i.e., observing, judging and recording) by the observers themselves (Campbell, 1958). Errors from the first source largely stem from the validity of the instrument. While assessing validity is beyond the scope of the present paper, it is clear that some observational validity issues affect reliability as well as vice versa. Weick (1968) has identified the possible errors due to the observational scheme or instrument and these can be summarized as follows:

1. Inference. Errors in observational research can stem from the degree of inference left to the observer in classifying a behavior. Thus, the nature and number of observational categories become crucial for reliability. As a rule of thumb, the fewer the categories, the more precise their definition, and the less inference required in making classifications, the greater will be the reliability of the data (Gellert, 1955). Since observation is largely concerned with "public, visible, and external events" (Weick, 1965, p. 358), the observed behaviors should be accessible to the senses. It follows, then, that to the extent that a behavior is not clearly visible nor audible (i.e., is more hidden or private), there is likely to be disagreement between observers regarding whether it occurred and concerning its exact nature. Similarly, observers generally respond to the manifest content of behavior, even when the covert or implied meaning may be more relevant or accurate. This is especially true in delicate interpersonal situations such as superior/subordinate interactions, disciplinary action or conflict management. To the extent that the categorization scheme clearly and unambiguously defines which behaviors should be classified where, the instrument itself can

certainly help overcome (but not entirely eliminate) the inference problem.

2. Time Sampling. Time sampling represents another form of the problem of inference. That is, since the selected time periods may not contain the natural limits of a behavior, an observed behavior may be somewhat arbitrarily forced into a category (or, more likely, left out), thus making it less likely that another observer would report the behavior in the same way. By sampling a greater number of time periods, a more adequate final "picture" is likely to be captured, but such repeated time sampling brings with it other sources of error. In the first place, of course, time pressures on observers may cause them to hurry, thus possibly causing inaccuracies. Second, extended intervals between observational periods lead to varying behaviors by the observers themselves, thus causing variability due to their own activities (e.g., more or less attention may be paid to observing, or recent interests/activities may cause them to classify the observed behaviors differently, etc.).

3. Context. The context errors result from situations surrounding the observation influencing the observer's accuracy. The literature is somewhat ambiguous regarding exactly how familiarity with context affects observation. On the one hand, Weick (1968) argues, "the categories should be so explicit as to discourage the use of context in classifying observed behaviors" (p. 423). Yet both he, Nunnally (1978) and others have noted that a thorough knowledge of the context of a behavior is essential in order to assess it accurately. For example, a thorough familiarity with a leader's

habitual patterns of speech may help an observer to distinguish between manifest and latent, or intended, meaning in conversation. In order to resolve the seeming incongruity regarding the value of a knowledge of context, what seems useful is that the observer have a substantial sense of field independence. That is, the observer needs to be sufficiently familiar with the gestalt (context) to make accurate judgments, yet also must be capable of recognizing specific unique behavior episodes in order to separate and classify them accurately. By the same token, overfamiliarity with the situation may also cause problems. For example, if there is any ambiguity in the behavior, overfamiliarity is likely to cause biases toward certain categories. This is similar to the phenomenon of observers developing early hypotheses and then categorizing subsequent observations in ways which follow those hypotheses. For example, a subordinate has known the leader a long time and feels the leader criticizes subordinates constantly; then, a behavior which might legitimately be considered "reinforcing" may be categorized by the subordinate-observer as "discipline". Overfamiliarity can also lead to boredom which is likely to cause errors itself through inattention, carelessness, etc.

Besides errors stemming from the instrument and procedures in using it, the observer him/herself may be the source of errors. As Kerlinger (1973) has noted: "The major problem of behavioral observation is the observer himself" (p. 538). Campbell (1958) described observer errors as largely information-processing problems. That is, they occur when

fallible humans are involved in a communicating system (duplicating, transmitting and translating information). Eight of the most common errors pertinent to inter-personal perception were used by Thornton and Zorich (1980) in an experimental training study to improve observer accuracy. These included:

(1) loss of detail through simplification;

(2) making snap judgments (e.g., using easily codable or familiar categories, and overlooking more abstract or unfamiliar ones);

(3) middle message loss (concentrating only on the first or last part of a behavior, overlooking during the mid-portion);

(4) categorization error or mind-reading (trying to second-guess another's thoughts or motives);

(5) contamination from prior behavior (tendency to code behaviors on the basis of activity immediately prior to the observation period);

(6) contextual or setting errors (allowing the situation to distort observations);

(7) stereotyping and prejudice (assuming that certain groups or classes of people will behave in ceratin ways);

(8) halo (being overly influenced by one characteristic of a person, either positively or negatively).

While this list is not exhaustive, it is quite representative of the kinds of errors due to observers themselves, apart from the observational measurement instrument itself.

An increasing number of studies have analyzed the use of training to improve observer accuracy (Bernardin, 1978;1979; Bernardin & Walter, 1977; Borman, 1975; Latham, Wexley & Pursell, 1975; Warmke & Billings, 1979; Wexley, Saunders, & Yukl, 1973). While in an old study Levine and Butler (1952) found that simple awareness of the possible errors did not reduce observers' tendency to commit them, researchers who have used more comprehensive training--including giving the observers a chance to practice observing and rating (Wexley, et al., 1973)--have shown significant increases in accuracy. Such practice exercises stem from the modeling principles of social learning theory (Bandura, 1977). Using videotaped exercises, so that the "correct" observations were already known, Thornton and Zorich (1980) were able to obtain a significant improvement in observer accuracy due to training their subjects on the eight specific observer error tendencies mentioned previously.

In summary, the literature clearly indicates that when observers are made aware of their typical error tendencies, coached on how to avoid making those errors, and given supervised practice in observing, they can improve their accuracy significantly beyond untrained observers.

The purpose of this study is to assess the reliability of trained participant observers using a thirteen category leadership behavior instrument. The derivation and description of this instrument is briefly discussed in the method section, but its validity is currently being assessed in other studies. As has been pointed out, there are a number of ways that observer reliability could be assessed, but the most common and accepted (Bijou, Peterson & Ault, 1968; Kelly, 1977; Mitchell, 1979; Weick, 1968) inter-rater agreement approach is mainly used here. Both the traditional and two alternative ways of calculating percentage of agreement are used. In

addition, the generalizability approach to reliability suggested by Cronbach (Cronbach, Gleser, Nanda, & Rajaratham, 1972; Cronbach, Rajarathnam & Gleser, 1963) is drawn upon.

Method

Settings and Subjects

The study was conducted in five purposely diverse organizational settings: a fairly large financial institution, a state agency, a medium sized manufacturing plant, a campus police department, and the Navy and Army R.O.T.C. units of a university. All those with supervisory responsibilities (from the very top to first line supervision) in the financial institution (N=52), campus police (N=16), and R.O.T.C. (Army and Navy) units (N=15) and systematically selected supervisors/managers in the state agency (N=18) and manufacturing plant (N=19) served as the target leaders (N=120) in the study. These target leaders typically had been with their respective organizations 6-10 years and in their present positions 1-5 years. Almost all of them fell in the 26-55 age range and a great majority had a college education. Their jobs covered the whole range of functions found in their respective organizations.

The participant observers (N=88) were selected jointly by the researchers and the personnel managers of the respective organizations (or the designated project officers in the case of the campus police and R.O.T.C. units) according to the following criterion: Does this person have maximum visual and audible contact with the target leader and have a good understanding of the functions, terminology, and nature of the work performed by the target leader? The target leader's approval was also needed and secured in all cases. The selected

participant observer in a large majority of cases turned out to be the target leader's secretary. If not a secretary, then a key subordinate was used. Eleven (12%) of the participant observers were responsible for observing two target leaders and 7 (8%) of the participant observers had three target leaders. This was discouraged as much as possible and only occurred when in the opinion of the researchers/personnel managers it was better to meet the criterion of selection as a participant observer in the study and observe more than one target leader than select another observer but not meet the selection criterion nearly as well. This usually was the case where one secretary served more than one target leader. These participant observers had considerable job experience but little formal education and, of course, had no knowledge of the literature on leadership research or theory. Except for the training they received (which will be described), they had little or no knowledge of the specifics of the study.

The 8 outside observers were graduate students in management. Three were assigned to the financial institution, two to the manufacturing plant, and one to the state agency and one to the campus police and R.O.T.C. units. An experienced graduate student (a Ph.D. student with the most knowledge of the study) observed at all the sites and largely coordinated the efforts of the other 7 outside observers. These outsider observers had briefly studied leadership theory and research in their course work in management, received training which will be described later and had a very general idea of the objectives and procedures of the study.

The Observation Instrument

The Leader Observation System or simply LOS instrument used by the participant observers is part of an ongoing larger study and will only be briefly described here.

The 13-category instrument was developed in 2 major steps. First, 44 managers at all levels in all types of organizations (not the 132 target managers observed in this study) were freely observed for an hour each day over a two week period of time (i.e., 440 hours of unstructured observation of managers in their natural settings). The 44 observers were management students who were given a training workshop pointing out the systematic errors commonly found in observing others (i.e., it followed the procedures suggested by Thornton & Zorich, 1980). In addition, they practiced writing descriptive observational logs from several role playing exercises that were then critiqued by the trainers/researchers. The observers were trained to observe continuously the behavior of the target manager over the hour; to record specific, identifiable behaviors on their logs; and to be reporters concentrating on objective description rather than trying to judge or evaluate the behaviors observed. These observers had not yet studied leadership theory or research.

While true randomization of the observation times was not always possible, the observers varied their hours as much as possible to help assure representativeness. After the two weeks, the managers were provided with copies of the observational logs and were asked to rate to what extent the recorded behaviors were typical. On a scale of 1-5, the mean rating was 3.9, which indicated the behaviors on the average, were typical "to a considerable extent." The managers were also asked to suggest any additional behaviors which they considered typical. These additions mainly consisted of activities which might be best described as of a sensitive nature, i.e., important policy meetings, disciplining, managing conflict, etc., which would generally be more infrequent than the reported behaviors.

The second major step involved in deriving the instrument used in this study was the considerable job of constructing comprehensive and workable categories to accommodate (contain) the 440 hours of freely observed behaviors. This task was accomplished by a Delphi process (Delbecq, VandeVen, & Gustafson, 1975). The Delphi panel consisted of four persons with considerable academic work in management/leadership and three graduate students from outside disciplines who were completely naive with respect to prior leadership research. All panel members were required to read and become familiar with the processes of constructing adequate behavioral categories as outlined by Kerlinger (1973) and Crano and Brewer (1973).

In the first Delphi round, the panelists independently reviewed the extensive behavioral logs recorded by the observers and suggested categories which would contain the observed behaviors. These categories with accompanying comments were collected and fed back to the panelists, and then through several iterations they further collapsed the categories into smaller but comprehensive sets which could be readily used by participant observers. Thus, the final surviving 12 categories incorporated a multiplicity of opinions and critiques whose purpose was not only to be representative but also exhaustive and mutually exclusive. The 12 final categories on the LOS instrument were broadly labeled the following: (1) planning/coordinating, (2) staffing, (3) training/developing, (4) decision-making/problem solving, (5) processing paperwork, (6) exchanging routine information, (7) monitoring/controlling performance, (8) motivating/reinforcing, (9) disciplining-punishing, (10) interacting with outsiders, (11) managing conflict, (12) socializing/politicking, and an "other" category. Each of these broad categories was then further defined by specific behaviors. For example, monitoring/controlling performance was described behaviorally as (a) inspecting work, (b) walking

around and checking things out, touring, (c) monitoring performance data (e.g., computer printouts, production or financial reports) (d) preventative maintenance.

The format of the instrument listed the behavioral categories along the left hand side and the random times along the top. The random times were for 10 minutes every hour over two weeks or a total of 80 observations. There was a sheet for each day. A nominal measuring format was used; i.e. the observers recorded either the behavior was present ("1") or absent ("0") for each 10 minute time slot. By judging whether the behavior was present or absent the problem of inferring degrees or magnitude of behavior exhibited was avoided (Medley & Mintzel, 1963). The validity of this instrument is currently being assessed and is beyond the scope of this paper.

Observer Training and Procedures

A training workshop conducted by the researchers was held on the premises of each of the participant observers respective organizations. Each session followed the same format, used the same trainers, and took approximately 2 1/2 hours to complete.

About the first half of the observer training workshop was devoted to three areas: First, to provide a very general explanation of the purpose of the observations (i.e. to gather data for input into a profile of the manager's behavior); second, to go over in detail the observational instrument, giving special attention and analysis to the 12 behavioral categories and the procedures for filling out the instrument including what to do if the manager was absent; and third, to give careful instruction of potential observational errors (following Thornton & Zorich, 1980) and how to overcome them. In particular, the potential errors of description versus evaluation and distortion to please the person being observed were

deemed to be particularly relevant to these participant observers and were stressed in the training. For example, the observers were instructed to avoid letting their evaluative biases color their observations, since there are no "good" or "bad" categories on the instrument, the observations would be useful only if they were accurate. By careful explanation and example the trainers showed how the observers could avoid these errors.

The second half of the training was devoted to demonstration and practice. The trainers employed a number of role-playing skits which illustrated the specific leader behavior categories, and the trainees used the instrument to record the behaviors they observed. By following the principles of social learning theory (Bandura, 1976; Latham & Sari, 1979) this aspect of the training was intended to increase observer accuracy through modeling, rehearsal, and repetition. After each role-playing skit, the trainers discussed which behavior category was being illustrated and which specific errors might have been committed during that observation.

In a final role-playing skit, which was rather lengthy and elaborate but realistic, 6 behavioral categories were represented. The observers performance on this last exercise served as an evaluation check for the training. A precise evaluation of observer accuracy is possible, of course, only when there is an objective criterion, i.e., when the "correct" observations are known. Since such an objective criterion was possible in this training exercise, an evaluation of trainee accuracy could be made. Although this data was unavailable in one of the organizations, in the remaining four organizations the participant observer trainees had an overall mean accuracy of 92.5%, with no significant differences between organizations. This accuracy was considerably higher than the 69% obtained by Thornton and Zorich (1980) in their training group. However, it should be noted

that their observers watched a 30-minute episode, whereas the trainees in the present study watched for approximately 10 minutes. The 10-minute time period was chosen to coincide with the time of the sampling observation periods used by the participant observers in the actual data gathering.

The 8 outside observers used in the study were given the same training as the inside, participant observers. After the training, they were given a tour of the facility and were introduced and chatted with the participant observers they would be working with over the two week observation period.

The study was set up to have each target manager have the LOS instrument filled out on him/her 80 times by a participant observer over the two week period of time (a random 10-minute period each working hour over two weeks). This represented a total of 9600 (120 target managers X 80 observation periods) possible observation periods when the instrument was to be filled out. The actual number of observations were somewhat less than this because the target managers were not always available to be observed. Since it was not feasible to have a second observer present at all times, a time sampling technique was employed to gather interrater reliability data. The trained outside observer would randomly appear unannounced and simultaneously record the observed behaviors of the target leaders. After both the outside observer and participant observer independently completed the LOS instrument for the 10 minute time period, the two would then compare notes and discuss what was going on and answer questions of one another. Importantly, they never changed their just-completed recording of the observed behavior, but the post-observation review was used to reinforce the training they had received. A total of 253 such simultaneous observations took place. With a couple of exceptions,

each of the participant observers had three reliability checks (i.e., three sat down with him/her to simultaneously record the behavior of the target manager).

Results

The traditional way to calculate interrater reliability (see Bijou, Peterson & Ault, 1968) is to take the number of agreements divided by the number of joint observations (i.e., the number of agreements plus the number of disagreements). [This will be referred to hereafter as the Traditional Method.] In this study there was a 93.5% interrater agreement using this traditional approach. It is important to note, however, that this traditional method includes the percentage of agreement both on the behaviors that were observed and those that were not observed. For example, suppose that in a given joint observation the participant observer marks seeing "planning" behavior and "motivating/reinforcing" behavior and marks not observing the other 10 categories of behavior. The outside observer marks seeing "planning" behavior and "decision making" behavior and marks not observing the other 10 categories. In this one joint observation there is agreement on 1 behavior and 8 nonbehaviors for an interrater agreement of 75% (9/12). However, such an approach may give misleading results, since it is likely to be so heavily weighted by nonbehaviors.

An alternative approach to assessing interrater reliability is to report only agreement on observed behaviors, leaving out agreement on behaviors that did not occur. Table 1 summarizes this data by category of behavior. Even using this approach there appear to be two logical ways to report the percentage of interrater agreement. If the reliability of the participant observers is being assessed, then it follows that one should report

the interrater agreement percentage as the number of agreements divided by the number of observations made by the participant observers. [This will be referred to hereafter as Behavioral Method A.] This yields an 87.4% interrater reliability. A more conservative way to calculate the reliability would be to assess the accuracy of both the participant observer and the outside observer. This could be done by viewing the participant observers' marked behaviors as one set of data and the outside observers' marked behaviors as another set of data. To get percentage of agreement the following calculation would be made:

$$\frac{\text{Number of Agreements}}{\text{Total Observations - Number of Agreements}}$$

The Total Observations would consist of the Participant Observer Observations plus the Outside Observer Observations. In essence, the agreements represent the intersection of the sets of data. [This will be referred to hereafter as Behavioral Method B.] Using this approach yields a 75.2% interrater agreement.

[Insert Table 1 About Here]

Besides the two behavioral interrater agreement percentages, Table 1 also shows the reliabilities by behavioral category. Disregarding the categories of disciplining and managing conflict because of their low frequency of occurrence, the percentage of interrater agreement (by averaging Behavioral Methods A and B) went from highest to lowest in the following order: exchanging routine information (90.4%), processing paperwork (88.3%), staffing (87.6%), decision making/problem solving (83%), planning/coordinating (79.6%), interacting with outsiders (79.2%), training/developing (77.4%), socializing/politicking (72%), monitoring/controlling performance (67.7%), and motivating/reinforcing (64.4%). In other words, some categories clearly

had higher interrater reliabilities than others. It is also interesting to note that the outside observer used the "other" category twice as often as the participant observer.

Another way of looking at the category reliabilities is in terms of the relative frequency of the observations. Perhaps the frequency in which a behavior is observed will affect its reliability. To determine this possibility, a Spearman rank order correlation between the frequency that the behavior was observed and each of the three ways of determining interrater agreement (i.e. the Traditional Method and Behavioral Methods A and B) was calculated. Table 2 shows that there is not a significant correlation between frequency of observation and interrater agreement. Although not significant, it is interesting to note the negative correlation between frequency of observation and the Traditional Method of calculating interrater agreement but a positive correlation between frequency and interrater agreement calculations from Behavioral Methods A and B.

[Insert Table 2 About Here]

To go beyond the strict interrater agreement assessment of reliability of observational data as suggested by Mitchell (1979), the Cronbach et.al. (1963; 1972) generalizability notion was examined. In particular, a hierarchical regression procedure (Cohen & Cohen, 1975) was used to assess interrater agreement variance as contributed by the type of organization, by outside observer, and over time. These three variables were selected on the basis of previous literature indicating that they would be most likely to contribute to the variance. The results are shown in Table 3.

[Insert Table 3 About Here]

Once again, the results are presented with respect to all three types of interrater reliability. Using a dummy-coded variable in the first step of the regression, the results indicate that there are significant

differences between organizations regardless of how the interrater agreements are calculated. The organization accounts for 17.7% of the variance in overall interrater agreement as calculated by the Traditional Method, 8.3% of the variance in overall interrater agreement as calculated by Behavioral Method A, and 18.3% of the variance in overall interrater agreement as calculated by Behavioral Method B. When entering a dummy-coded variable representing the outside observer, the second step of the regression shows a non-significant increase of the accounted for variance of only 1.3% (bringing the total up to 19%) using the Traditional Method, 1.2% (total to 9.5%) using Behavioral Method A, and 1.4% (total to 19.7%) using Behavioral Method B. In the final step of the regression analysis, the sequencing of the observation checks over the two week time period was entered. In all cases this added nothing to the accounted for variance of the interrater reliabilities.

Discussion

Organizational behavior scholars generally agree today that there is a need for different and better measuring techniques (Cummings, 1981). The study of leadership in particular may be hampered by the almost sole dependence on indirect questionnaire measures with questionable reliability and validity. Alternative methods such as direct observation of leaders in natural settings is talked about and even advocated in methodological discussions, but it is not being used in leadership research. The reasons for its nonuse generally revolve around the issues of reliability and practicality. This study found the use of observational measures of leader behavior to be quite reliable and, although there were some

practical problems with the use of participant observers, in general, can be realistically used.

The results of the reliability assessment indicates 93.5% interrater agreement from the traditional method of calculation and 87.4% and 75% from calculations that only include agreement on observed behaviors (not agreement that a behavior did not occur) on a 12 category instrument. This percentage of agreement between the trained participant observers and the trained outside observers appears to be quite high. Although it is difficult to make direct comparisons because of the lack of observational measures in leadership research, Bass (1954) reports interrater agreement on 12 LGD (leaderless group discussion) studies ranging from 53% to 90% with an average of 75%. In the few behavior management studies which employ observation measures, the reported interrater agreements run 90% or better. For example, Komaki, Waddell & Pearce (1977) report 93.8% agreement on 6 interrater reliability checks on the observed performance behavior of a game room attendant; Komaki, Barwick & Scott (1978) report 97.4% and 99.6% agreements in two departments in an observational study of manufacturing employees' safety behaviors; Shook, Johnson & Uhlman (1978) report 100% agreement on 6 checks and 98.3% agreement on 27 checks in two experiments involving observed staff performance behaviors in a special education unit; and Luthans, Paul & Baker (1981) report 94% agreement on 487 joint observations of salespersons' performance behaviors. However, it is also important to note that a methodological review of 19 behavioral research studies in business settings found only 5 (26%) reporting any reliability assessments of the mostly observational measures used (Andrasik, 1979). In addition, with the exception of a recent observational study of the managerial activities of police chiefs which reported a 91% interrater agreement (Bussom,

Larson, Vicars, & Ness, 1980), the widely recognized observational studies of managerial work by Mintzberg (1973) and others (Bussom, et.al., 1980) report no reliabilities. In other words, the reliabilities found in this leadership study, and even the fact that it was assessed at all, compares favorably with the reliability assessments of observational measures reported in LGD, behavioral management, and managerial activities research.

Even though there is quite high interrater agreement in this study, some of the potential shortcomings with such an assessment of reliability should be noted. For example, Mitchell (1979) suggested that the agreement percentage may be insensitive to the degree of agreement. In this study, however, each behavioral category was a nominal or "dummy" variable. It was scored by the observer as being either present ("1") or absent ("0") for each random ten minute time period. Therefore, the degree of agreement as suggested by Mitchell is not relevant. Instead, the degree of agreement is captured by measuring the amount of specific behaviors the two observers agreed were present (or absent in the case of the traditional approach) out of 12 possible behaviors within the ten-minute period (rather than the extent to which a specific behavior was present or absent).

Some other possible shortcomings of interrater agreement reliability cannot be so easily taken care of. For example, Mitchell (1979) also notes that some interrater agreement can occur by mere chance. This potential problem was minimized in this study by carefully developing, structuring and defining the behavioral categories and then thoroughly training the observers until they were very knowledgable and comfortable with the categories.

Another potential problem occurs when the observer is sensitive to and thus is more accurate when another observer joins them on an interrater reliability check (Reid, 1970). This was minimized in the study by stressing

during the training that the outside observers would be around periodically to "help" them, not "check-up" on them. After the joint observations were completed, the participant observer and outside observer chatted about the process and answered each others questions in a helpful, friendly exchange.

Finally, Spool (1978) and others have stressed that interrater agreement is sensitive only to random or unsystematic error, but systematic error can go undetected. As Thornton and Zorich (1980) note: "systematic errors that affect all raters in the same way and decrease the accuracy of observation are not evaluated by indices of interrater agreement" (p. 351). This error, of course, is the problem with any measuring instrument and must be solved by the validity of the instrument. Whether reliability or validity issues should be given precedence in observational research is still controversial (Weick, 1968). However, Byrne's (1964) advice that precedence be given to measurement reliability so that "one's experimental results are not based on the shifting sands of error variance" (p. 57) is followed here. However, it is recognized, in the final analysis, that validity of this observational system of measuring leader behavior or any other measurement instrument is the key to understanding and progress.

Besides trying to overcome some of the obvious problems of an interrater reliability assessment, this study also went beyond the traditional method of calculating the agreement percentage by two alternative, more conservative, methods that only included agreement on the actual behaviors observed. Although one could make a case for the importance of accurately observing when particular behaviors do not occur (e.g. observing that a leader does not exhibit certain behaviors may be as important as observing that she/he does exhibit certain behaviors), it is much more difficult to observe and, especially, identify/categorize behaviors that do occur. Assessments that

do not specifically look at agreement of observed behaviors may lead to misrepresented, inflated reliability estimates.

In addition to analyzing some alternative ways to calculate agreement percentage, this study also examined the various categories of observed behavior, particularly the relationship between frequency of occurrence of various behavioral categories and interrater agreement. Although certain behavioral categories had higher interrater reliabilities than others, there was no significant relationship between frequency of occurrence and interrater agreement. As one would expect, the more routine, relatively straightforward behavioral categories (e.g. exchanging routine information, processing paperwork, and staffing) had higher reliabilities than did the more complex, richer behavioral categories such as monitoring/controlling performance or motivating/reinforcing. Importantly, however, even these latter behaviors still had about two-thirds interrater agreement and behaviors such as training/developing and socializing/politicking had about three-fourths interrater agreement.

The more generalizable notion of reliability that examined interrater agreement variance contribution yielded some significant results. Remembering that the study took place in four quite different organizational settings (financial institution, manufacturing plant, state government agency and quasi-military campus police and ROTC units of a university), it was found that this did make a significant contribution to the agreement variance. On the other hand, who the outside observer was and the sequencing of the reliability checks over time did not help explain the variance in the interrater agreement. The fact that it did not make a difference who the outside rater was supports the effectiveness of the training they all received, but it is somewhat surprising that the sequencing of the checks did not matter. One would probably guess that the agreement percentage

would increase over time. The fact that it did not again could be explained by the training effectiveness of the observers. They did not seem to agree more (nor, importantly, less) over time. They were ready to go at the end of the simulated training sessions and practicing in the real setting did not affect their agreement percentage.

Overall, this study demonstrated that trained participant observers can reliably measure leader behavior in ongoing, natural settings. There was high agreement between relatively research-sophisticated outside observers (graduate students in management given observation training) and participant observers (secretaries or staff assistants with no knowledge of academically-based leadership theory/research given observation training). This finding supports the value of observer training and the use of participant observers. However, there were some practical problems that both the participant and outside observers brought out in post observation analysis.

They were asked in an open-ended, free response format what, if any, problems they had encoutered filling out the observation instrument, giving special attention to things that may have affected their accuracy. The most frequently reported problem (about a third of the 65 participant observers who answered this open-ended question) was that they were occasionally unable to observe the target manager in the assigned random time slot because he/she had left the area and was out of sight. The observers were, of course, instructed to follow prescribed procedures in such instances, but they still felt this was a problem. Another fourth of the participant observers reporting problems stated that their own work was so demanding that they found it difficult and inconvenient to take the time out to do the observations in the random ten minute time slot every hour. Remember, this problem was cited even though all their supervisors and top management had given permission to do the observatio

and they were aware that it might detract from their regular work. About 15 percent of the participant observers reporting problems said that they, and sometimes the target managers, felt the observations were too intrusive and an invasion of privacy. Other problems mentioned only a couple of times included things such as the behavior categories not being appropriate to their particular organization or department; that it was impossible to separate their feelings from their observations; and that the target manager was uncooperative, that they were allowed to see but not really hear in all cases what was going on. In addition, about 5 percent of the participant observers were negative about their experience, about 10 percent reported they encountered no problems at all and about 25 percent did not comment.

Debriefings of the outside (graduate student) observers mostly centered upon the behavioral categories. They did report that the observation training had been very useful to them; they generally were confident in their ability to categorize the observed behaviors, even in foreign settings; and, except for a few cases, everyone was cooperative and felt or reported that it was a positive experience.

In summary, the observers encountered some practical problems, but certainly not enough to negate the use of this approach to measurement of leadership behavior. With the dissatisfaction surrounding current leadership theory and research methods and the new theoretic assumptions stressing leader-environment-behavior interactionism, the search for reliable leadership measurement techniques may be found in observation techniques such as those described in this study.

References

Andrasik, F. Organizational behavior modification in business settings: A methodological and content review. Journal of Organizational Behavior Management, 1979, 2, 85-102.

Bandura, A. Social learning theory. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Barrow, J.C. Worker performance and task complexity as causal determinants of leader behavior style and flexibility. Journal of Applied Psychology, 1976, 61, 433-440.

Bass, B.M. The leaderless group discussion. Psychological Bulletin, 1954, 51, 465-492.

Bernardin, H.J. Rater training: A critique and reconceptualization. Paper presented at the meeting of the Academy of Management, Atlanta, Georgia, August 1979.

Bernardin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, 63, 301-308.

Bernardin, H.J., & Walter, C.S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Bijou, S.W., Peterson, R.F., & Ault, M.H. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1968, 1, 175-191.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Bussom, R.S., Larson, L.L., Vicars, W.M., & Ness, J.J. The nature of a police executive's work: Executive summary. Carbondale, Ill.: College of Business and Administration, Southern Illinois University, 1980.

Byrne, D. Assessing personality variables and their alteration. In P. Worchel & D. Byrne (Eds.), Personality change. New York: Wiley, 1964.

Campbell, D.T. Systematic error on the part of human links in communication systems. Information and control, 1958, 1, 334-369.

Campbell, J.P. Summary comments. In J.G. Hunt & L.L. Larson (Eds.), Leadership: The cutting edge. Carbondale, Ill.: Southern Illinois University Press, 1977.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. Managerial behavior. New York: McGraw-Hill, 1970.

Carroll, S.J., Jr., & Taylor, W.H., Jr. A study of the validity of a self-observational central-signaling method of work sampling. Personnel Psychology, 1968, 21, 359-364.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Crano, W.D., & Brewer, M.B. Principles of research in social psychology. New York: McGraw-Hill, 1973.

Cronbach, L.J., Nageswari, R., & Gleser, G. Theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 2, 137-163.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cummings, L.L. Organizational behavior in the 1980's. Decision Sciences, 1981, 12, 365-377.

Davis, T.R.V., & Luthans, F. A social learning approach to organizational behavior. Academy of Management Review, 1980, 5, 281-290.

Davis, T.R.V. & Luthans, F. Leadership reexamined: A behavioral approach. Academy of Management Review, 1979, 4, 237-248.

Delbecq, A.L., VandeVen, A.M., & Gustafson, D.H. Group techniques for program planning. Glenview, Ill.: Scott, Foresman, 1975.

Gellert, E. Systematic observation: A method in child study. Harvard Education Review, 1955, 25, 179-195.

Graen, G., & Cashman, J.G. A role-making model of leadership in formal organizations: A developmental approach. In J.G. Hunt, & L.L. Larson (Eds.), Leadership frontiers. Kent, Ohio: Comparative Administration Research Institute, Kent State University, 1975.

Greene, C.N. The reciprocal nature of influence between leader and subordinate. Journal of Applied Psychology, 1975, 60, 187-193.

Hemphill, J.K. Job descriptions for executives. Harvard Business Review, 1959, 37 (5), 55-67.

Johnson, S.M., & Bolstad, O.D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In L.A. Hamerlynck, L.C. Handy, & E.J. Mash (Eds.), Behavior change: Methodology, concepts and practice. Champaign, Ill.: Research Press, 1973.

Kelly, J. Organizational behavior. Homewood, Ill.: Irwin-Dorsey, 1969.

Kelly, M.B. A review of the observational data-collection and reliability procedures in the Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 1977, 10, 97-101.

Kerlinger, F.N. Foundations of behavioral research. New York: Holt, Rinehart, and Winston, 1973.

Kerr, S. Substitutes for leadership: Their meaning and measurements. Proceedings of the annual meeting of the American Institute for Decision Sciences. San Francisco, 1976.

Komaki, J., Barwick, K.D., & Scott, L.R. A behavioral approach to occupational safety: Pinpointing and reinforcing safe performance in a food manufacturing plant. Journal of Applied Psychology, 1978, 63, 434-445.

Komaki, J., Waddel, W.M., & Pearce, M.G. The applied behavior analysis approach and individual employees: Improving performance in two small businesses. Organizational Behavior and Human Performance, 1977, 19, 337-352.

Korman, A.K. Contingency approaches to leadership: An overview. In J.G. Hunt, & L.L. Larson (Eds.), Contingency approaches to leadership. Carbondale, Ill.: Southern Illinois University Press, 1974.

Latham, G.P., & Sari, L.M. Application of social-learning theory to training supervisors through behavioral modeling. Journal of Applied Psychology, 1979, 64, 239-246.

Latham, G.P., Wexley, K.N., & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Levine, J., & Butler, J. Lecture versus group discussion in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

Luthans, F. Leadership: A proposal for a social learning theory base and observational and functional analysis techniques to measure leader behavior. In J.G. Hunt, & L.L. Larson (Eds.), Crosscurrents in leadership. Carbondale, Ill.: Southern Illinois University Press, 1979.

Luthans, F., Paul, R., & Baker, D. An experimental analysis of the impact of contingent reinforcement on salespersons' performance behavior. Journal of Applied Psychology, 1981, 66, 314-323.

Medley, D., & Mintzel, H. Measuring classroom behavior by systematic observation. In N. Gage (Ed.), Handbook of research in teaching. Skokie, Ill.: Rand McNally, 1963.

Miner, J.B. The uncertain future of the leadership concept: An overview. In J.G. Hunt, & L.L. Larson (Eds.), Leadership frontiers. Kent, Ohio: Comparative Administration Resources Institute, Kent State University, 1975.

Mintzberg, H. The nature of managerial work. New York: Harper & Row, 1973.

Mitchell, S.K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 1979, 86, 376-390.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1978.

Penfield, R.V. Time allocation patterns and effectiveness of managers. Personnel Psychology, 1974, 27, 245-255.

Reid, J.B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Schriesheim, C.A., Bannister, B.D., & Money, W.H. Psychometric properties of the LPC scale: An extension of Rice's review. Academy of Management Review, 1979, 4, 287-290.

Schriesheim, C.A., & Kerr, S. Theories and measures of leadership: A critical appraisal of current and future directions. In J.G. Hunt, & L.L. Larson (Eds.), Leadership: The cutting edge. Carbondale, Ill.: Southern Illinois University Press, 1977.

Shook, Gerald L., Johnson, C.M., & Uhlman, W.F. The effect of response effort reduction, instructions, group and individual feedback, and reinforcement on staff performance. Journal of Organizational Behavior Management, 1978, 1, 206-215.

Spool, M.D. Training programs for observers of behavior: A review, Personnel Psychology, 1978, 31, 853-888.

Terborg, J.R., Richardson, P., & Pritchard, R.D. Person-situation effects in the prediction of performance: An investigation of ability, self-esteem, and reward contingencies. Journal of Applied Psychology, 1980, 65, 574-583.

Thornton, G.E., III, & Zorich, S. Training to improve observer accuracy. Journal of Applied Psychology, 1980, 65, 351-354.

Warmke, D.L., & Billings, R.S. Comparison of training methods for improving psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 1979, 64, 124-131.

Weick, K. Systematic observational methods. In G. Lindzey, & E. Aronson (Eds.), The handbook of social psychology, (Vol. 2). Reading, Mass.: Addison-Wesley, 1968.

Wexley, K.N., Sanders, R.E., & Yukl, G.A. Training interviews to eliminate contrast effects in employment interviews. Journal of Applied Psychology, 1973, 57, 233-236.

Table I

Reliability Data by Behavioral Category

Behavioral Category	Agreements Between Participant & Outside Observers of Behaviors Only[a]	Total Number of Observations of Behavior (Participant & Outside Observers)	Participant Observers Raw Frequency	Outside Observers Raw Frequency	% Agreement Behavioral Method I (Reliability of Participant Observers)[b]	% Agreement Behavioral Method I (Reliability of Participant and Outside Observers)[c]
Exchanging Routine Information	158	350	165	185	.958	.823
Processing Paperwork	131	285	143	142	.916	.851
Decision Making/ Problem Solving	76	174	86	88	.884	.775
Socializing/ Politicking	64	161	82	79	.780	.660
Planning/ Coordinating	65	152	77	75	.844	.747
Interacting with Outsiders	56	132	66	66	.848	.737
Monitoring/Controlling	35	94	46	48	.761	.593
Training/Developing	18	44	21	23	.857	.692
Motivating/Reinforcing	16	43	23	20	.696	.592
Staffing	17	38	18	20	.944	.809
Managing Conflict	3	8	4	4	.750	.600
Disciplining/Punishing	1	2	1	1	1.000	1.00
Other	8	27	9	18	.889	.420
Totals	648	1510	741	769	.874	.752

[a]The data in this and the other columns reflect only those behaviors that were actually observed and does not include agreement on behaviors that were not observed.

[b]This assesses the interrater reliability of the participant observer only (agreements/the observations of the participant observer) and is referred to as Behavioral Method I.

[c]This assess the interrater reliability of the participant and outside observers (agreements/total number of observations-agreements) and is referred to as Behavioral Method II.

Table II

Rank Order Correlations Between Behavioral Frequencies and Interrater Reliabilities

Method of Interrater Reliability Calculation	Rank-Order Correlation Coefficient	Significance
Traditional Method (Agreements on behaviors and nonbehaviors/total number of observations-agreements)	$r_s = -.44$	n.s.
Behavioral Method I (Agreements on behaviors/ the observations of the participant observer only)	$r_s = .32$	n.s.
Behavioral Method II (Agreements on behaviors/ total number of observations-agreements)	$r_s = .11$	n.s.

Table III

Generalizability of Interrater Reliabilities:

Hierarchical Regression Analysis

Method of Calculating Interrater Reliability[a]	Hierarchical Regression Steps	ΔR^2	Total R^2
Traditional Method	Step 1: Type of Organization	.177*	.177
	Step 2: Outside Observer	.013	.190
	Step 3: Sequencing (Time) of Reliability Check	.000	.190
Behavioral Method I	Step 1: Type of Organization	.083*	.083
	Step 2: Outside Observer	.012	.095
	Step 3: Sequencing (Time) of Reliability Check	.000	.095
Behavioral Method II	Step 1: Type of Organization	.183*	.183
	Step 2: Outside Observer	.014	.197
	Step 3: Sequencing (Time) of Reliability Check	.000	.197

[a]See Table II that explains how the three methods of interrater reliability are calculated.

*$p < .01$