REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFIT/CI/NR 88-53
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): RATING DELAY AND RATING OUTCOMES IN LABORATORY-BASED PERFORMANCE EVALUATIONS
5. TYPE OF REPORT & PERIOD COVERED: MS THESIS
6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): RICHARD S. TALLARIGO
8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS: AFIT STUDENT AT: BOWLING GREEN STATE UNIVERSITY
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:
12. REPORT DATE: 1988
13. NUMBER OF PAGES: 69

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): AFIT/NR, Wright-Patterson AFB OH 45433-6583
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report): DISTRIBUTED UNLIMITED: APPROVED FOR PUBLIC RELEASE

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

18. SUPPLEMENTARY NOTES: Approved for Public Release: IAW AFR 190-1
LYNN E. WOLAVER 18 July 88
Dean for Research and Professional Development
Air Force Institute of Technology
Wright-Patterson AFB OH 45433-6583

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

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

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Abstract

Intervals of rating delay and levels of true halo were examined for relationships with rater errors, rater accuracy (Cronbach, 1955), and convergent/discriminant validity. Relationships among these rating outcomes were also examined. Delay intervals did not affect systematic distortion (SD), convergent, or discriminant validity, but resulted in lower Differential Elevation (DEL) accuracy and higher Absolute Halo Error (AHE) after a two-day interval. Significant intercorrelations were found among measures of rater error and rater accuracy. Elevation accuracy correlated positively with Observed Halo (OH), Observed Leniency (OL) and convergent validity. DEL correlated negatively with OH and positively with Restriction of Range. Both Stereotype Accuracy and Differential Accuracy were positively related to OH and negatively related to AHE. True halo, rather than rating delay, moderated relationships between several rater errors and rater accuracy measures. It was concluded that the delay intervals studied had few strong influences on rating outcomes and that rater "error" measures could not in all cases serve as meaningful measures of rating inaccuracy. Levels of true halo in rated performances need to be considered as well.

RATING DELAY AND RATING OUTCOMES IN LABORATORY-BASED PERFORMANCE EVALUATIONS

Richard S. Tallarigo, Captain, USAF

Air Force Institute of Technology

Civilian Institutions Program

Bowling Green State University

RUNNING HEAD: RATING DELAY

Author Notes

The guidance and assistance of Dr William K. Balzer throughout this project is gratefully acknowledged. Many helpful comments and suggestions provided by Dr. Sebastiano Fisicaro are also appreciated.

Leslie Hammer's diligent help with data collection was instrumental in the timely completion of one phase of this study.

Special thanks are due Nancy Shafer of the Bowling Green State University Statistical Consulting Center for her expert data analysis programming assistance.

This paper is based on a research project by the author as part of graduate studies in Industrial/Organizational Psychology, Bowling Green State University.

Comments and conclusions in this paper are those of the author and do not represent official positions of the United States Air Force.

Abstract

Intervals of rating delay and levels of true halo were examined for relationships with rater errors, rater accuracy (Cronbach, 1955), and convergent/discriminant validity. Relationships among these rating outcomes were also examined. Delay intervals did not affect systematic distortion (SD), convergent, or discriminant validity, but resulted in lower Differential Elevation (DEL) accuracy and higher Absolute Halo Error (AHE) after a two-day interval. Significant intercorrelations were found among measures of rater error and rater accuracy. Elevation accuracy correlated positively with Observed Halo (OH), Observed Leniency (OL) and convergent validity. DEL correlated negatively with OH and positively with Restriction of Range. Both Stereotype Accuracy and Differential Accuracy were positively related to OH and negatively related to AHE. True halo, rather than rating delay, moderated relationships between several rater errors and rater accuracy measures. It was concluded that the delay intervals studied had few strong influences on rating outcomes and that rater "error" measures could not in all cases serve as meaningful measures of rating inaccuracy. Levels of true halo in rated performances need to be considered as well.

Rating Delay and Rating Outcomes in Laboratory-based Performance Evaluations

An important aspect of performance ratings is their dependence on a rater's memory for ratee work performance (Barnes-Farrell & Couture, 1984; Bernardin & Beatty, 1984; Murphy, Balzer, Lockhart, & Eisenman, 1985; Nathan & Lord, 1983). The importance of studying memory-based ratings is evident considering typical rating environments, where evaluative ratings can be made days, weeks, or months after observed performance (Landy & Farr, 1980). A number of rating outcomes could likely be affected when a rater depends upon his/her memory for a ratee's performance. These include rater errors such as systematic distortion (SD), leniency, restriction of range, and halo. Other outcomes such as interrater agreement/convergent validity and discriminant validity may also logically be affected by reliance on memory for performance. Moreover, when criterion true scores for ratee performance are available, the effects of rating delay on rating accuracy may be examined and true halo may be estimated.

Systematic Distortion (SD) is a rater error which refers to biases in memory-based ratings such that correlations between rated dimensions occur in the direction of the implicit covariance theory (ICT) of the rater (Borman, 1983; Cooper, 1981a, 1981b). ICT is an individual characteristic which describes a rater's inferences or beliefs about how performances in specific rating categories are likely to covary among ratees. ICT has its roots in implicit

personality theory research (Borman, 1983; Bruner & Tagiuri, 1954; Ilgen & Favero, 1983; Schneider, 1973). While SD has been found most frequently when ratings are made under difficult memory conditions, lack of job knowledge and/or ratee familiarity (e.g., Kozlowski & Kirsch, 1987; Kozlowski, Kirsch, and Chao, 1986) have also been associated with the presence of SD. It is not clear, however, what minimum intervals of rating delay are necessary for SD to occur when ratee familiarity and job knowledge are unknown. In studies which have reported SD effects, rating delays have ranged from one-day (Murphy & Balzer, 1986) to several weeks (Shweder, 1975), to six months (Kozlowski & Kirsch, 1987).

An outcome thought to be related to SD of ratings is halo (Cooper, 1981a). Although its conceptual definition as a failure to discriminate among rating dimensions is relatively consistent (Saal, Downey, and Lahey, 1980), its operational definitions are numerous. Pulakos, Schmitt, and Ostroff (1986) showed that, as measures of halo, the average standard deviation across rating dimensions is equivalent to the average interdimension correlation when ratings are first standardized within dimensions. They recommended the use of each rater's average observed intercorrelation among the dimensions as a measure of halo. As Pulakos et al. (1986) pointed out, this measure of halo will always be perfectly correlated with halo error, with halo error defined as the difference between observed and true dimension intercorrelations. True dimension intercorrelations are computed using criterion true scores for each ratee on each of several performance dimensions. True halo is a

constant when all raters view the same ratees. It is not clear, though, what general relationship may hold between observed halo and halo error when raters view different sets of ratees, and the relationships are examined across all raters. Furthermore, because halo errors may be either positive or negative, another meaningful measure of halo is the absolute value of halo error, or absolute halo error (AHE; Fisicaro, 1987). AHE provides an overall index of the amount of halo error present in ratings independent of direction of the error. Halo Error, on the other hand, provides information about both the direction (positive or negative) and the intensity of the halo error. Systematic Distortion has been suggested as one source of halo error in ratings (Cooper, 1981a).

The rater error measures of leniency and restriction of range have also been commonly applied to performance ratings. Leniency has usually been defined as mean ratings above the scale midpoint. The central idea of leniency, according to Saal, Downey, and Lahey (1980), is that ratings are consistently too high or too low (severity). Restriction of range is commonly defined as the average of the standard deviations of ratings across ratees (Saal, Downey, and Lahey, 1980). Where leniency reflects a level effect, range restriction reflects a rater's ability to discriminate among different ratees.

Additional rating outcomes relevant to performance ratings include convergent and discriminant validity. Convergent validity indicates the overall amount of agreement on ratees across raters and dimensions; and discriminant validity indicates the extent to

which raters distinguish among performance dimensions (Kavanagh, MacKinney, and Wolins, 1971).

The rater accuracy measures examined in the present study are those described by Cronbach (1955) and used frequently in performance evaluation research for which criterion true scores are available. Cronbach (1955) demonstrated that a measure of the overall distance from criteria when a rater evaluates multiple ratees on multiple dimensions consists of four components: Elevation (EL), Differential Elevation (DEL), Stereotype Accuracy (SA), and Differential Accuracy (DA). In general terms, EL is a measure of the closeness of a rater's grand mean of ratings to the grand mean of criterion scores. DEL reflects how closely a rater's overall ranking of ratees corresponds to the ranking based on criterion scores. SA indicates how accurately a rater ranks the dimensions across ratees. Finally, DA indicates how accurately a rater can distinguish among ratees within each dimension. Operational definitions of all the rating outcomes used in this study are presented below in the Method section.

Because SD is not a commonly measured rating outcome, background to the SD concept is presented below in some detail. After this background, hypotheses are presented which predict changes among the various rating quality outcomes over intervals of rating delay.

What is Systematic Distortion?

Systematic distortion occurs when ratings conform more to the rater's preconceptions of "what goes with what" than to what

actually covaries among ratees' behavior. The effects of rater preconceptions about how traits covary has had a long research history. Early theorists in social psychology found evidence of halo (Thorndike, 1920, p. 29) and logical error (Newcomb, 1931) in interpersonal ratings. Building on this earlier work, Bruner and Tagiuri (1954) and Schneider (1973) used the term "implicit personality theory" to describe the inferences a perceiver makes about the attributes of others. Similar descriptions are found in person perception research. For example, raters have been found to make "implicit interpretations" of others' traits which may affect the index of stereotype accuracy (Cline, 1964; Cronbach, 1955; Gage & Cronbach, 1955). Research in the area of clinical psychology and personality has also found that ratings of others may be influenced by illusory correlations (Chapman & Chapman, 1967, 1969), correlational biases of the raters (Berman & Kenny, 1976; Nisbett & Wilson, 1977), or the rated likelihood of co-occurrence of personality traits (Hakel, 1974).

The systematic distortion hypothesis (SDH) predicts that traits which are semantically or conceptually similar will be recalled as if they covaried (Shweder & D'Andrade, 1979). In numerous studies, when judgmental ratings were made after varying retention intervals, the averaged covariance structure of the memory-based ratings was more similar to the raters' averaged pre-existing covariance beliefs than to the actual covariance structures of the target behaviors (D'Andrade, 1974; Shweder, 1975, 1977, 1980, 1982, 1983; Shweder & D'Andrade, 1979, 1980; cf. Lamiell, 1980).

Similar results were found in studies of leadership behavior ratings (Lord, Foti, & DeVader, 1984). Subjects have been found to distort leadership ratings to be consistent with the memory schemata manipulated by the experimenters (Phillips & Lord, 1982). In these studies, SD was explained in terms of the schematic memory of the rater (Foti, Fraser, & Lord, 1982; Nathan & Lord, 1983; Phillips & Lord, 1981).

Systematic Disortion in the Performance Appraisal Process

Cooper (1981a, 1981b) and Borman (1983) presented theoretical and empirical evidence suggesting that SD occurs in performance ratings. Cooper (1981b) applied Shweder and D'Andrade's (1980) paradigm to study SD in job performance ratings. His subjects directly rated the interdimension similarities of performance dimensions. These interdimension similarity ratings were averaged across subjects and resulted in a single interdimension similarity matrix. For two studies, the results showed that the similarity matrices correlated significantly with rated behavior matrices. In a third study, Cooper used videotapes and true scores developed by Borman (cited in Cooper, 1981b). Although the rated behavior matrix again correlated (r = .55) with the similarity matrix, it correlated even more strongly with the true score interdimension matrix (r = .89). This pattern did not follow the typical SD pattern in which memory-based ratings are more highly related to implicit theory beliefs than to criterion ratings.

One reason for Cooper's (1981b) failure to detect SD may have been that the rated behavior matrix consisted of ratings made immediately after viewing each videotape, whereas a presumed prerequisite for SD is a significant delay between observation and rating. Furthermore, Kozlowski and Kirsch (1987) suggested that the use of criterion true score/videotape methods are inappropriate for studying SD because the pooled expert judgments from which criterion scores are derived may be subject to the same cognitive distortion processes as the observed ratings. This may result from inconsistency among studies in how criterion true scores are developed (Sulsky & Balzer, 1987). As discussed below, however, the processes used by Borman et al. (1978) to develop criterion true scores are designed to minimize memory demands and job and ratee unfamiliarity on the part of the expert raters. True score estimates derived from pooled expert ratings thus are likely to minimize the influence of shared stereotypes.

Cooper (1981b) recommended that further investigations of systematic distortion in memory-based performance ratings should elicit similarity and performance rating matrices from each rater, rather than single matrices based upon group averaged ratings. Such an individual-level approach to studying systematic distortion was applied by Kozlowski and Kirsch (1987). The present study applied both an individual and group-level analysis of SD.

Borman (1983) reviewed implications of implicit personality theory for performance ratings by reanalyzing data from previously published studies (Borgatta, Cottrell, & Mann, 1958; Mann, 1959;

cited in Borman, 1983). He concluded that "both inflation in correlations and distortion towards semantic similarity occur in personality and behavior rating, at least under the conditions studied" (p. 133). Borman (1983) offered two major criticisms of SD research. First, using different raters to generate the similarity, criterion, and memory-based performance ratings produces differing frames-of-reference in interpreting the rating dimensions. This problem is reduced greatly if the same raters provide all ratings using standardized dimension definitions. The second criticism was concerned with developing the true or criterion matrix of ratings. The Shweder and D'Andrade (1980) method was to use on-line (i.e., during the observation process) behavior frequency ratings or judgmental ratings collected with minimal time-delay between observation and rating. As research has demonstrated, such behavior frequency and judgmental ratings are subject to the same impression-based biases as memory-based ratings (Murphy et al., 1982).

With the true score methods developed by Borman et al. (1978) and others (e.g., Murphy, Balzer, Kellam, & Armstrong, 1984), the validity of a criterion matrix may be substantially enhanced. True scores permit the estimation of true halo (Cooper, 1981a), the average intercorrelation among the true scores on the performance categories. Cooper (1981a, 1981b) and Borman (1983) discussed the importance of true halo when studying distortion in ratings of job performance. They argued that jobs tend to represent more homogeneous behaviors than personality traits or leadership behaviors. Essentially, this implies that the true halo of job

performance categories is likely to be somewhat high. Estimates of true job performance levels are useful for measuring (a) the accuracy of ratings, (b) levels of halo error, and (c) the relative influence of implicit covariance theories versus true levels of performance on actual performance ratings.

The Borman et al. (1978) method for constructing criterion true scores consisted of procedures to validate expert ratings of videotaped performances. Essentially, experts estimated means, standard deviations, and interdimension correlations of job performance on behaviorally-anchored rating scales. Based on these expert estimates, "intended" true scores were established, scripts were written, and performance videotaped for each ratee. Final true scores were then obtained from experts who studied the videotapes and assigned performance evaluation ratings. These final true scores were validated by (a) correlating them with the intended true scores, (b) measuring interrater agreement among the experts, and (c) analysis of convergent and discriminant validity in which the ratee main effect indicated convergent validity and the ratee x dimension interaction indicatee discriminant validity. If such statistical validation is acceptable (in terms of interrater agreement, convergent and discriminant validity), one then uses the means of the expert ratings as criterion true scores (Borman et al., 1978) in the computation of accuracy scores (e.g, EL, DEL, SA and DA; Cline, 1964; Cronbach, 1955).

Measuring Systematic Distortion

The usual test for SD is based upon comparisons among a co-occurrence (or ICT) matrix, a criterion intercorrelation matrix, and an intercorrelation matrix of the rated performance categories. The typical comparison indicates a higher correlation between the co-occurrence matrix and the performance ratings matrix than between the performance ratings matrix and the criterion matrix. As depicted in Figure 1, the correlation between the co-occurrence matrix and the performance ratings matrix represents the extent to which the correlation matrix of performance category ratings covaries with implicit theories about those category interrelations.

Insert Figure 1 about here

This correlation can be denoted as a Systematic Distortion Index (SDI). Moreover, the correlation between the performance rating intercorrelations and the criterion score intercorrelations can be denoted as an index of Correlational Structure Accuracy (COSTAC). If SDI exceeds COSTAC, the result is interpreted to mean that performance ratings are more similar to implicit notions of rating category covariance than to the actual covariance in the categories.

Are Systematic Distortion, Halo and Accuracy Related?

Recent empirical evidence suggests a link between SD, halo and rating accuracy. Murphy and Balzer (1986) found that one-day delayed performance ratings contained higher mean interdimension correlations and higher SA and DA than immediate ratings. Cooper (1981a, 1981b), however, proposed that inflated correlational

structures of memory-based ratings contain illusory halo - - a source of inaccuracy (Cooper, 1981a, 1981b; Thorndike, 1920). Murphy and Balzer (1986) explained the increased accuracy within memory-based ratings in terms of raters' reliance on accurate schematic memory via the SD process.

Kozlowski and Kirsch (1987) found that their SDI measure (i.e., conceptual similarity-rating covariation) significantly and positively correlated with observed halo (the "standardized" average variance measure); and that halo positively correlated with DA and negatively with SA. Moreover, ratee familiarity appeared to influence the direction of significant correlations of SDI with SA and DA.

Cooper (1981a, 1981b) had suggested, in discussing the paradoxical weak positive correlations between halo and accuracy reported in a number of studies, that some raters use their implicit covariance matrix as a heuristic which aids their accuracy (Cooper, 1981a, p. 239). While such may have been the case in the two studies reported above, others have concluded on the basis of the empirical literature that commonly used rater error measures have little or no relationship to rating accuracy (Becker & Cardy, 1986). Of course, the typical usage of rater error measures, as their name implies, presumes a negative relationship with rating accuracy.

Conclusions and Research Objectives

Rating delay intervals have been associated with the rater error of Systematic Distortion (SD), in which performance evaluations are influenced by raters' presuppositions regarding

rating dimension covariance. The SD effect, in turn, has been associated both theoretically and empirically with increased halo. And finally, halo has been associated with increased accuracy in some studies. The primary objective of the present study is to investigate the effects of rating delay on several rater errors, rater accuracy, convergent and discriminant validity, and on relationships among these rating outcomes.

Research Hypotheses

Effects of Rating Delay

1a. Rater error measures (halo, leniency, restriction of range) should increase with increased rating delay intervals.

1b. Rating accuracy (EL, DEL, SA, and DA) should decrease (i.e., the distance scores will increase) as a function of rating delay.

1c. Systematic Distortion will increase as the delay between observation and rating increases. The systematic distortion index (SDI) and the difference between SDI and Correlational Structure Accuracy (COSTAC) should both increase with increased rating delay.

1d. Convergent validity and discriminant validity will decrease with increased rating delay.

Relationships among rating outcomes.

2a. Rater error measures will be negatively related to rating accuracy.

2b. Halo measures will be positively correlated with the Systematic Distortion Index (SDI).

Effects of rating delay on the relationships among rating outcomes.

No hypotheses are made regarding the effects of rating delay on the relationships among rating outcomes.

Method

Stimulus Materials and True Scores

Videotapes. Eight videotapes of classroom lectures produced and used in prior research (Murphy & Balzer, 1981, 1986; Murphy et al., 1984; Murphy et al., 1982) were used in the present study. Four drama students role-playing graduate students in psychology were videotaped with each presenting two five to seven minute lectures on the topics of Self-fulfilling Prophecies and Crowding and Stress. As with the Borman et al. (1978) tapes, predetermined varying levels of effectiveness were scripted into each lecture. The eight lectures represented varying levels of clarity and organization (good or bad), presentation style (dynamic or hesitant), and responsiveness to questions (responsive or evasive, Murphy et al., 1984).

Evaluation Rating Scale. The performance rating scale developed by Murphy and colleagues in the development of the videotapes was used in this study. This scale consisted of eight performance dimensions (e.g., Thoroughness of Preparation, Grasp of Material) which are rated on a scale from 1 (Very Bad) to 5 (Very Good). In the present study, each rating form included a photograph of the ratee along with the lecture topic as a means of identifying the ratees for the subjects/raters. Coefficient alpha for this scale in the present study was .94.

Performance Category Co-Occurrence Rating Form. A Performance Category Co-Occurrence Rating Form was developed to measure raters' implicit covariance theories regarding typical classroom lecturer performance. This rating format was similar to those typically used in multidimensional scaling, implicit personality theory, and systematic distortion research (Cooper, 1981b; Schneider, 1973; Schultz & Siegel, 1964; Shweder, 1975). Each subject rated, on a scale of 0 (Not Likely to Co-Occur) to 7 (Very Likely to Co-Occur), the extent to which similar levels of performance on all 56 possible pairs of the eight dimensions are likely to be found together. Each subject's implicit covariance theory was thus defined in terms of the rated likelihood that paired performance categories covary in the general college classroom lecturer population.

Criterion scores. Murphy et al. (1984) obtained criterion scores on the performance evaluation scale rating dimensions for each of the eight videotaped performances in a manner following Borman et al. (1978). The intraclass coefficient for the ratee main effect in a rater by ratee by dimension analysis of variance was .70, and the ratee x dimension interaction intraclass coefficient was .47. As measures of convergent and discriminant validity, respectively, these values compare well with studies reporting rater x ratee x dimension analyses of performance ratings (cf. Borman, 1978; Kavanagh, MacKinney, & Wolins, 1972; Lee, Malone, & Greco, 1981). The mean expert ratings also converged with the intended true scores as indicated by a median correlation (across tapes) of .84 (Murphy & Balzer, 1981).

Experimental Manipulation

Four experimental delay conditions were created to examine the effects of delay between observing and rating performance on the strength of systematic distortion. The four delay groups included an immediate-rating group (D0), whose members rated the tapes immediately after viewing the four tapes; a one-day delay group (D1), whose members returned the following day to rate the tapes; a two-day delay group (D2), whose members returned after two days to rate the tapes; and a seven-day delay group (D7) whose members returned after seven days.

Subjects

Three-hundred and thirty-seven subjects completed the study. These consisted of 333 introductory psychology undergraduates who participated in exchange for course credit, and four volunteer graduate students in Business Administration. Fifty-two of these subjects participated under slightly different conditions than the remainder of the subjects: Instead of viewing a random sample of four of the eight videotapes, they viewed a specific set of four tapes. Because the analysis below utilizes measures of true halo, only a random sample of eight subjects from these 52 are utilized for this analysis as a means of equating the delay groups on the true halo levels contained in the stimulus videotapes. Of the 293 subjects in the present sample, the mean age was 19, ranging from 17 to 47 years of age; approximately two-thirds of the sample was female.

Procedures

Students who volunteered for this study were informed that they may be asked to return to provide memory-based ratings. Upon arrival for the initial session, subjects were given a description of the general nature of the study. Subjects participated in small groups of 1-4 and were randomly assigned as a group to one of the four delay conditions. Following an explanation of each of the performance rating dimensions, subjects completed the co-occurrence ratings.

All subjects (with the exceptions noted above) were then shown a randomly-selected sample of four of the eight videotapes, subject to the constraint that only one lecture by each actor was viewed. This procedure resulted in 14 (out of 16 possible) tape combinations viewed by raters. While viewing the tapes, raters were not permitted to take notes. After viewing the tapes, immediate-rating subjects (D0) were asked to provide performance ratings for each lecturer. Subjects assigned to the delayed rating conditions (D1, D2, and D7) were told where and when to return to make their memory-based performance ratings. These delay condition subjects were given written instructions to return to a specified room after either a 24 hour interval (D1 condition) , a 48 hour interval (D2 condition), or a one-week interval (D7 condition). Subjects were asked to return to make their ratings at an hour as close as possible to the desired interval. When the delayed condition subjects returned for the second part of the experiment, they were provided with a rating packet (which included instructions and materials for the rating task), completed their memory-based ratings, and returned their

ratings to a designated location. All subjects were provided with either an oral (for the DO subjects) or written (for delayed-condition subjects) debriefing of the experiment.

Measures

Systematic Distortion Index. Each subject's mean Performance Category Co-Occurrence Rating for each dimension pair was paired with the obtained correlation between performance ratings on each dimension pair. This resulted in a single correlation, the Systematic Distortion Index (SDI) for each subject/rater. SDI represents the degree of association between logical presuppositions about the covariance structure of the performance dimensions and a rater's obtained covariance structure of the performance dimensions.

Rating accuracy. The algebraic difference score formulas of Cronbach (1955) were used in this study. Because they are difference scores, higher values indicate less accuracy. Although presented in their squared form, the square roots of each of the components was used in the analyses.

Elevation $(EL^2) = (\bar{r}.. - \bar{t}..)^2$,

Differential Elevation $(DEL^2) = 1/n \sum [(\bar{r}_i. - \bar{r}..) - (\bar{t}_j. - \bar{t}..)]^2$,

Stereotype Accuracy $(SA^2) = 1/k \sum [(\bar{r}._j - \bar{r}..) - (\bar{t}._j - \bar{t}..)]^2$,

Differential Accuracy =

$$1/kn \sum \sum [(r_{ij} - \bar{r}_i. - \bar{r}._j + \bar{r}..) - (t_{ij} - \bar{t}_i. - \bar{t}._j + \bar{t}.)]^2,$$

where r_{ij} and t_{ij} = rating and true score for ratee i on dimension j; $\bar{r}_{i.}$ and $\bar{t}_{i.}$ = mean rating and mean true score for ratee i; $\bar{r}_{.j}$ and $\bar{t}_{.j}$ = mean rating and mean true score for dimension j; and $\bar{r}_{..}$ and $\bar{t}_{..}$ = mean rating and mean true score over all ratees and dimensions.

Rater errors. Observed Halo (OH) was computed as the median interdimension correlation of each rater's performance ratings, across the four ratee videotapes viewed by each rater. Halo Error (HE) was computed for each rater by subtracting the median true score interdimension correlation for the tapes viewed by a rater from the rater's OH. Absolute Halo Error (AHE) is simply the absolute value of HE (Fisicaro, 1987). Observed Leniency (OL) was computed as the difference obtained by subtracting the scale midpoint (3.0) from the mean ratings within ratees, averaged across ratees. Leniency Error (LE) utilized the mean true score instead of the scale midpoint. (Absolute Leniency Error is equal to Elevation Accuracy, and was therefore not computed.) Restriction of Range (ROR) was computed as the standard deviation of ratings across ratees, averaged across dimensions. Higher values indicate greater variability in ratings across ratees and less restriction in range.

Convergent/Discriminant Validity. Procedures described by Kavanagh, MacKinney, and Wolins (1971) were used to obtain the intraclass correlation coefficients associated with the ratee main effect (convergent validity), ratee x dimension interaction (discriminant validity), and the rater x ratee interaction. The absence of a significant ratee main effect may also indicate the

presence of range restriction. The rater x ratee interaction term has received a number of conceptual labels. The absence of a significant rater x ratee interaction has been interpreted as evidence for interrater agreement (Saal, Downey, & Lahey, 1980). The presence of a significant interaction has been interpreted as evidence for halo (Saal et al., 1980) or relative halo (Johnson, 1963; Willingham & Jones, 1958). In either case, it is consistently viewed as an undesirable source of variance. For purposes of this study, the rater x ratee interaction is referred to as relative halo, and reflects the tendency for different raters to rate ratees differently. Because not all subjects viewed the same ratee tapes, these analyses were performed within groups of raters who viewed the same tapes. The sampling procedures involved in assigning ratees to raters resulted in 13 unique tape combinations viewed by multiple raters. One tape combination was viewed by only one rater and thus was not included in the convergent/discriminant validity analyses.

Results

Performance Category Co-Occurrence Ratings

With eight performance dimensions, there are two sets of 28 pairs of performance dimension co-occurrence ratings. Each set is a symmetrical opposite of the other. The mean of each of the symmetrical opposites was taken as the implicit theory measure for each pair of performance dimensions, resulting in 28 implicit theory ratings for each subject. Table 1 presents the overall means on each of these 28 performance dimension pairings. These scores suggested

that the subjects's preconceived notions about the covariance of the performance categories were rather restricted.

Insert Table 1 about here

The values ranged from 3.11 (Rapport with Audience / Responsiveness to Questions) to 6.07 (Speaking Ability / Organization and Clarity) with an overall mean of 4.89 and standard deviation of .66. Considering the eight point (0 - 7) scale used, the size of the standard deviations associated with each covariance estimate (less than 2.0) suggested good agreement among the subjects in those estimates. The overall mean indicated that as a group, the subjects thought the "likelihood of co-occurrence" of the performance dimensions was just slightly above the midpoint of the 0 - 7 scale of co-occurrence likelihood. In addition, the mean co-occurrence ratings for each of the 28 pairs was calculated for each rating delay group in order to assess the equivalence of the implicit theories across groups. Table 2 presents the group intercorrelations of these mean ratings. The results suggested a very high degree of agreement among the groups in how they viewed the likely co-occurrence of the performance categories.

Insert Table 2 about here

Effects of Rating Delays on Rating Outcomes

Table 3 includes descriptive statistics for all rating outcomes (except convergent and discriminant validities) for the combined and separate rating delay groups.

Insert Table 3 about here

A number of observations are noteworthy in Table 3. First, the overall and within-group mean levels of True Halo are extremely high. In Pearson correlation form, the overall True Halo was approximately .95. Second, note the difference in interpretation of halo and leniency depending on whether true scores are considered. For example, as can be seen in Table 3, the Observed Leniency (OL) measure indicated that raters, on average, were neither lenient nor severe in their ratings, using the scale midpoint as a criterion. But when criterion true scores were used as criteria, the majority of raters were severe in their ratings and the percentage of raters exhibiting leniency error dropped from approximately one-half to one-quarter. In a similar manner, Observed Halo was present at high levels. According to traditional interpretation, raters failed to discriminate sufficiently among the dimensions. But Halo Error levels indicated that raters discriminated too much. That is, Halo Error was consistently negative. The majority of subjects (88%) exhibited negative halo error, i.e., OH lower than true halo. Thus, one can arrive at differing conclusions regarding rating quality depending on whether or not criterion true scores are considered in rating quality indices. Third, the pattern of change in the means

across the rating delay intervals is not clearly indicative of systematic change in one direction or another.

Rater errors and rater accuracy. Differences attributable to the four intervals of rating delay were examined with a one way multivariate analysis of variance (MANOVA). This test determined if, for a linear combination of all 11 rating outcomes, significant delay group differences existed. Table 4 lists the results of the one-way MANOVA and the follow-up univariate F tests.

Insert Table 4 about here

The overall effects of the four intervals of rating delay were significant (Wilks' Λ = .75, $p < .0001$). Among the dependant measures, post hoc univariate procedures determined that only DEL accuracy and AHE significanty changed over time. Scheffe tests were performed to determine which contrasts among group means were significantly different for these two rating outcomes. As noted above, inspection of the means (Table 3) indicated clearly that, for these two dependent measures, means did not systematically increase or decrease as a simple linear function of increased rating delay. The significant pairwise and complex contrasts in Table 5 suggested that DEL accuracy decreased and AHE increased with increasing rating delay after a two-day delay in ratings.

Insert Table 5 about here

Note, however, that the contrast between the immediate and one-day ratings for both DEL and AHE were not significant. Thus, it would not be proper to conclude that DEL accuracy increased and AHE decreased after a 24 hour delay. The contrast between the average of the immediate and one-day delay groups versus the two-day and seven-day groups was non-significant for both DEL ($F(3, 289) = 1.82$) and AHE ($F(3, 289) = 2.14$), indicating that a simple linear increase in retention interval cannot explain the mean differences. The failure of this contrast may also have to do with the conservative nature of the Scheffe test as with a possible curvilinear relationship between rating delay interval and rating outcomes. When only one-day to seven-day intervals are considered, however, the significant contrasts indicated less accuracy and greater absolute halo error with increasing retention intervals. Thus, hypotheses 1a and 1b are partially supported: One rater error measure (Absolute Halo Error) increased with increasing rating delay intervals, but only after 48 hours; and one rater accuracy measure (Differential Elevation) decreased with increasing rating delay intervals after 48 hours.

Systematic distortion. The multivariate results and the results depicted in Table 6

Insert Table 6 about here

---

failed to confirm Hypotheses 1c (that SD will increase over time). This hypothesis was tested at both the individual and group levels of analysis. Even though the level of SDI (the correspondence

between co-occurrence ratings and performance rating covariance) did not increase across the retention interval groups (see Table 4), it was necessary to examine its relative distance from COSTAC (the correlation between performance rating covariance and true score covariance (or Correlational Structure Accuracy; see Figure 1). Table 6 lists the matrix intercorrelational data as referenced in Figure 1) in testing the SDH across each delay group in each level of analysis. For individual-level analysis, a SDI and COSTAC score was computed for each rater. The means of these scores are presented in the table at the individual-level of analysis. For the group-level analysis, a single SDI and COSTAC score was computed for each rating delay group based on group mean co-occurrence ratings and the group means of the true score and rating intercategory correlations. Recall that systematic distortion is indicated when SDI exceeds COSTAC; and this increment in SDI should expand over time according to the SDH.

As the Table 6 data indicate, however, the increment between the mean SDI and COSTAC scores did not increase with increasing rating delay within either level of analysis as one would expect if SD in the direction of implicit theories were occurring. None of the differences across groups were statistically significant (all z's < 1.96); and the median SDI score across all analyses (.18) was less than the median COSTAC score (.26). Thus, Hypothesis 1c could not be supported: Systematic Distortion did not increase as rating delay increased from immediate to one-week intervals. Figure 2 depicts the pattern of the SDI - COSTAC values across rating delay conditions

for both individual- and group-level analyses. The lack of a systematic increase in SD is readily apparent.

Insert Figure 2 about here

Convergent and discriminant validity. Table 7 lists the multi-trait multi-rater analyses of variance, collapsed across delay groups, for each of the 13 ratee combinations employed in the present study to which multiple raters were exposed.

Insert Table 7 about here

Table 7 includes the variance components and intraclass indices for the ratee main effect (convergent validity/interrater agreement), ratee x dimension interaction (discriminant validity), and rater x ratee interaction (relative halo). As described in Kavanagh et al. (1971), each intraclass coefficient was converted to Fisher's z, averaged, then converted back to Pearson's correlation coefficient. The average intraclass coefficient for the ratee main effect, across the 13 ratee combinations, was .54. The average discriminant validity index was only .08. A large relative halo effect was indicated by the average intraclass correlation for the rater x ratee effect of .49. Table 8 lists the mean intraclass coefficients associated with each of the three sources of variance in each of the four rating delay groups. In all cases, the convergent validity indices were greater than the relative halo indices; and the

discriminant validity indices were consistently and extremely low. A one-way MANOVA indicated no significant overall rating delay effect on any of the variance sources (Wilks' Λ = .74; $F(9, 80.5) = 1.18$, $p > .10$). Discriminant validity indices systematically decreased with increasing rating delay, but the changes were not statistically significant.

Insert Table 8 about here

Relationships between True Halo and Rating Outcomes

True halo, rater errors and rater accuracy. Table 9 lists the intercorrelations among the rating outcomes and includes correlations with the true halo levels associated with the 14 ratee combinations. In addition, correlations are provided for the combined and for each delay group.

Insert Table 9 about here

(It should be noted that the signs on the correlations with the distance accuracy scores and with ROR in Tables 9, 10 and 11 have been reversed to reflect relationships with accuracy and lack of discriminability, respectively.) True Halo (TH) was not, in general, significantly correlated with rater accuracy measures. In the combined sample, only EL was related to TH, and in a positive direction. Among the delay groups, when a significant relationship was detected between TH and a rater accuracy score, the relationship

was positive and occurred only for EL and DEL accuracy. More consistent and strong relationships were found between TH and OH, AHE, HE, and ROR. Significant but weak relationships were detected with LE. As true halo increased, raters apparently increased the observed halo in their ratings, but not sufficiently so. As a result, the AHE also increased. It should be noted that HE also increased with TH. The negative sign to the HE/TH correlation means that the direction of HE "increases" in a negative direction, with raters increasingly underestimating TH as TH increases. In general, as TH increased, raters were more likely to exhibit halo. Paradoxically, though, raters tended to increase their discrimination among the ratees with increasing TH, as indicated by the negative relationship between TH and ROR.

True halo and multi-trait multi-rater (MTMR) indices. Table 10 presents intercorrelations among the multi-trait multi-rater (MTMR) analysis indices and mean accuracy and rater error outcomes for those analysis subgroups.

Insert Table 10 about here

Interrater agreement increased with increasing true halo in the tapes. The correlation between true halo and the intraclass coefficient for the ratee main effect was .75, ($df = 11$, $p < .01$). Although the relative halo indices did not significantly covary with true halo, it is noteworthy that, upon inspection of the intraclass coefficients in Table 7, that for the six analyses at the "lower"

end of the true halo range, relative halo exceeded convergent validity. The middle analysis (true halo 1.798) resulted in convergent validity and relative halo bias sharing equal amounts of variance. For the six analyses on the upper end of the true halo range, convergent validity exceeded relative halo bias. Table 10 also suggests that relative halo cannot serve as a surrogate measure for observed halo. The use of convergent validity as a measure of range restriction, however, appears worthy of support. The significant negative relationship suggested that higher levels of convergent validity may indicate reduced restriction of range. These analyses also indicated that, with two exceptions, there was no consistent relationship between MTMR outcomes and rater accuracy outcomes. Convergent validity correlated positively with Elevation; and Relative Halo correlated negatively with Differential Elevation.

Relationships between Rater Error and Rater Accuracy

Halo, leniency, restriciton of range and accuracy. The data in Table 9 suggest that there was no consistently negative or positive relationship between rater error and rater accuracy. These relationships depended upon the particular measures employed. In addition, the direction of the significant relationships included both expected (i.e., negative) and paradoxical (i.e., positive) outcomes. Analyses below will consider the potential moderating influences of rating delay interval and level of true halo on relationships between rater error and rater accuracy. Inspection of Table 9 at the combined group level, however, indicated that the three halo measures, two leniency measures, and restriction of range

each significantly related to one or more of the accuracy component scores.

Observed Halo (OH) was positively related to EL, SA, and DA, and negatively related to DEL. Halo Error (HE) followed a similar pattern, except that the correlation with EL was not significant. One reason for this similarity in pattern is likely due to the fact that when all raters evaluate the same ratees, HE and OH are perfectly correlated because true halo is a constant (Pulakos, Schmitt, & Ostroff, 1986). In the present study, for example, the correlation between OH and HE was 1.00 within each unique tape combination. In the present study, however, true halo is a variable. Even so, Table 9 indicates that OH and HE correlated positively and strongly. Absolute Halo Error (AHE), as opposed to OH and HE, served more appropriately as a rater error measure. AHE was negatively related to SA and DA but did not correlate significantly with EL or DEL.

As with the relationship between OH and HE, the relationship between OL and LE was highly positive even though the true score criterion (i.e., the ratee true score mean) was a variable rather than a constant. Moreover, both OL and LE correlated positively with EL. The more leniency raters exhibited, the more their leniency aided their EL accuracy. This was no doubt due to the fact that most rater exhibited severity with respect to the true score means. Thus, the more positive the LE scores, the closer they approximated the true scores, resulting in a positive correlation.

Finally, Restriction of Range (ROR) correlated negatively (as expected) with EL accuracy but positively with DEL. In failing to distinguish among ratees, raters were also less able to accurately assess the overall level of ratee performance (EL, LE). The positive correlation between ROR and DEL was surprising, indicating that by discriminating less among ratees, raters also enhanced their ability to accurately rank order the ratees. In general, hypothesis 2a must be rejected: a general negative correlation does not exist between rater error and rater accuracy.

Halo and systematic distortion. Table 9 clearly indicates the extremely low correlations between the systematic distortion index (SDI) and all other measures, including all of the halo measures. Hypothesis 2b cannot be supported: SD cannot be considered as a likely correlate of halo in the present data. A more likely influence on halo may be true halo, which correlated significantly with Observed Halo, Halo Error, and Absolute Halo Error.

Rating Delay and True Halo as Moderators of Rater Error-Rater Accuracy Relationships

As mentioned above, the use of procedures in the present study in which four videotapes were sampled from a population of eight tapes provided an opportunity to study the true halo in the stimulus performances as a variable. Fourteen of the sixteen possible combinations of four tapes were utilized in the present study. Each combination was characterized by a unique true halo value. The mean true halo values for each delay group and for the combined group are found in Table 3. These values did not differ significantly across

delay groups ($F(3, 289) = .62$, $p > .10$). In Fisher z form, the true halo in the videotapes ranged from .429 to 2.58 in each rating delay group. The 14 values of true halo were negatively skewed (skewness = -.85, n = 293, $p < .01$). However, an analysis was made of the extent to which both the rating delay interval and the level of true halo affected the functional relationships between OH and rating accuracy, ROR and rating accuracy, and OL and rating accuracy. This was done by (a) hierarchical moderated regression to determine if rating delay and true halo moderated the rater error/rater accuracy relationship (see Table 11), and (b) computing the correlations within levels of true halo for those variables which the regression analysis suggested significant moderator effects were present (see Table 12).

Insert Tables 11 and 12 about here

For each accuracy component, a separate series of hierarchical regression equations were contructed to test increments in overall R-squareds associated with three interaction terms: e.g., OH x DELG (to test the moderating influence of rating delay), OH x TH (to test the moderating influence of True Halo), and OH x TH x DELG (to test the joint moderating influences of rating delay and True Halo). A note of caution must be included here regarding this analysis in two respects. First, because of the number of dependent measures and the number of regression analyses performed in Table 11, significant increments could be expected by chance at a rate in excess of α =

.05. Therefore, in order to reduce the experimentwise error rate, the increments were tested at α = .0125. Second, estimates of the three way interaction are likely to be unstable. With 14 levels of True Halo and four rating delay groups, there are 56 True Halo/Delay group conditions. Twenty-five of these cells had fewer than four observations. Thus, interpretation of the three-way interaction is not likely to be reliable due to the small n's within some of the TH x DELG cells.

The analyses in Table 11 suggested, however, that rating delay did not act as an important moderator of rater error/rater accuracy relationships under the conditions examined. None of the Rater Error x Delay Group interactions contributed significantly to explained variance over the main effects models. True Halo, on the other hand, did appear to moderate some of the relationships. The rater error x True Halo interaction increments were significant in predicting DEL from both OH and from ROR; in predicting EL accuracy from OL; and in predicting DA from OH. Table 12 depicts how the correlations associated with these regressions change as a function of level of True Halo. For the OL/EL relationship, the moderating influence of True Halo is not at all easy to decipher. The relationship between OL and EL is high at both ends of the True Halo range, but low for a few levels in the middle and at the upper extreme. The interpretation for the other three moderated relationships appeared more direct. Restriction of Range was positively correlated with Differential Elevation only at the lowest levels of True Halo, and was not significantly correlated at other levels. Observed Halo was

positively correlated with Differential Accuracy at the highest levels of True Halo, and for the most part was not significantly correlated at other levels.

Discussion

Effects of Rating Delay

The minimal effects of rating delay on rater errors and rater accuracy added to an already clouded picture given the previous research which has found increased accuracy over delay periods of one-day (Murphy & Balzer, 1986), decreased accuracy over periods of one week (Heneman & Wexley, 1983; using the overall D-squared measure of accuracy) and no clear effect of rating delay for a two-day delay (Nathan & Lord, 1983). A similarly mixed picture exists for the effects of rating delay on observed halo. Certainly more integretive research and reviews are needed which cumulate findings taking into careful consideration the operationalizations of rater error and rater accuracy.

Lack of a significant increase in Observed Halo may have been due to ceiling effects. The mean levels of Observed Halo were high throughout the rating delay groups (approximateley $\underline{r}$ = .80 and higher) and possibly represented the upper bound of halo likely to be found under the present experimental conditions.

The significant increase found in Absolute Halo Error over time pointed out how this measure behaved compared to uncorrected, or observed halo. Absolute Halo Error, rather than Observed Halo, increased over longer rating intervals even though the levels of true halo were essentially equal at each interval.

Systematic Distortion

The present study tested the SDH at both individual and group levels of analyses, and found what had been interpreted as SD in previous studies: SDI greater than COSTAC in (2-day) delayed rating conditions (see Table 6). When viewed in the context of a continuum of rating delay, however, SD did not increase as expected by the SDH in neither the group nor individual-level analysis. One possibility is that this study represented a "baseline" at the lower boundary of conditions likely to elicit SD. For example, the rating delay intervals in the present study may have been insufficient to provoke and systematically increase the SD response. Kozlowski and Kirsch (1987), for example, found SD effects when performance ratings were rendered up to six months after the target performances. Also, unfortunately, no measures were taken in the present study of subjects' familiarity with either ratee or the job of classroom lecturer. However, subjects in the present study, as students, were likely to be highly familiar with the job of classroom lecturer, which would have also reduced the likelihood of detecting SD. An additional consideration is that subjects in the present study had great difficulty in discriminating differences in ratee performance among the performance dimensions, as indicated by the extremely low discriminant validities and the low variance found for DA and SA compared to EL and DEL. Because SD, as operationalized here, is dependent upon variance patterns among the performance dimensions (in measuring the SDI), the use of videotapes characterized by positively skewed true halo may inhibit the ability to detect SD by

restricting across-dimension rating variance. Thus, the present study may be viewed as a demonstration of the lower bound conditions for SD. The most parsimonious interpretation of the levels and trends of SD in Figure 2 would be straight lines with slope = 0 through the plots for individual and group analyses. Tests of the SD hypothesis which seek to relate intensity of SD to intervals of delay should utilize greater retention intervals and measure raters' familiarity with the ratees and with the jobs being evaluated.

Of note also is the measure of implicit theory: the Performance Category Co-Occurrence Ratings. This instrument appeared to perform quite well in capturing what appeared to be a consistent "schema" of category covariance among the experimental groups as evidenced by the high inter-group correlations in the category co-occurrence ratings (Table 2).

Inter-Outcome Relationships

The data suggest that paradoxical as well as expected relationhips occur between rater error and rater accuracy measures. Moreover, at least some of the paradoxical relationships appear to be explainable as a function of the moderating influence of true levels of halo.

Although the correlations among the outcome measures did not appear to change as a result of the rating delay intervals, several relationships changed as a function of the level of true halo in the ratee performances evaluated. As argued by Becker and Cardy (1986), when true halo is high (as in the present data), then observed halo may enhance rating accuracy. This was the case in the present data

only for the relationship between OH and DA. When TH was high, this relationship was significant and positive; when TH was relatively low, the relationship was no longer significant. In the present study, the implications of given levels of Observed Halo, Restriction of Range, and Observed Leniency for Differential Accuracy, Differential Elevation, and Elevation were different, depending on the true halo in ratee performance. These conclusions are drawn cautiously, however, for several reasons. First, true halo in this study was positively skewed. The generalizeability of the results linking rater error measures with rating accuracy needs to be more closely examined in target performances which are normally distributed in true halo. Second, examination of the nature of the moderating influences of true halo needs further analysis. In this study, the use of subgroup bivariate correlations to explicate changes in prediction functions across levels of a presumed moderator variable presents problems with statistical power and in explication of the moderator effects. Problems in statistical power develop because of the small sample sizes within each subgroup. Also, more accurate methods are available (such as the Johnson-Neyman technique) for estimating regions along the range of the predictor variables for which differences in true halo exert the greatest effects.

Finally, the results of this study linking rater errors among themselves and with rater accuracy need to be further explicated and reconciled with prior studies which have examined those relationships. For example, Kozlowski and Kirsch (1987), among

others, found a different pattern of relations between rater error and rating accuracy and between halo and the systematic distortion index than found in the present study.

Summary of Major Findings

The results of the present study suggested that (a) systematic distortion does not influence performance ratings when rating delays are short (i.e., up to one week) and when raters are likely to be familiar with the ratees' jobs, (b) rating delays of up to one week do not influence most of the Cronbach (1955) accuracy measures, most rater errors, or convergent/discriminant validity, (c) rater errors may covary among themselves and with rating accuracy, but not always in the expected directions, (d) convergent and discriminant validity indices may not have direct implications for most rater accuracy outcomes, and (e) true halo may exert a strong role as a correlate of several rater error measures, as a correlate of interrater agreement, and as a possible moderator of rater error/rater accuracy relationships.

Final Comment

This study suggested that, under the experimental conditions created, true halo levels may influence rating outcomes to a greater degree than rating delay intervals. As Cooper (1981a, 1981b) observed, real world performance in many jobs is likely to contain true halo. Although the distribution of true halo and true levels of performance in jobs may be difficult to determine, theoretical models relating rater errors with other outcomes must confront the reality of positively skewed true halo in certain performance

evaluation environments. To be useful, performance evaluation methods, instrumentation, and evaluation quality criteria should be operationalized and interpretable when true performance variance is restricted and true performance levels are inflated and unidimensional. These conditions may occur when a workforce has been subjected to formal and demanding selection processes. Implications also exist for the design of appraisal formats and feedback systems under such circumstances. The use and interpretation of commonly-used rater errors such as halo, leniency, and range restriction under these, and perhaps most circumstances, is not a simple and direct matter.

References

Barnes-Farrell, J, & Couture, K. (1984). Effects of appraisal salience on immediate and memory-based judgments. (Report No. 84-1). Arlington, VA: Office of Naval Research. (DTIC AD No. A140334).

Becker, B., & Cardy, R. (1986) Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. Journal of Applied Psychology, 71, 662-671.

Berman, J., & Kenny, D. (1976). Correlational bias in observer ratings. Journal of Personality and Social Psychology, 34, 263-273.

Borman, W. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, and J. Cleveland (Eds.), Performance measurement and theory. Hillsdale: Erlbaum

Borman, W., Hough, L, & Dunnette, M. (1978). Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater errors. (Report No. TR-78-A12). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (DTIC AD No. A061149).

Bruner, J., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), Handbook of social psychology. Cambridge, MA: Addison-Wesley.

Chapman, L., & Chapman, J. (1967). Genesis of popular but erroneous psychodiagnostic observations. Journal of Abnormal

Psychology, 72, 193-204.

Chapman, L., & Chapman, J. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 74, 271-280.

Cline, V. (1964). Interpersonal perception. B. A. Maher (Ed.), Progress in experimental personality research, Vol. 1. (pp. 221-284). New York, NY: Academic Press.

Cooper, W. (1981a). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Cooper, W. (1981b). Conceptual similarity as a source of illusory halo in job performance ratings. Journal of Applied Applied Psychology, 66, 302-307.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". Psychological Bulletin, 52, 177-193.

D'Andrade, R. (1974). Memory and the assessment of behavior. In H. M. Blalock, Jr. (Ed.), Measurement in the social sciences. Chicago: Aldine.

Fisicaro, S. (1987). A re-examination of the relationship between halo error and accuracy. Manuscript submitted for publication.

Foti, R., Fraser, S., & Lord, R. (1982). Effects of leadership labels and prototypes on perceptions of political leaders. Journal of Applied Psychology, 67, 326-333.

Gage, N., & Cronbach, L. (1955). Conceptual and methodological problems in interpersonal perception. Psychological Review, 62, 411-422.

Hakel, M. (1974). Normative personality factors recovered from

ratings of personality descriptors: The beholder's eye. Personnel Psychology , 27, 409-421.

Ilgen, D., & Favero, J. (1983). Methodological contributions of person perception to performance appraisal. (Report No. 83-4), Arlington, VA: Office of Naval Research. (DTIC AD No. A1228638).

Johnson, D. (1963). Reanalysis of experimental halo effects. Journal of Applied Psychology, 47, 46-47.

Kavanagh, M., MacKinney, A, & Wolins, L. (1971). Issues in managerial performance: Multitrait - multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Kozlowski, S., & Kirsch, M. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. Journal of Applied Psychology, 72, 252-261.

Kozlowski, S., Kirsch, M., & Chao, G. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: An exploration. Journal of Applied Psychology, 71, 45-49.

Kruskal, J., & Wish, M. (1978). Multidimensional scaling. Beverly Hills: Sage.

Lamiell, J. (1980). On the utility of looking in the "wrong" direction. Journal of Personality, 48, 82-88.

Landy, F., & Farr, J. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Landy, F., & Farr, J. (1983). The measurement of work: Methods, theory, and applications. New York: Academic Press.

Lee, R., Malone, M., & Greco, S. (1981). Multitrait-multimethod-multirater analysis of performance ratings for law enforcement

personnel. Journal of Applied Psychology, 66, 625-632.

Lord, R., Foti, R., & DeVader, C. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. Organizational Behavior and Human Performance , 34, 343-378.

Murphy, K., & Balzer, W. (1981). Rater errors and rating accuracy. Paper presented at the American Psychological Association annual convention, Los Angeles, CA.

Murphy, K., & Balzer, W. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Murphy, K., Balzer, W., Kellam, K., & Armstrong, J. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76, 45-54.

Murphy, K., Balzer, W., Lockhart, M., & Eisenman, E. (1985). Effects of previous performance on evaluations of present performance. Journal of Applied Psychology, 70, 72-84.

Murphy, K., Martin, C., & Garcia, M. (1982). Do Behavioral Observation Scales measure observation? Journal of Applied Psychology, 67, 562-567.

Nathan, B., & Lord, R. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Newcomb, T. (1931). An experiment designed to test the validity

of a rating technique. Journal of Educational Psychology, 22, 279-289.

Nisbett, R., & Wilson, T. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of Personality and Social Psychology, 35, 250-256.

Phillips, J., & Lord, R. (1981). Causal attributions and perceptions of leadership. Organizational Behavior and Human Performance, 28, 143-163.

Phillips, J., & Lord, R. (1982). Schematic information processing and perceptions of leadership in problem-solving groups. Journal of Applied Psychology, 67, 486-492.

Pulakos, E., Schmitt, N., and Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. Journal of Applied Psychology, 71, 29-32.

Schneider, D. (1973). Implicit personality theory: A review. Psychological Bulletin, 79, 294-309.

Schultz, D., & Siegel, A. (1964). The analysis of job performance by multidimensional scaling techniques. Journal of Applied Psychology, 48, 329-335.

Shweder, R. (1975). How relevant is an individual difference theory of personality? Journal of Personality, 43, 455-484.

Shweder, R. (1977). Likeness and likelihood in everyday thought: Magical thinking in judgments about personnality. Current Anthropology, 18, 637-658.

Shweder, R. (1980). Factors and fictions in person perception: A reply to Lamiell, Foss, and Cavenee. Journal of Personality,

48, 74-81.

Shweder, R. (1982). Fact and artifact in trait perception: The systematic distortion hypothesis. In B. A. Maher and W. B. Maher (Eds.), Progress in experimental personality research, (vol 2). New York: Academic Press.

Shweder, R. (1983). In defense of surface structure. In F. Landy, S. Zedeck, and J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, N.J: Erlbaum.

Shweder, R., & D'Andrade, R. (1979). Accurate reflection or systematic distortion? A reply to Block, Weiss, and Thorne. Journal of Personality and Social Psychology, 37, 1075-1084.

Shweder, R., & D'Andrade, R. (1980). The systematic distortion process. In R. A. Shweder and D. W. Fiske (Eds.), New directions for methodology of social and behavioral science, (vol. 4). San Francisco: Jossey-Bass.

Sulsky, L., & Balzer, W. (1987). The meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Manuscript submitted for publication.

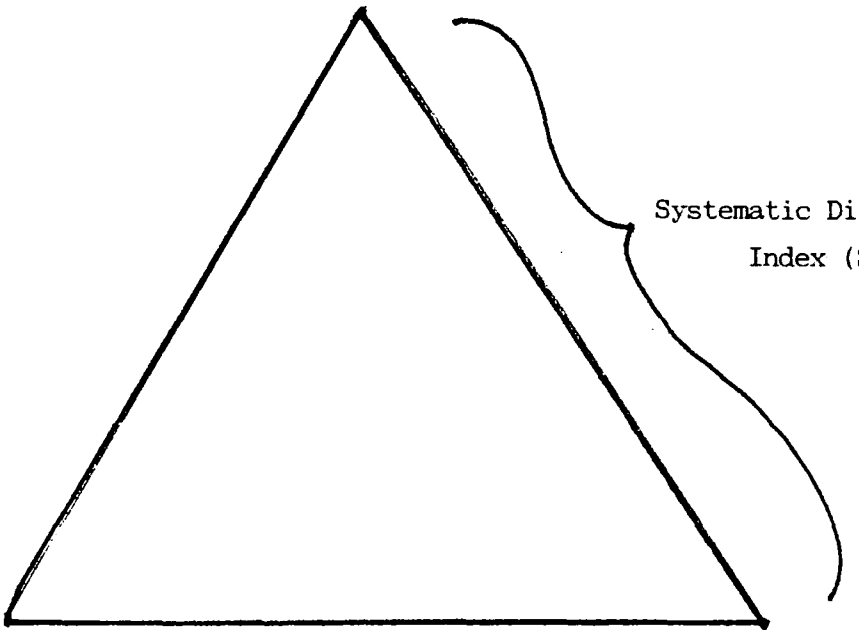
Thorndike, E. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.

Willingham, W., & Jones, M. (1958). On the identification of halo through analysis of variance. Educational and Psychological Measurement, 18, 403-407.

Figure Caption

Figure 1. Correlational matrices in the analysis of systematic distortion effects.

Category Similarity or Co-Occurrence Matrix



Correlational Structure Accuracy (COSTAC)
of Performance Ratings

Table 1

Means and Standard Deviations of Performance Category Co-Occurrence Ratings

		1	2	3	4	5	6	7	8
1. Thoroughness of Preparation		-							
2. Grasp of Material	M	5.78	-						
	SD	1.11							
3. Organization & Clarity	M	6.01	4.53	-					
	SD	.83	1.32						
4. Poise & Demeanor	M	5.45	5.07	4.59	-				
	SD	1.20	1.46	1.32					
5. Responsiveness to Questions	M	4.55	5.62	4.42	5.63	-			
	SD	1.37	.96	1.46	1.28				
6. Educational Value of Lecture	M	5.17	4.62	4.35	4.65	4.74	-		
	SD	1.46	1.54	1.38	1.21	1.38			
7. Rapport with Audience	M	4.86	4.26	5.11	4.49	3.11	5.14	-	
	SD	1.40	1.29	1.21	1.27	1.43	1.22		
8. Speaking Ability	M	5.66	4.46	6.07	4.53	4.10	4.03	5.45	-
	SD	1.01	1.42	.92	1.32	1.49	1.46	1.09	

Note. n = 293. Ratings were made on a scale of 0 (Not Likely to Co-Occur) to 7 (Very Likely to Co-Occur).

Table 2

Means, Standard Deviations, and Correlations for Mean Performance Category Co-Occurrence Ratings among the Delay Groups

	M	SD	1	2	3	4
1. Immediate-group	4.79	.68				
2. One-day delay group	4.99	.62	.982			
3. Two-day delay group	4.89	.71	.977	.975		
4. Seven-day delay group	4.81	.75	.963	.977	.976	

Note. n = 28. Ratings were made on a scale of 0 (Not Likely to Co-Occur) to 7 (Very Likely to Co-Occur).

Table 3

Descriptive Statistics for Rating Outcomes and True Halo for Combined and Delay Groups

		Combined	Immediate	1-day	2-day	One-week
	n	293	80	79	80	54
Elevation	M	.37	.33	.38	.36	.45
	SD	.30	.26	.27	.30	.38
Differential Elevation	M	.60	.58	.53	.66	.61
	SD	.30	.29	.24	.35	.30
Stereotype Accuracy	M	.33	.33	.34	.33	.33
	SD	.11	.12	.11	.12	.09
Differential Accuracy	M	.53	.52	.53	.54	.52
	SD	.12	.11	.13	.12	.13
Observed Leniency	M	.01	.02	.01	-.05	.10
	SD	.39	.38	.35	.44	.39
Leniency Error	M	-.28	-.22	-.30	-.24	-.38
	SD	.39	.36	.35	.40	.46
Restriction of Range	M	1.10	1.11	1.14	1.12	1.02
	SD	.29	.30	.29	.28	.30
Observed Halo	M	1.07	1.07	1.13	1.08	1.00
	SD	.57	.60	.55	.56	.30
Halo Error	M	-.76	-.82	-.64	-.79	-.81
	SD	.70	.65	.51	.79	.85
Absolute Halo Error	M	.88	.91	.70	.94	.99
	SD	.55	.51	.43	.59	.63
True Halo	M	1.83	1.89	1.76	1.86	1.81
	SD	.63	.67	.60	.66	.58
Systematic Distortion Index (SDI)	M	.13	.15	.12	.16	.08
	SD	.24	.25	.22	.23	.24

Note. Halo and SDI values are in Fisher z form. Higher accuracy score values indicate lower levels of accuracy. Higher Restriction of Range values indicate lower levels of range restriction.

Table 4

One-Way Multivariate Analysis of Variance: Effects of Rating Delay on Rating Outcomes

Wilks' Lambda for "Rating Delay": .75 $F(33,822.7) = 2.54^{***}$

Post-hoc Univariate Analyses:

Dependent Variables	Source	df	ms	F[a]
Elevation Accuracy	Delay	3	.180	2.04
	Error	289	.088	
Differential Elevation	Delay	3	.247	2.87*
	Error	289	.086	
Stereotype Accuracy	Delay	3	.001	.05
	Error	289	.012	
Differential Accuracy	Delay	3	.004	.25
	Error	289	.015	
Observed Halo	Delay	3	.160	.49
	Error	289	.329	
Halo Error	Delay	3	.55	1.13
	Error	289	.49	
Absolute Halo Error	Delay	3	1.24	4.31**
	Error	289	.289	
Observed Leniency	Delay	3	.220	1.46
	Error	289	.15	

(Table continues)

Table 4 (Continued)

One-Way Multivariate Analysis of Variance: Effects of Rating Delay on Rating Outcomes

Dependent Variables	Source	df	ms	F[a]
Leniency Error	Delay	3	.33	2.18
	Error	289	.15	
Restriction of Range	Delay	3	.17	1.96
	Error	289	.09	
Systematic Distortion Index	Delay	3	.081	1.47
	Error	289	.055	

[a]Exact F equivalent of Wilks' Lambda

* p $\leq$ .05 **p $\leq$ .005 *** p $\leq$ .0001

Table 5

Significant Post Hoc Scheffe Contrasts among Delay Groups for Differential Elevation Accuracy and Absolute Halo Error

Dependent Variable	Contrast	Mean Difference	df	F
Differential Elevation	D1 vs D2	-.13	3, 289	2.71*
	D1 vs $\overline{D2,D7}$	-.11	3, 289	2.31 ($p < .10$)
Absolute Halo Error	D1 vs D2	-.24	3, 289	2.72*
	D1 vs D7	-.29	3, 289	3.14**
	D1 vs $\overline{D2,D7}$	-.27	3, 289	3.99***
	$\overline{D0,D2,D7}$ vs D1	.25	3, 289	3.99***

Note. D0 = Immediate rating group; D1 = one-day delay group; D2 = two-day delay group; D7 = one-week delay group.

*p < .05

**p < .025

***p< .01

Table 6

Systematic Distortion among the Delay Groups: Individual and Group-level Analyses

Level of Analysis	n	Delay Group	SDI[a]	COSTAC[b]
Individual[c]	80	Immediate	.15	.13
Individual[c]	79	One-day delay	.12	.15
Individual[c]	80	Two-day delay	.16	.08
Individual[c]	54	Seven-day delay	.08	.10
Group	28	Immediate	.65	.41
Group	28	One-day delay	.29	.41
Group	28	Two-day delay	.59	.37
Group	28	Seven-day delay	.25	.42

Note. All correlation values are in Fisher z form. None of the correlational differences across groups (within levels of analysis) are significantly different.

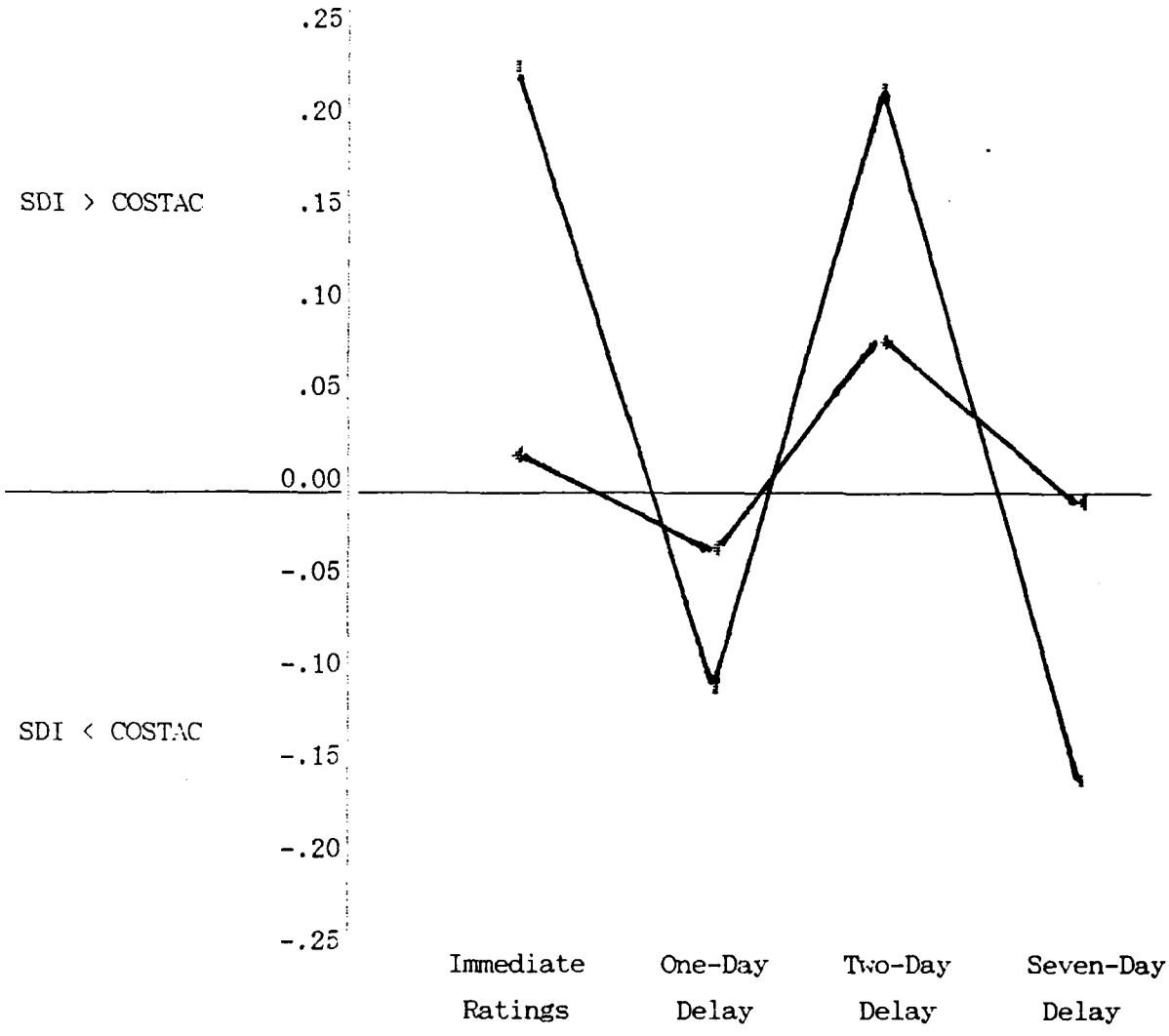
[a]SDI = systematic distortion index.

[b]COSTAC = correlational structure accuracy.

[c]SDI and COSTAC values for individual level analysis represent means.

Figure Caption

Figure 2. Systematic Distortion Trends for Individual and Group-Level Analyses.



= Group-level

= Individual-level

Table 7

Multi-Trait Multi-Method Analyses of Variance for 13 Tape Combinations

Tapes	True Halo[a]	Source	df	MS	F	VC[b]	IC[c]
1A 2B	.429	Ratee	3	41.17	93.49***	.204	.32
3B 4A		Ratee x Dimension	21	1.33	3.02***	.036	.08
		Rater x Ratee	72	3.56	8.31***	.403	.48
		Error	504	.44		.44	
1A 2B	.956	Ratee	3	18.14	42.21***	.082	.16
3B 4B		Ratee x Dimension	21	2.44	5.68***	.075	.15
		Rater x Ratee	78	6.02	14.00***	.70	.62
		Error	546	.43		.43	
1B 2B	1.467	Ratee	3	52.59	114.68***	.31	.40
3B 4A		Ratee x Dimension	21	1.15	2.51***	.03	.07
		Rater x Ratee	60	4.26	9.28***	.48	.51
		Error	420	.46		.46	
1A 2B	1.632	Ratee	3	34.97	81.53***	.39	.48
3A 4B		Ratee x Dimension	21	1.52	3.55***	.10	.19
		Rater x Ratee	30	6.22	14.51***	.73	.63
		Error	210	.43		.43	

(Table continues)

Table 7 (Continued)

Multi-Trait Multi-Method Analyses of Variance for 13 Tape Combinations

Tapes	True Halo	Source	df	MS	F	VC	IC
1A 2A	1.758	Ratee	3	49.06	137.91***	.32	.47
3B 4B		Ratee x Dimension	21	.79	2.21**	.02	.08
		Rater x Ratee	54	2.80	7.88***	.31	.59
		Error	378	.36		.36	
1B 2B	1.761	Ratee	3	2.81	7.04***	.05	.12
3B 4B		Ratee x Dimension	21	.42	1.04	.00	.06
		Rater x Ratee	15	2.66	6.65***	.28	.46
		Error	105	.40		.40	
1A 2A	1.798	Ratee	3	115.04	345.10***	.49	.60
3A 4B		Ratee x Dimension	21	1.14	3.41***	.03	.08
		Rater x Ratee	84	4.21	12.64***	.49	.59
		Error	588	.33		.33	
1B 2B	1.978	Ratee	3	250.50	612.26***	.80	.66
3A 4A		Ratee x Dimension	21	1.93	4.72***	.04	.09
		Rater x Ratee	114	3.75	9.15***	.42	.51
		Error	798	.41		.41	

(Table continues)

Table 7 (Continued)

Multi-Trait Multi-Method Analyses of Variance for 13 Tape Combinations

Tapes	True Halo	Source	df	MS	F	VC	IC
1B 2A	2.199	Ratee	3	236.06	528.38***	.95	.68
3A 4A		Ratee x Dimension	21	.92	2.05**	.02	.03
		Rater x Ratee	90	2.60	5.83***	.27	.38
		Error	630	.45		.45	
1B 2A	2.236	Ratee	3	81.47	168.35***	.60	.56
3B 4A		Ratee x Dimension	21	.96	1.98***	.03	.05
		Rater x Ratee	48	4.70	9.70***	.53	.52
		Error	336	.48		.48	
1B 2B	2.472	Ratee	3	16.42	54.61***	1.01	.77
3A 4B		Ratee x Dimension	21	.37	1.23	.03	.10
		Rater x Ratee	3	.40	1.32	.01	.04
		Error	21	.30		.30	
1B 2A	2.536	Ratee	3	175.79	433.00***	.56	.58
3B 4B		Ratee x Dimension	21	.66	1.63*	.01	.02
		Rater x Ratee	114	3.73	9.19***	.42	.51
		Error	798	.41		.41	

(Table continues)

Table 7 (Continued)

Multi-Trait Multi-Method Analyses of Variance for 13 Tape Combinations

Tapes	True Halo	Source	df	MS	F	VC	IC
1B 2A 3A 4B	2.582	Ratee	3	238.78	757.59***	1.15	.78
		Ratee x Dimension	21	1.12	3.55***	.03	.09
		Rater x Ratee	75	4.12	13.08***	.48	.60
		Error	525	.32		.32	

[a]in Fisher's z form

[b]VC = variance component

[c]IC = intraclass correlation

Table 8

Descriptive Statistics for Intraclass Correlations in each Delay Group

Variable	Delay Group	n[a]	M	SD
Ratee	Immediate	11	.55	.29
Main	One-day	10	.63	.32
Effect	Two-day	9	.57	.26
	Seven-day	9	.51	.35
Ratee x	Immediate	11	.11	.09
Dimension	One-day	10	.09	.07
Interaction	Two-day	9	.06	.08
	Seven-day	9	.04	.04
Rater x	Immediate	11	.51	.14
Ratee	One-day	10	.43	.20
Interaction	Two-day	9	.54	.42
	Seven-day	9	.40	.35

Note. All coefficients were converted to Fisher z form prior to analysis. Means have been converted back to Pearson coefficient values.

[a]n = the number of rater groups viewing a unique combination of videotapes in each rating delay condition.

Table 9

Correlations for Combined and Delay Groups

		EL[a]	DEL[a]	SA[a]	DA[a]	OH	AHE	TH	SDI
1. Elevation (El)									
2. Differential Elevation (DEL)	C	01							
	D0	-07							
	D1	04							
	D2	10							
	D7	-05							
3. Stereotype Accuracy (SA)	C	16**	-05						
	D0	26*	-16						
	D1	15	28**						
	D2	18	-13						
	D7	01	-14						
4. Differential Accuracy (DA)	C	10	06	32***					
	D0	01	14	31**					
	D1	22*	09	41***					
	D2	01	-02	21					
	D7	18	05	41**					
5. Observed Halo (OH)	C	25**	-23**	28**	45***				
	D0	17	-33**	46***	44***				
	D1	37***	-18	20*	37***				
	D2	17	-22*	26*	58***				
	D7	30*	-20	11	43***				
6. Absolute Halo Error (AHE)	C	-07	02	-19***	-35***	-48***			
	D0	-13	29**	-23*	-31**	-42***			
	D1	01	04	-20**	-40***	-41***			
	D2	02	-17	-23	-39***	-49***			
	D7	-17	13	-08	-35***	-60***			
7. True Halo (TH)	C	15**	00	05	03	33***	44***		
	D0	13	01	18	12	49***	48***		
	D1	28**	-02	03	03	60***	38***		
	D2	21	-16	00	06	18	39**		
	D7	-05	39**	-07	-16	-10	52***		
8. Systematic Distortion Index (SDI)	C	00	03	-03	-13*	-05	-04	-09	
	D0	-10	07	-09	-29**	-16	09	-04	
	D1	-02	09	08	05	02	-14	-03	
	D2	-07	09	-12	-21	-06	01	-14	
	D7	13	-16	06	-05	02	-19	-21	

(Table continues)

Table 9 (Continued)
Correlations for Combined and Delay Groups

		EL[a]	DEL[a]	SA[a]	DA[a]	OH	AHE	TH	SDI	ROR	OL	LE
9. Restriction of Range[b] (ROR)	C	-25***	22***	-08	11	-66***	27***	-30***	-04			
	D0	-28**	35**	-29**	-03	-67***	33**	-39***	-11			
	D1	-16	22*	06	13	-74***	19	-55***	04			
	D2	-35**	15	-16	06	-59***	19	-17	08			
	D7	-16	21	-19	23	-62***	35**	-01	-16			
10. Observed Leniency (OL)	C	63***	06	07	06	04	-02	-01	05	-07		
	D0	65***	-14	16	-02	02	-15	-05	06	-18		
	D1	61***	11	11	15	01	-05	-10	19	08		
	D2	67***	15	-04	-06	03	14	15	02	-20		
	D7	73***	09	09	21	16	-09	-10	-01	-00		
11. Leniency Error (LE)	C	80***	06	13	06	20***	-05	17**	03	-22***	75***	
	D0	77***	-15	31**	-09	15	-07	15	-11	-27**	79***	
	D1	84***	20	14	21	34**	01	30**	08	-19	74***	
	D2	72***	15	02	-07	09	-01	18	-01	-31**	78***	
	D7	88***	07	04	21	22	-14	-02	12	-03	83***	
12. Halo Error (HE)	C	07	-19***	18**	34***	52***	-78***	-63***	04	-27***	04	01
	D0	02	-31**	24*	29**	43***	-89***	-58***	-10	-23*	07	-02
	D1	07	-18	18	36***	38***	-88***	-51***	06	-16	12	02
	D2	-06	-03	18	36***	56***	-68***	-71***	08	-29**	-10	-08
	D7	21	-10**	13	40**	74***	-76***	-71***	15	-41**	18	16

Note. Decimals omitted. C = Combined sample, n = 293. D0 = Immediate rating group, n = 80. D1 = One-day delay group, n = 79. D2 = Two-day delay group, n = 80. D7 = Seven-day delay group, n = 51. Halo and SDI values are based on Fisher z transformations.

[a] signs have been reversed to reflect correlations with accuracy.

[b] signs have been reversed to reflect correlations with restricted range, i.e. decreased averaged standard deviations.

* p < .05

** p < .01

*** p < .001

Table 10

Group-Level Correlations among True Halo, Variance Indices and Mean Rating Outcomes

	TH	CV	DV	RH	OH	ROR	EL	DEL	SA
True Halo (TH)									
Convergent Validity (CV)	.75**								
Discriminant Validity (DV)	-.38	-.18							
Relative Halo (RH)	-.30	-.38	.31						
Observed Halo (OH)	.69**	.94***	-.08	-.13					
Restriction of Range (ROR)	-.53	-.79***	-.03	-.03	-.89***				
Elevation (EL)	.31	.73**	.03	-.10	.78	-.71**			
Differential Elevation (DEL)	.22	.29	-.20	-.93***	.004	.23	.03		
Stereotype Accuracy (SA)	.16	.11	.07	-.53	.30	.00	.14	.60*	
Differential Accuracy (DA)	.23	.20	.11	-.09	.14	.20	.03	.34	.56*

Note. n = 13.

* p < .05 **p < .01 ***p < .001

Table 11

Hierarchical Moderated Regression Analyses

Dependent Variable	Independent Variables Added to Equation	R^2 Increment	df	F
Elevation Accuracy	OH TH DELG	.085	5, 287	5.33***
	OH x DELG	.011	3, 284	1.18
	OH x TH	.006	1, 283	1.81
	OH x TH x DELG	.006	3, 280	.64
	ROR TH DELG	.082	5, 287	5.14**
	ROR x DELG	.008	3, 284	.84
	ROR x TH	.008	1, 283	2.56
	ROR x TH x DELG	.012	3, 280	1.25
	OL TH DELG	.464	5, 287	49.58***
	OL x DELG	.019	3, 284	3.56
	OL x TH	.039	1, 283	23.15***
	OL x TH x DELG	.003	3, 280	.57
Differential Elevation Accuracy	OH TH DELG	.092	5, 287	5.81***
	OH x DELG	.004	3, 284	.42
	OH x TH	.059	1, 283	19.80***
	OH x TH x DELG	.033	3, 280	3.80*
	ROR TH DELG	.090	5, 287	5.69***
	ROR x DELG	.004	3, 284	.45
	ROR x TH	.071	1, 283	24.02***
	ROR x TH x DELG	.036	3, 280	4.18*

(Table continues)

Table 10 (Continued)

Hierarchical Moderated Regression Analyses

Dependent Variable	Independent Variables Added to Equation	R^2 Increment	df	F
Differential Elevation Accuracy	OL TH DELG	.032	5, 287	1.90
	OL x DELG	.014	3, 284	1.36
	OL x TH	.009	1, 283	2.79
	OL x TH x DELG	.032	3, 280	3.27
Stereotype Accuracy	OH TH DELG	.078	5, 287	5.08**
	OH x DELG	.022	3, 284	2.22
	OH x TH	.002	1, 283	.54
	OH x TH x DELG	.002	3, 280	.21
	ROR TH DELG	.008	5, 287	.05
	ROR x DELG	.032	3, 284	3.10
	ROR x TH	.001	1, 283	.22
	ROR x TH x DELG	.005	3, 280	.48
	OL TH DELG	.009	5, 287	.51
	OL x DELG	.008	3, 284	.77
	OL x TH	.005	1, 283	1.55
	OL x TH x DELG	.007	3, 280	.64
Differential Accuracy	OH TH DELG	.23	5, 287	16.67***
	OH x DELG	.001	3, 284	.69
	OH x TH	.027	1, 283	10.18*
	OH x TH x DELG	.004	3, 280	.55

(Table continues)

Table 10 (Continued)

Hierarchical Moderated Regression Analyses

Dependent Variable	Independent Variables Added to Equation	R^2 Increment	df	F
Differential Accuracy	ROR TH DELG	.018	5, 287	1.05
	ROR x DELG	.005	3, 284	.53
	ROR x TH	.001	1, 283	.16
	ROR x TH x DELG	.020	3, 280	1.91
	OL TH DELG	.007	5, 287	.37
	OL x DELG	.014	3, 284	1.43
	OL x TH	.001	1, 283	.20
	OL x TH x DELG	.001	3, 280	.12

Note. OH = Observed Halo; TH = True Halo; DELG = Delay Group; ROR = Restriction of Range; OL = Observed Leniency.

* $p \leq .0125$

** $p \leq .001$

*** $p \leq .0001$

Table 12

Correlations between Common Rater Errors and Rating Accuracy for Levels of True Halo

True Halo	n[a]	Correlation between: OL, EL	ROR, DEL	OH, DEL	OH, DA
.429	25	.96****	.84***	-.81****	.37
.956	27	.95****	.79***	-.73****	.33
1.467	21	.98****	.60**	-.62**	.52*
1.632	11	.52	.32	-.50	-.10
1.751	1	-	-	-	-
1.758	19	.17	-.22	-.01	.43
1.761	6	.99****	.05	.05	.56
1.798	29	.70****	-.17	.07	.29
1.978	39	.89****	.31	-.07	.64****
2.199	31	.40*	-.10	-.03	.67****
2.236	17	.90****	-.20	.08	.79***
2.472	2	-	-	-	-
2.536	39	.98****	.07	.05	.52***
2.582	26	.17	.01	-.14	.39*

Note. True Halo and Observed Halo (OH) values are based on Fisher z equivalents. OL = Observed Leniency, EL = Elevation Accuracy, ROR = Restriction of Range, DEL = Differential Elevation Accuracy, DA = Differential Accuracy.

[a]indicates the number of raters who viewed tapes with the corresponding level of true halo.

* p < .05

** p < .01

*** p < .001

**** p < .0001