BASIC AND APPLIED SOCIAL PSYCHOLOGY, 1995, *16*(4), 417–434

The Systematic Distortion Hypothesis: A Confirmatory Test of the Implicit Covariance and General Impression Models

David J. Woehr
Texas A & M University

David V. Day
Pennsylvania State University

Winfred Arthur, Jr.
Texas A & M University

Arthur G. Bedeian
Louisiana State University

Supervisory ratings of subordinate performance were used to compare two variations of the systematic distortion hypothesis: an *implicit covariance* model and a *general impression* (GI) model. Our purpose was to assess whether a model specifying a set of correlated performance dimensions or a model specifying a single, GI factor provided the best representation of actual supervisory performance ratings ($N = 491$). Confirmatory factor analysis was used to test a first-order correlated factor (i.e., implicit covariance) model against a second-order hierarchical (i.e., GI) model. The results generally supported the latter perspective whereby a GI is thought to serve as the basis for subsequent dimensional judgments, which in turn serve as the basis for individual item ratings. The implications of these results with respect to rater training and rating accuracy are discussed.

The validity of performance-appraisal rating systems is typically construed to be a function of at least three factors (Ilgen & Feldman, 1983): (a) the

Requests for reprints should be sent to David J. Woehr, Department of Psychology, Texas A & M University, College Station, TX 77843–4235.

nature of the appraisal setting, (b) the behavior of those being appraised, and (c) the cognitive processes of appraisers. Whereas the first two factors boast rich and lengthy research traditions (see Landy & Farr, 1980), interest in how appraisers cognitively represent and process performance information has developed more recently (Feldman, 1981).

With regard to understanding how raters represent and process performance information, several models have been advanced including a *systematic distortion hypothesis* whereby ratings are systematically distorted in the direction of preexisting conceptual schemas (Shweder & D'Andrade, 1980). Research suggests that such schemas are more likely to be used in appraisal contexts when raters lack knowledge about a particular job or ratee, or when there is a delay between the observation of performance and subsequent ratings (for summaries of relevant literature, see Cooper, 1981b; Kozlowski, Kirsch, & Chao, 1986). This distortion process has also been proposed as an explanation for illusory halo (nonperformance-based general impression) in performance ratings (Cooper, 1981a, 1981b). Within the systematic distortion hypothesis, two alternative explanations for how ratings are systematically distorted relative to preexisting conceptual schemas have been advanced: (a) an implicit covariance model and (b) a general impression model.

IMPLICIT COVARIANCE MODEL

One explanation for how ratings are systematically distorted is that raters organize and retrieve performance information around conceptual schemas consisting of multiple correlated performance dimensions. According to this explanation, raters adopt an implicit model of performance whereby the multiple dimensions assumed to constitute performance are conceptualized as separate but interrelated. This model implies that raters form and store separate dimensional representations of an individual's performance. Subsequent performance ratings are then based on these dimensional evaluations as well as on the rater's beliefs about the conceptual similarities among the dimensions being rated. Thus, the intercorrelation of performance ratings is a function of both the ratee's actual performance and the rater's implicit theory of performance. Moderately to highly intercorrelated ratings are usually construed as evidence of substantial halo, although it is unclear to what extent this constitutes error (Cooper, 1981b; Lance & Woehr, 1986; Nathan & Lord, 1983). The implication, however, is that the greater the degree to which rating scales are congruent with raters' existing cognitive performance dimensions, the more likely performance appraisals are to be accurate (Borman, 1987; Woehr, 1992).

The implicit covariance model implies that halo error is largely a function of the rater and thus should be relatively consistent for each rater across ratees (Murphy & Anhalt, 1992). However, it is likely that a rater's reliance on an implicit theory of performance will increase as the cognitive demand associated with the rating task increases. For example, early research investigating the implicit covariance model (see Shweder & D'Andrade, 1980, for a summary of this literature) has shown that ratings made following a delay between observation and rating (i.e., with higher memory demand) tend to display greater intercorrelations than do ratings that are made immediately following performance observation. Furthermore, Cooper (1981a) provided evidence of illusory covariance in job performance ratings. His results were attenuated, however, probably because of a laboratory methodology and a procedure requiring ratings immediately after viewing individual videotapes. Cooper suggested that stronger support for the implicit covariance model might be found in a study that "would more closely approximate conventional job performance rating conditions" (1981a, p. 306), especially when the opportunity for memory decay was present (i.e., more time between observation and rating).

GENERAL IMPRESSION MODEL

Another possible explanation for the systematic distortion of performance ratings is that raters might maintain a single, general impression (GI) of a ratee's performance, which serves as the basis of subsequent performance ratings. This explanation is consistent with findings stemming from the social cognition literature (e.g., Lingle, Geva, Ostrom, Leippe, & Baumgardner, 1979; Lingle & Ostrom, 1979; Skowronski & Carlston, 1987) and the subsequent model of person perception proposed by Wyer and Srull (e.g., Srull & Wyer, 1989; Wyer, 1989; Wyer & Srull, 1989). These researchers suggest that individuals often spontaneously encode observed behaviors into idiosyncratic dimensional clusters and then extract a general evaluative concept of the target individual. Thus, a model in which dimensional representations are subsidiary to a more general evaluative representation is suggested. This model implies that an overall evaluative impression of a ratee (as well as dimension-related clusters of behavioral information) is formed and stored in memory. Subsequently, either the GI or the specific behavioral information may serve as the basis for performance appraisal ratings. However, because the general evaluative concept is the last representation formed and used, and because it is more cognitively efficient to store and use a single evaluative representation, this representation is more likely to be available when ratings are required (Lingle & Ostrom, 1979; Skowronski & Carlston, 1987; Srull & Syer, 1989; Wyer & Srull, 1989). Consequently, unless the actual rating items match and thus "cue" the more specific dimensional information, raters will retrieve and use this evaluative impression as the basis for performance ratings.

The Wyer and Srull (1989) model is consistent with a similar model proposed by Feldman and Lynch (1988), who suggested that inputs used to generate a particular evaluative response will depend on the accessibility and diagnosticity of information in memory. *Diagnosticity* refers to the degree to which some previous judgment or stored information is perceived to be relevant to a subsequent judgment. In performance evaluation, for example, an overall evaluative judgment may be moderately diagnostic for subsequent dimensional (but still evaluative) ratings. *Accessibility* refers to the ease with which a given episode, affective response, prior judgment, knowledge structure, or other cognitive/affective construct can be brought to awareness. It is further postulated that diagnosticity and accessibility interact, such that the first sufficiently diagnostic information retrieved is used as the basis for a particular judgment. Such a "race-for-information model" of the evaluation process is often supported in the information processing literature (e.g., Logan & Klapp, 1991). This model implies that an accessible, moderately diagnostic judgment (e.g., an overall evaluation) may be used as the basis for evaluation even if more diagnostic but less accessible information (memory for specific behavioral information) exists in memory.

Both the Wyer and Srull (1989) and Feldman and Lynch (1988) models imply that, under typical organizational conditions, an overall GI of a ratee (rather than impressions on multiple dimensions) will be retrieved from memory and form the basis for subsequent judgments on individual performance dimensions. Thus, the intercorrelation of dimensional performance ratings is attributable to the global evaluative impression. In addition, this GI affects what will be recalled about a ratee (Woehr & Feldman, 1993) and also helps direct what future behaviors a rater will attend to, encode, and store (Lord, 1985b).

LIMITATIONS OF PREVIOUS RESEARCH

Research by Landy, Vance, Barnes-Farrell, and Steele (1980) and a reanalysis of their data by Hulin (1982) have suggested that a single (global) GI factor best explained supervisory ratings of middle-level managers on 15 performance dimensions. However, both Landy et al. and Hulin used an exploratory factor-analytic technique and, thus, were unable to test the appropriateness of the GI model or compare it to the implicit covariance model. To date, no study has directly evaluated the appropriateness of either model or adopted a confirmatory technique to compare the two models directly.

A more general limitation of previous research in the area of performance appraisal, but not specifically of the Landy et al. (1980) and Hulin (1982) studies, is the tendency to use "paper people" (i.e., written scenarios or videotapes) in laboratory settings. Although a laboratory approach enables the experimenter to manipulate ratee behavior and assure equal exposure across raters, much of the informational richness (Daft & Lengel, 1984) in which ratee performance is embedded is lost, with a potential concomitant reduction in the generalizability of results (Ilgen & Favero, 1985). In line with Cooper (1981b), others have argued that because organizations provide a "noisy" environment that imposes numerous demands on raters' limited cognitive resources, performance information may be more likely encoded and stored by means of automatic processing (e.g., Feldman, 1981; Lord, 1985b; Woehr & Lance, 1991). If this is indeed the case, the explanatory model that provides for the most efficient use of cognitive resources will appear to be most tenable. Given the principle of *cognitive economy,* which states that categorization preserves as much information as possible while minimizing cognitive load (Rosch, 1978), it appears logical that raters will tend to adopt the GI model. This is likely to be the case because the implicit covariance model posits that raters develop and maintain numerous dimensions of performance in memory to encode and organize information, thus imposing a greater cognitive load on raters than does the GI model.

The purpose of the present study was to assess—using confirmatory factor analysis (CFA)—the adequacy of the implicit covariance and GI models in describing actual supervisory performance ratings. It was hypothesized that the GI model suggested by the Wyer and Srull (1989) and Feldman and Lynch (1988) perspectives, whereby correlations among a set of first-order factors are due to a single underlying factor (i.e., the GI) would provide a better fit to job performance ratings than a first-order correlated factor model suggested by the implicit covariance perspective. In addition, we attempted to overcome the most common methodological problems and limitations associated with prior research. That is, as suggested by Cooper (1981b) and Ilgen and Favero (1985), among others, we collected actual rather than simulated performance ratings from a large number of superior–subordinate dyads in a field setting. Hence, performance was observed and evaluated within the context of frequent face-to-face interactions between raters and ratees. Moreover, as a consequence of being conducted in an everyday employment situation, ratee performance was assessed as it naturally varied (i.e., it was not held constant or artificially manipulated). Because of the memory-based nature of the ratings and the opportunity for information decay, it was expected that performance ratings would be systematically distorted (Shweder, 1980). It is important to note that although moderately to highly intercorrelated and/or skewed ratings are usually construed as evidence of substantial "distortion," it is impossible to determine to what extent the covariation among ratings actually represents "distortion" or "error" without knowledge of the actual distribution of the target performance (Balzer & Sulsky, 1992). Such knowledge is, of course, not available in this study. However, our primary concern is not to determine the extent to which the supervisory ratings are distorted, but rather to examine the processing model that best represents supervisory performance ratings in a field setting.

METHOD

Sample

Data for this study were drawn from a larger sample originally collected to examine background and personality characteristics of accountants. The larger sample consisted of 1,821 accountants who were randomly drawn from the membership lists of the American Society of Certified Accountants, the National Association of Accountants, the American Association of Women Accountants, and the Association of Government Accountants and who had previously agreed to participate. A total of 1,145 research packets were returned from this larger sample, of which supervisory performance ratings were also returned for 491 (43%), representing a final participation rate of approximately 27%. Of these accountants, there were approximately an equal number of men and women. Ages were distributed as follows: approximately 24% between 20 and 29, 41% between 30 and 39, 19% between 40 and 49, 13% between 50 and 59, and approximately 3% age 60 or older. The average length of tenure in their present job was reported to be about 3.1 years (SD = 5.83 years). For the supervisors providing the ratings, approximately 90% were men. Approximately 4% were between the ages of 20 and 29, with 39% between 30 and 39, 30% between 40 and 49, 20% between 50 and 59, and approximately 7% age 60 and older. The average length of time reported as supervising a ratee was 5.43 years (SD = 5.14 years).

Performance Ratings

The accountants were independently rated by their immediate supervisors on 23 items, which were derived from the job performance literature and applicable to the accounting profession. The sampling adequacy of the selected items vis-à-vis the intended content domain was of special concern. To ensure a sufficient, but parsimonious, domain sampling, item generation was conducted following the guidelines outlined by Meister (1986) for human performance measurement. Accordingly, items were selected not only to be directly relevant across organizations and type of accounting performed (e.g., public, industrial, and government), but also with regard to their being easily collected, without specialized instrumentation, at the appropriate supervisory level, while still being meaningful for research purposes. Each item was rated on a scale ranging from *unsatisfactory* (1) to *excellent* (7). All ratings were returned directly to the researchers to guarantee respondent confidentiality. All items rated by the supervisors were included in the two models represented in the figures. A complete list of performance appraisal items is incorporated in Table 1.

TABLE 1
Standardized Estimates of Relations of Manifest Performance Rating Items on Four First-Order Factors

	Factors			
Performance Rating Items	*Ability*	*Commitment*	*Interpersonal Relations*	*Attendance/ Punctuality*
Ability	.665[a]			
Creativity	.725			
Job knowledge	.667			
Quality of work	.687			
Promotability	.840			
Judgment	.651			
Responsibility	.672			
Productivity	.713			
Accuracy	.577			
Gets job done	.669			
Professional image	.718			
Initiative	.723			
Effort	.475	.183		
Commitment to organization		.910		
Loyalty to organization		.849		
Commitment to job		.816[a]		
Loyalty to supervisor		.781		
Attitude		.349	.472	
Honesty/Integrity	.246	.192		
Interpersonal relationships			.701	
Cooperation			.778[a]	
Punctuality				.735
Attendance				.576[a]

[a]Fixed parameter.

Analysis

Consistent with the implicit covariance model, preliminary analyses revealed four factors (Ability, Commitment, Interpersonal Relationships, and Attendance/Punctuality) accounting for approximately 70% of the variance among the set of 23 items. CFA was then used to test a first-order model with four correlated factors (see Figure 1). Similarly, consistent with the GI perspective, we tested a hierarchical model with one second-order factor (i.e., General Impression, GI) underlying the four first-order factors (see Figure 2). That is, the correlations among the four first-order factors were specified to be accounted for by one underlying GI factor.

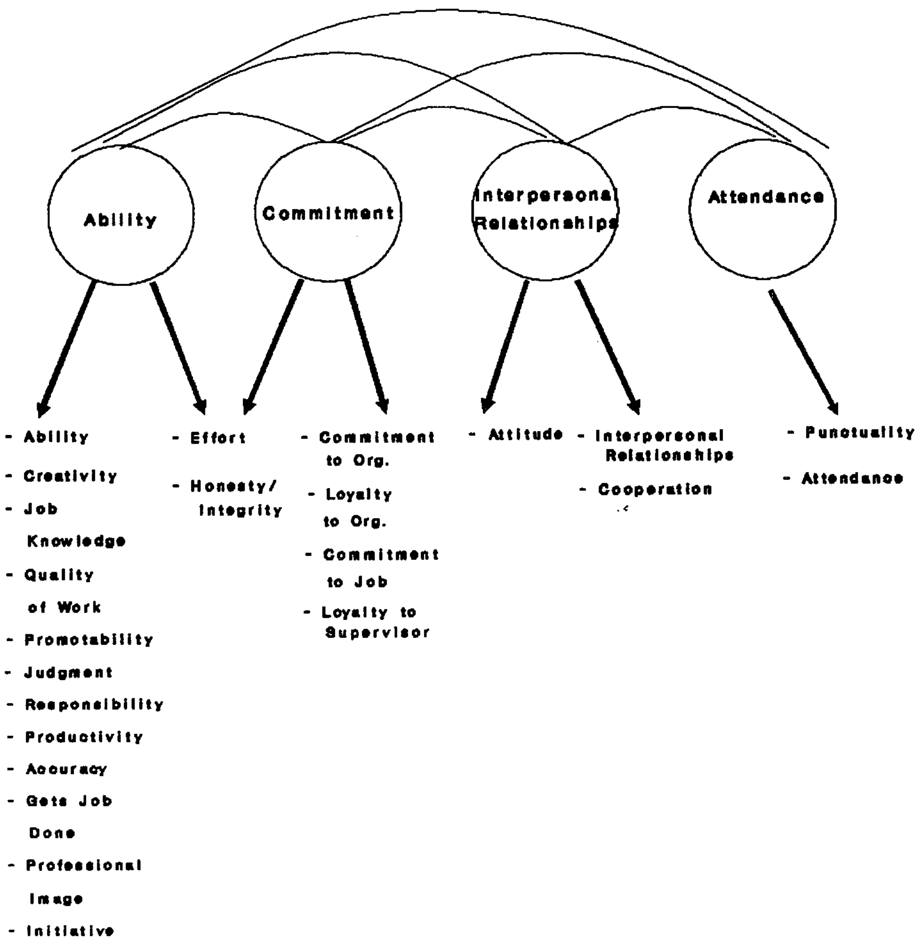

FIGURE 1 A first-order (implicit covariance) factor model.

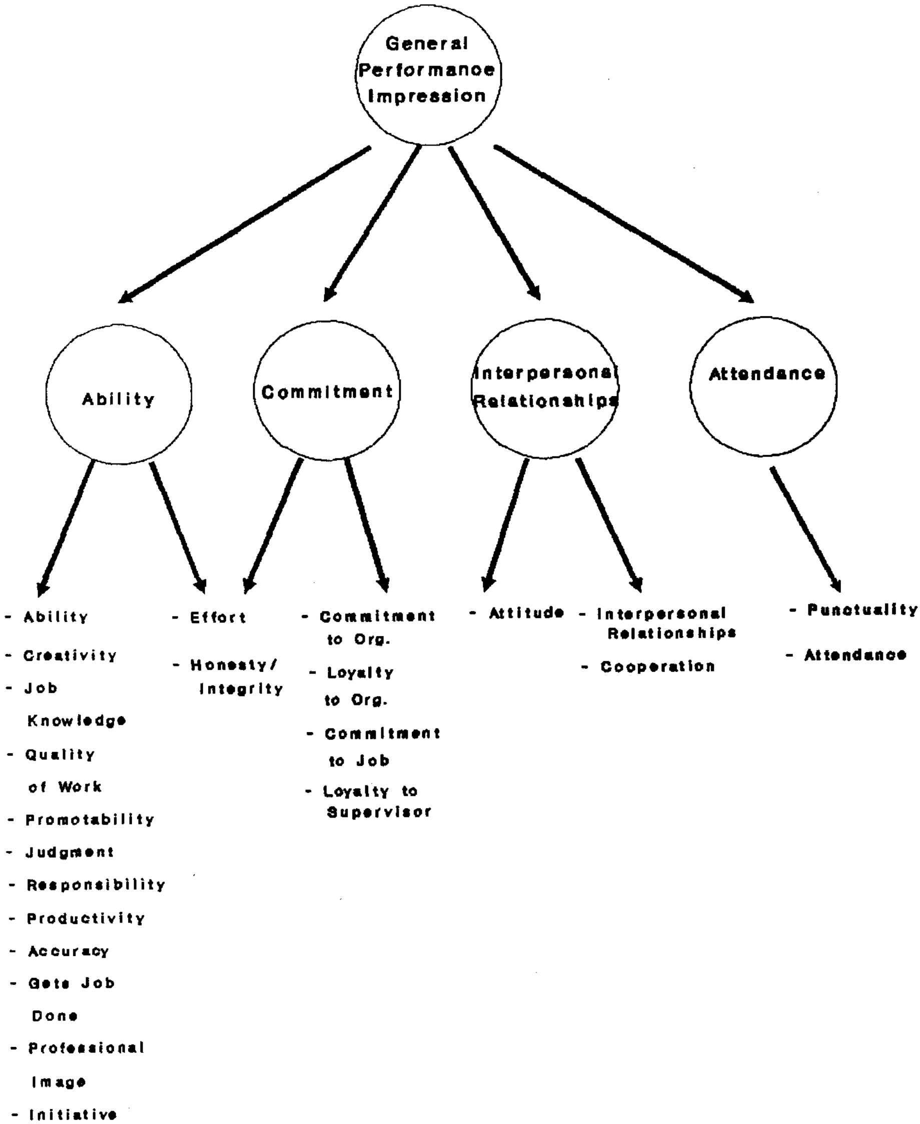

FIGURE 2 A hierarchical (GI) model.

RESULTS

The supervisory ratings of subordinate performance demonstrated moderate negative skew (-0.72) as well as moderate item intercorrelation (mean r among the 23 items = .49) and restricted range and variation (e.g., M = 5.95; SD = 0.68). This frequency distribution is consistent with the presence of systematic rating "distortion." That is, whereas distortion implies "error" only in the sense that the performance rating distribution is more (or less) interrelated or skewed than actual performance, distortion is often inferred in cases of skewed and/or moderately to highly intercorrelated rating distributions (Cooper, 1981b; Balzer & Sulsky, 1992). In this sense, the rating distribution is consistent with the presence of distortion (i.e., a necessary but not sufficient condition). In addition, although criticisms of this inference with respect to the extent to which such distributions actually reflect error are valid (e.g., Balzer & Sulsky, 1992; Murphy & Anhalt, 1992), our focus is not on this inference. Rather it is on examining which model best describes the observed rating distribution.

The goodness of fit for the two models (implicit covariance and GI) thought to account for the systematic distortion was evaluated using a CFA application of LISREL–VI (Jöreskog & Sörbom, 1986). Examination of descriptive statistics for the performance ratings indicated a potential deviation from multivariate normality. The models were therefore evaluated using a generalized least squares (GLS) solution as opposed to the more typical maximum likelihood solution. A benefit of using a GLS solution is that it yields an approximate chi-square distribution, which can be used to test a model's overall goodness of fit under somewhat less restrictive assumptions than multivariate normality (Browne, 1977). The observed *covariance* matrix among the 23 performance items was used to determine individual chi-square values for the first-order (i.e., covariance) factor model (χ_1^2) and the second-order (i.e., GI) factor model (χ_2^2). χ_1^2 was used as a stand-alone index of fit for the covariance factor model. Other fit indices examined for this model were the Goodness of Fit Index (GFI) calculated by the LISREL program and the χ_1^2/df ratio suggested by Jöreskog (1969).

Because the GI model represents a higher order model that directly corresponds to the first-order covariance model (i.e., the GI model posits a single, higher order factor to account for the covariation among the four first-order factors), the goodness of fit for the GI model cannot be better than that of the covariance model (Marsh & Hocevar, 1985). This is because the GI model attempts to explain all the covariation among the first-order factors with fewer parameters. Thus, the overall fit of the GI model was assessed in two ways. First, the fit of the GI model was assessed with a difference-of-chi-square test ($\chi_2^2 - \chi_1^2$, with degrees of freedom equal to $df_2 - df_1$). Because the GI model represents a more restricted model than the covariance model, a significant difference of chi-squares would indicate support for the latter whereas a nonsignificant difference of chi-squares would support the former (James, Mulaik, & Brett, 1982).

The fit of the GI model was also assessed with the target coefficient (T) index described by Marsh and Hocevar (1985). This index, designed specifically to assess the fit of a second-order model relative to a first-order model, is the ratio of the chi-square value from the first-order model to that of the second-order model. *T* has an upper limit of 1.0, which would occur if the covariation among the first-order factors were completely accounted for by the second-order model. An advantage of *T* is that it separates lack of fit due to the second-order structure from lack of fit in the definition of the first-order factors. Thus, it is possible to have a high *T* value even when the overall fit of the first-order model is only modest.

The GLS standardized solution provided by the covariance-model CFA is presented in Table 1. The data generally supported the proposed model (i.e., the parameters specified to be nonzero are in fact nonzero and those specified to be zero are zero). The correlations among the four first-order factors are presented in Table 2. As expected, however, there was a high degree of correlation among the four factors. The GFI (presented in Table 3) indicates that the covariance model's fit to the data is acceptable. The chi square for the model, $\chi_1^2(220, N = 491) = 657.9$, appears satisfactory given the study's sample size and the number of study variables (Hayduck, 1987).

Tests of overall fit (see Table 3) indicate that the GI model reasonably fits the data. The difference of chi-squares between the covariance and GI models is small and nonsignificant (see Table 3), indicating no significant reduction in fit to the data from the first-order, correlated factor model to the more restricted second-order, hierarchical model. In addition, the *T* value is .989 indicating that approximately 99% of the covariation among the four first-order factors can be accounted for by a second-order GI factor. Together these results indicate that the GI model provides the best description of the relations among the rating items. Additional support for the GI model is provided by the generally high magnitude of the estimated factor loadings for the first-order factors on a second-order, general factor (presented in Table 4). These results suggest that the GI factor explains substantial portions of the variance in each of the four first-order factors. Additionally, the indirect effects of the GI factor on the 23 manifest performance appraisal items are presented in Table 5. The indirect effects are a product of the direct effects of the GI factor on the first-order factors and the direct effects of the first-order factors on the manifest variables.

TABLE 2
Correlations Among Four First-Order Factors

Factors	*1*	*2*	*3*	*4*
1. Ability	—			
2. Commitment	.758	—		
3. Interpersonal Relations	.738	.793	—	
4. Attendance/Punctuality	.671	.643	.572	—

TABLE 3
Goodness-of-Fit Indices for First-Order and Hierarchical Models of Performance Impressions

	First-Order Model	*Hierarchical Model*
χ^2	657.9	664.8
df	220	222
N	491	491
GFI	0.883	0.882
RMSR	0.092	0.092
χ^2/df	2.990	2.995
Difference of chi-squares[a] = 6.90, $p > .01$		
$T = .989$		

Note. RMSR = Root Mean Square Residual.
[a]Difference of chi-squares is between first-order model and higher order model with 2 *df*.

TABLE 4
Estimated Factor Loadings for First-Order Factors on a Single Higher Order, GI Factor

Factor	*GI*
Ability	.867
Commitment	.893
Interpersonal Relations	.869
Attendance/Punctuality	.739

These results indicate that a reasonable proportion of variance in the performance ratings is accounted for by the GI factor.

These results provide empirical support for the hierarchical model of performance ratings. Thus, within the context of CFA, one may say that this model was confirmed. Such confirmation suggests that the model provides a useful basis for explaining the supervisory performance ratings. It does not, however, imply that the model is unique or has been proven to be true (James et al., 1982). More specifically, alternate models may explain the data equally well. An alternate explanation of particular concern here is that the set of rating items were actually unidimensional and only tapped one factor. Thus we also assessed the fit of a single-first-order-factor model. If such a model provides a reasonable fit to the data, it is possible that the results with respect to the GI model may simply be an artifact attributable to a unidimensional rating scale. CFA of a single, first-order factor, however, indicated a significantly lower level of fit, $\chi^2(230, N = 491) = 872.91$, GFI = .850, Root Mean Square Residual (RMSR) = .113, than either the four-factor or second-order-factor models.

TABLE 5
Indirect Effects of the GI Factor on Performance-Rating Items

	Factors			
Items	*Ability*	*Commitment*	*Interpersonal Relations*	*Attendance/ Punctuality*
Ability	.577			
Creativity	.629			
Job knowledge	.578			
Quality of work	.596			
Promotability	.728			
Judgment	.564			
Responsibility	.583			
Productivity	.618			
Accuracy	.500			
Gets job done	.580			
Professional image	.623			
Initiative	.627			
Effort	.412	.163		
Commitment to organization		.813		
Loyalty to organization		.758		
Commitment to job		.729		
Loyalty to supervisor	.697			
Attitude		.312	.410	
Honesty/integrity	.213	.172		
Interpersonal relationships			.609	
Cooperation			.676	
Punctuality				.543
Attendance				.426
Summary Statistics				
Range	.213 to .728	.163 to .813	.410 to .676	.426 to .545
M	.559	.521	.565	.485

In addition, to provide some construct validity for the GI factor, the squared correlations between the GI factor and four additional variables were examined. These variables were (a) the supervisor's overall rating of the ratee, (b) the supervisor's ranking of the ratee in the current work group, (c) the number of years the supervisor had been in a supervisory role, and (d) the supervisor's years of education. It was expected that the squared correlations should be higher for the first two variables and relatively small for the second two variables. This pattern was in fact supported. The squared correlations were .38, .13, .06, and .07, respectively.

DISCUSSION

Results of this field study comparing two alternative models underlying the systematic distortion of performance ratings generally supported the GI perspective. More specifically, support was found for a model in which a single (global) GI serves as the basis for subsequent dimensional judgments and individual item ratings over a model in which ratings derive from separate but correlated performance factors. This is consistent with the view that performance judgments are made through the formation of an overall impression—most likely using category-based and relatively automatic processes—based on what a rater believes is the relevant information required to make an accurate assessment. As such, a rater's GI of a ratee may not be solely a source of idiosyncratic bias, but may contain meaningful performance information (Lance & Woehr, 1986; Nathan & Tippins, 1990).

An area of possible concern with the present study is that it aggregated data over individuals, thereby inflating the systematic distortion effect (Kozlowski & Kirsch, 1987). Such distortion, however, is likely to occur whenever there is a time lag between observation and rating, thereby allowing for information decay (Shweder, 1980). In addition, it could be argued that our sample of raters may not have been motivated to provide accurate, unbiased ratings (Banks & Murphy, 1985); however, the ratings were collected expressly for research purposes which should reduce sources of rater bias affecting performance appraisal results when used for personnel decisions (Zedeck & Cascio, 1982).

Our study has several implications for rater training and for rater information-processing issues. First, it is likely that raters will form and use a GI of ratee performance. More important, research indicates that such a GI contains a substantial amount of valid performance information and leads to better performance ratings (Lance, Woehr, & Fisicaro, 1991; Nathan & Tippins, 1990). Thus, it is important to consider the nature of raters' performance-related categories or prototypes. It is likely that such prototypes serve as the basis for the construction of ratee impressions and that this process occurs relatively automatically.

It might be argued that training raters to engage in effortful, controlled processing might result in more differentiated ratings; however, research suggests that controlled processing is not always linked to more differentiated ratings (McKelvey & Lord, 1986; Woehr, 1992; Woehr & Feldman, 1989). On the other hand, it may be possible to train raters to become familiar with relevant performance dimensions as well as with the relations among dimensions. For example, previous research with expert judges found that experts are able to use multiple dimensions because they know the intercorrelations between dimensions (Phelps & Schanteau, 1978;

Schanteau, 1988). Rater training procedures, such as Frame-of-Reference (FOR) training (Bernardin & Buckley, 1981), may prove beneficial in this respect. FOR training provides raters with job-relevant schemas of performance by training them to recognize which behaviors are important for each performance dimension and to provide information about the effectiveness level of different behaviors. With practice, these dimensions may come to be used automatically to encode, store, and retrieve performance information, resulting in improved rating accuracy. Further, assuming that performance dimensions are interrelated to some extent and evidence "true" halo (Cooper, 1981b; Murphy, 1982), training raters to use multiple dimensions in organizing and retrieving performance information may result in more veridical ratings. FOR training has, in fact, been shown to result in more consistent and accurate ratee evaluations when compared with other training strategies or no rater training (Athey & McIntyre, 1987; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986). However, if an organization is concerned only with accuracy at the classification or categorical level (Lord, 1985a), relying on a GI as a basis for ratings may provide satisfactory accuracy with optimal efficiency.

In conclusion, we concur with Nathan and Lord (1983) that improving performance ratings first requires an understanding of how raters represent and process performance information. Our study was intended to further this understanding by comparing the appropriateness of the implicit covariance and GI explanations of how performance ratings are systematically distorted. Further, results of this study using actual supervisory ratings of subordinate performance tend to support previous laboratory-based findings with respect to the cognitive processes underlying performance evaluations.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of Lorne M. Sulsky, whose comments and suggestions regarding this article were significant.

REFERENCES

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level-of-processing theory and social facilitation perspectives. *Journal of Applied Psychology, 72,* 239–244.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77,* 975–985.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335–345.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40,* 307–322.

Browne, M. W. (1977). Generalized least-squares estimators in the analysis of covariance structures. In D. J. Aigner & A. S. Goldberger (Eds.), *Latent variables in socio-economic models* (pp. 205–266). Amsterdam: North-Holland.

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

Cooper, W. H. (1981b). Ubiquitous halo. *Psychological Bulletin, 90,* 218–244.

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organizational design. *Research in Organizational Behavior, 6,* 191–233.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127–148.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73,* 421–435.

Hauenstein, N. M. A., & Foti, R. J. (1989). From laboratories to practice: Neglected issues in implementing frame-of-reference training. *Personnel Psychology, 42,* 359–378.

Hayduck, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore: Johns Hopkins University Press.

Hulin, C. L. (1982). Some reflections on general performance dimensions and halo rating error. *Journal of Applied Psychology, 67,* 165–170.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10,* 311–321.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. *Research in Organizational Behavior, 5,* 141–197.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis assumptions, models and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL-VI: Analysis of linear structural relations by maximum likelihood and least squares methods.* Mooresville, IN: Scientific Software, Inc.

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

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

Lance, C. E., & Woehr, D. J. (1986). Statistical control of halo: A clarification from two models of the performance appraisal process. *Journal of Applied Psychology, 71,* 679–685.

Lance, C. E., Woehr, D. J., & Fisicaro, S. (1991). Cognitive categorization processes in performance rating: A replication and extension. *Journal of Organizational Behavior, 12,* 1–20.

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

Landy, F. L., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance settings. *Journal of Applied Psychology, 65,* 501–506.

Lingle, J. H., Geva, N., Ostrom, T. M., Leippe, M. R., & Baumgardner, M. H. (1979). Thematic effects of person judgments on impression organization. *Journal of Personality and Social Psychology, 5,* 674–687.

Lingle, J. H., & Ostrom, T. M. (1979). Retrieval selectivity in memory-based impression judgments. *Journal of Personality and Social Psychology, 37,* 180–194.

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 179–195.

Lord, R. G. (1985a). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71.

Lord, R. G. (1985b). An information processing approach to social perceptions, leadership and behavioral measurement in organizations. *Research in Organizational Behavior, 7,* 87–128.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69,* 147–156.

McKelvey, J. D., & Lord, R. G. (1986, April). *The effects of automatic and controlled processing on rating accuracy.* Paper presented at the meeting of the Society for Industrial/Organizational Psychology, Chicago.

Meister, D. (1986). *Human factors testing and evaluation.* Amsterdam: Elsevier.

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67,* 161–164.

Murphy, K. R., & Anhalt, R. L. (1992). Is halo error a property of the rater, ratees, or the specific behaviors observed? *Journal of Applied Psychology, 77,* 494–500.

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

Nathan, B. R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology, 75,* 290–296.

Phelps, R. H., & Schanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21,* 209–219.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581–588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38,* 76–91.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68,* 203–215.

Shweder, R. A. (1980). Factors and fiction in person perception: A reply to Lamiell, Foss, & Cavanee. *Journal of Personality, 43,* 455–484.

Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder & D. W. Fiske (Eds.), *New directions for methodology of behavioral science: Fallible judgment in behavioral research* (pp. 37–58). San Francisco: Jossey-Bass.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52,* 689–699.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96,* 58–83.

Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior, 17,* 357–367.

Woehr, D. J., & Feldman, J. M. (1989, April). *The effect of information processing objective* *on performance appraisal judgments.* Paper presented at the meeting of the Society for Industrial/Organizational Psychology, Boston.

Woehr, D. J., & Feldman, J. M. (1993). Processing objective and question order effects on the causal relationships between memory and judgment: The tip of the iceberg. *Journal of Applied Psychology, 78,* 232–241.

Woehr, D. J., & Lance, C. E. (1991). Paper people versus direct observation: An empirical examination of laboratory methodologies. *Journal of Organizational Behavior, 12,* 387–397.

Wyer, R. S. (1989). Social memory and social judgment. In P. R. Solomon, G. R. Goethals, C. M. Kelley, & B. R. Stephans (Eds.), *Memory: Interdisciplinary approaches* (pp. 221–242). New York: Springer-Verlag.

Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zedeck, S., & Cascio, W. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67,* 752–758.