Implications of Personality Theory and Research for the Rating of Work Performance in Organizations

Walter C. Borman
Personnel Decisions Research Institute

Borman first discusses implicit personality theory and its implications for performance rating. One implication is that the correlations between performance dimensions, especially interpersonal aspects, are probably inflated and conform to the evaluator’s theory of behavior. He suggests one method of coping with such rater distortion through an adaptation of behaviorally anchored rating scales in which behavioral examples of dimensions are recorded by raters. Borman next describes the relationship of the trait-situation controversy to work performance. In particular, he outlines Epstein’s argument for behavior stability across time and situation. He reviews research suggesting that there may be more consistency in work behavior than in behavior in general, probably because a smaller variety of situations is encountered in the work setting than in daily experiences, and relatively stable abilities underlie performance in most jobs. Borman introduces personal-construct theory and suggests applications of construct systems to perceptions of work behavior. One implication of this theory for performance rating is that we should alter the individual’s personal construct systems to conform to the performance areas with which the raters should be concerned in evaluating an employee’s performance.

Swedler stresses the need to attend to the distinguishing features or the content of the performance situations. He says that it is incorrect to believe that personalities and minds consist of abstract underlying structures that influence any task or problem. Instead, he focuses on the importance of the content of the situation as a determinant of various behaviors.

This discussion has three major sections. The first considers implicit personality theory (IPT) and its implications for performance rating. The second explores the
nature of work performance in light of the trait–situation controversy in personality psychology. And the third section introduces personal construct theory and speculates about applications of interpersonal construct systems to perceptions of work behavior. The purpose of the paper is to examine each of these areas in personality theory and research, first to help provide new ways of looking at issues and research in performance measurement, and second to suggest potentially fruitful directions for performance-appraisal research and practice.

IMPLICIT PERSONALITY THEORY

In general, research on Implicit Personality Theory (IPT) has been concerned with persons' perceptions of relationships between traits—that is, with how people construe traits to covary in other persons (Schneider, 1973). In this section on IPT, I focus on: (1) studies of the halo effect (e.g., Berman & Kenny, 1976; Koltuv, 1962; Nisbett & Wilson, 1977); (2) investigations of factor structure similarity in personality ratings with differing ratee samples (e.g., Norman, 1963; Passini & Norman, 1966); and (3) studies of the systematic-distortion hypothesis along with the issue of IPT's "validity"—that is, how closely IPTs reflect reality in stimulus persons (e.g., D'Andrade, 1965; Lay & Jackson, 1969; Mulaik, 1964; Shwedler, 1975; Shwedler & D'Andrade, 1980).

Studies of the Halo Effect

Researchers have long noticed what seems to be inflated correlations between dimensions in ratings made of others' personalities or job performances (e.g., Newcomb, 1931; Thorndike, 1920)—that is, correlations higher than warranted by actual ratee behavior. Of course, one problem with studying halo is that the actual "true" level of the dimension intercorrelations is not typically known, rendering impossible any precise estimate of halo (Borman, 1975; Johnson, 1963; Schwab, Heneman, & DeCotiis, 1975). As late as 1963, it could be stated that no experimental validation of halo had been offered (Johnson, 1963). However, in recent studies in which actual levels of relationships have been known (e.g., Borman, 1975; Nisbett & Wilson, 1977), halo has been found, suggesting that individuals do distort the magnitude of relationships between dimensions of personality and job performance.

An example of a study focusing on halo is found in a paper by Borman (1975). He prepared six vignettes of hypothetical first-line supervisor raters, using behavioral examples that had been sorted reliably into a single performance dimension, and assigned a narrow range of effectiveness scale values by individuals knowledgeable about the job. Each vignette consisted of six such examples, one representing each of six performance dimensions. The performance examples, then, essentially "told a story" about each supervisor. An important feature of this approach is that the mean effectiveness scale values of the examples served as performance "true scores" for each supervisor vignette on each performance dimension, and the variance of the true scores within ratee across dimensions, then provided the target against which the halo of individual raters' evaluations could be assessed. The key result of this study for our purposes here is that the variances of the subjects' ratings were significantly lower than the actual (target) variances, even after raters were instructed in a brief training course to avoid the halo error. Thus, raters generally failed to make distinctions between performance on the dimensions compared to the true score standard. Results of this study and the few other studies allowing an assessment of halo suggest that people do seem to overestimate relationships between dimensions in evaluating others.

Factor Analysis Studies of Personality Ratings

A series of studies have shown that the factor structure of personality ratings is relatively unaffected by how well raters know ratees. Passini and Norman (1966) provide an exemplar of this finding. These researchers asked college students who were unacquainted previously (they had only been together for about 15 minutes) to rate each other on 20 scales drawn from the Cattell Personality sphere. Ratings were intercorrelated, and the $20 \times 20$ correlation matrix factor analyzed. The factor structure was rotated to the varimax criterion and a five-factor solution appeared to be most psychologically meaningful.

The most interesting aspect of this study is that the factor solution obtained here is remarkably similar to solutions obtained when persons rate others whom they know well or with whom they are merely acquainted (Norman, 1963; Tudes & Christal, 1958, 1961). Importantly, a similar structure emerges when subjects are asked essentially to rate the semantic similarity of these trait terms (Hakel, 1974; Norman & Goldberg, 1966). Passini and Norman (1966) conclude that persons who have only the most superficial information about one another can draw upon their more or less comparable prior experiences and whatever easily observable cues are available to them to yield peer rating structures that are highly similar to those obtained from subjects who are intimately acquainted with one another [p. 48].

In other words, the similarities in factor structure imply that the raters are using their own categories to judge others—that is, their implicit personality theories. A conclusion that may be drawn from this work, then, is that rather than reflecting "real structure" in ratee behavior, the relational structure of personality ratings is largely a function of the expectations raters have of correlations between dimensions in ratees.

However, it is also possible to conclude that semantic structure of personality ratings reflects reality regarding correlations between dimensions in target
rates. It seems logical that persons in a culture develop ideas about semantic similarity of descriptors based on actual experiences observing others, and thus form "what goes with what" ideas veridical to the way characteristics of persons actually covary in human interaction. Work on the systematic-distortion hypothesis (D'Andrade, 1965; Shweder, 1975; Shweder & D'Andrade, 1980) addresses this issue more directly.

Systematic Distortion in Ratings
This research related to distortions in correlations between dimensions in the direction of *semantic relationships* between those dimensions.

D'Andrade and Shweder (D'Andrade, 1974; Shweder, 1975; Shweder & D'Andrade, 1979, 1980) have investigated this aspect of IPT. They examined data from studies in which observers made recordings of behavior immediately after viewing individual rates, and observers also made memory-based ratings on dimensions meant to reflect constructs tapped by behavior. As in the research just reviewed, results show that memory-based ratings (rated behavior) yield correlation matrices quite similar in structure to *semantic similarity relational* matrices. Most important, however, is that comparisons of actual (recorded) behavior to rated behavior yield relatively little correspondence. In the studies reviewed, correlations between the elements of these two matrices are low, suggesting that correlations among memory-based ratings on different traits tell us more about the *conceptual covariance* of these traits than about the way behaviors related to these traits actually covary in human conduct. An example may help the reader to understand this important finding.

Newcomb (1929) conducted a study on introversion–extraversion with two groups of boys in a summer camp ($N=27$ and $24$). In the course of the study, counselors kept daily records of each boy's behavior related to each of 26 "dimensions." Examples of the dimensions are: "speaks with confidence of his own abilities," "engages in group misdemeanors," and "reads a half-hour or more during the day." These records yielded scores on each dimension (actual behavior scores). In addition, six observers evaluated each boy on each dimension upon completion of the camp (rated behavior scores). Finally, Shweder had 10 graduate students rate the conceptual similarity between each pair of dimensions. Analysis of the data involved correlating the corresponding elements of the three correlation matrices for each group. Results showed that relationships between the rated behavior correlation matrices and the conceptual similarity matrices were strikingly similar, and importantly, the rated behavior matrices were relatively dissimilar in structure to the *actual behavior matrices' structure.*

This result is typical of what has been found in other studies with this kind of data. Shweder (1980) reports average correlations across seven studies as shown in Fig. 6.1.

A somewhat different approach to assessing the "validity" of IPT is provided by Lay and Jackson's (1969) study. For each of the 21 Personality Research Form (PRF) dimensions, they identified an item that correlated highly with that item's total score and then asked subjects to rate, for each pair of such items, the probability that a respondent would answer an item true given a response of true on another item. They found that a multidimensional scaling of interitem distances generated from these data provided results similar to the structure that emerges from actual PRF self-report data. The conclusion is that the relational structure of IPT (based on the probability ratings) corresponds closely to actual personality structure (PRF responses).

A study by Stricker, Jacobs, and Kogan (1974) used a similar approach, except data were analyzed for individual subjects rather than for the group. Results suggested that, as in the Lay and Jackson study, IPT was substantially congruent with dimensions generated from personality test responses.

Of course, a different interpretation of the Lay and Jackson and Stricker et al. findings is that test responses are affected by IPT just as the conceptual similarity judgments are assumed to be, and therefore, rather than serving as a criterion for
assessing the validity of IPT structure, self-report personality test structure may be likewise "biased" by IPT (Schneider, 1973).

On balance, research on systematic distortion suggests that our implicit personality theories may be substantially affected by semantic meaning, and by assumptions about the dimensions that belong together. Unfortunately, these assumptions may be in error. This, of course, casts doubt on the validity of our IPTs.

Halo and Systematic Distortion Considered Together

So far, I have discussed two different rating errors: halo and systematic distortion. It may be instructive now to examine effects on ratings of the two biases taken together. As mentioned, the term halo implies inflation of correlations among dimensions, whereas systematic distortion implies that raters make memory-based ratings, with correlations among dimensions biased in the direction of semantic similarity.

In the rated-behavior and actual-behavior correlation matrices available to the author, rating correlations are in fact higher than behavior correlations. For example, the median rating correlation in the Borgatta, Cottrell, and Mann (1958) study is .28 and for the actual behavior matrix the corresponding median correlation is .11 (using absolute values of the correlations because several of the dimension pairs are logical opposites). In the Mann (1959) data, I picked the six dimension pairs that have the highest conceptual similarity, and the average actual behavior interdimension correlation is .15.

In the rated-behavior and actual-behavior correlation matrices available to the author, rating correlations are in fact higher than behavior correlations. For example, the median rating correlation in the Borgatta, Cottrell, and Mann (1958) study is .28 and for the actual behavior matrix the corresponding median correlation is .11 (using absolute values of the correlations because several of the dimension pairs are logical opposites). In the Mann (1959) data, I picked the six dimension pairs that have the highest conceptual similarity, and the average actual behavior interdimension correlation is .15, whereas +.49 is the average correlation for the rated-behavior data. There are four dimension pairs that were rated as very dissimilar conceptually, and for those pairs the average actual-behavior correlation is −.19; the corresponding average correlation for the rated behavior is −.32. Thus, the inflation in correlations for the behavior ratings works both ways. For conceptually similar dimensions the ratings result in inflated positive correlations and for conceptually dissimilar dimension pairs, the ratings yield inflated negative correlations.

This is even more clearly shown in a table from Shweder (1975), reproduced here as Table 6.1. There data reflect the same phenomenon; ratings (and conceptual similarity judgments) overestimate the relationships between behavioral dimensions both when dimension pairs are logically similar and when they are

### Table 6.1

<table>
<thead>
<tr>
<th>Quintiles</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>5.2</td>
<td>.20</td>
<td>.53</td>
<td>.28</td>
<td>.29</td>
<td>.21</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>5.2</td>
<td>.34</td>
<td>.34</td>
<td>.21</td>
<td>.18</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>4.0</td>
<td>.28</td>
<td>.03</td>
<td>.37</td>
<td>.08</td>
<td>.23</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>2.9</td>
<td>.31</td>
<td>.27</td>
<td>.25</td>
<td>.02</td>
<td>.25</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>2.1</td>
<td>.20</td>
<td>.50</td>
<td>.43</td>
<td>-.10</td>
<td>.22</td>
</tr>
</tbody>
</table>

*Means and standard deviations for groups of rated behavior correlations, actual behavior correlations, and conceptual similarity scores divided into quintiles in terms of the degree of conceptual similarity for pairs of extraverted and introverted items in Newcomb's 26-item behavior list.

logical opposites. Apparently, both inflation in correlations and distortion towards semantic similarity occur in personality and behavior ratings, at least under the conditions studied.

"Direct" Validity of Ratings Considered

It is critical to note that what we have discussed so far refers specifically to "validity" of the correlational structure of the ratings—that is, correspondence between rated- and actual-behavior relational matrices. Of course, these findings have implications for validity of ratings on individual dimensions, but more direct evidence for validity comes from correlations between ratings on individual dimensions and some index of "true scores" on the corresponding dimensions. The D'Andrade and Shweder data provide one source of such validity estimates.

In a study employing a 30-minute videotape of family member interaction and offering both (1) on-line scores for "dimensions" (such as Inform, Question, Ridicule) by having raters study the tape and a transcript of that tape and also (2) raters' memory-based evaluations on these same dimensions, correlations between the ratings and "actual behavior" is in the neighborhood of .30 according to Shweder and D'Andrade (1980). In the Newcomb (1929) study, validities average .49.

It is very important to discuss these validities because they bear strongly on a point made by Shweder and D'Andrade (1980). The systematic-distortion extent of IPT does not specify that ratings are necessarily distorted and inaccurate, but that in the absence of relevant cues for a rater (whether there are a lack of cues or the cues have been attended to and forgotten), there will be nonrandom
distortion in the direction of semantic, "what goes with what" relationships between dimensions.

Critiques of the Systematic-Distortion Hypothesis

As might be expected, the systematic-distortion hypothesis has been criticized especially in what Cooper (1981b) has called the "strong form" (Shweder, 1975), and the conclusions that have been drawn, again, from the "strong" systematic-distortion position, have been questioned (Block, Weiss, & Thorne, 1979; Lamie, Foss, & Cavenee, 1980). Block, Weiss, & Thorne (1979) have provided the main attack on systematic distortion and correlational bias thinking and research. Their arguments, which are very detailed, are summarized here. Block et al. point to the nonequivalence of dimensions used in some of the studies D’Andrade and Shweder consider—that is, different definitions used by observers recording behavior for the "actual-behavior" data and raters generating the "rated-behavior" data. This could, of course, serve to reduce the validities discussed previously as well as the rated-behavior—actual-behavior matrix correlations. Block et al. also suggest correcting the rated-behavior—actual-behavior validities for unreliability, and point out that such reliabilities are not available for some of the studies.

Perhaps the most important concern voiced by Block et al. (1979) is that the "actual behavior" measures may serve as questionable criteria for ratings. The main problem has to do with the adequacy of simple counts of behaviors for reflecting a person's standing on important behavioral dimensions. The problem is very similar to industrial psychologists' quest for "true scores" of work performance against which to evaluate the accuracy of performance ratings or other measures of individuals' effectiveness. It is a tricky business to develop performance scores that on the one hand are arguably objective and nonarbitrary and at the same time reflect meaningful behavioral categories of work performance. Likewise, justifying a particular method of obtaining true scores for rates on nontrivial dimensions of personality-related behavior presents a difficult problem.

In my judgment, the criteria used in this research provide reasonable estimates of true scores for purposes of testing the systematic distortion hypothesis. Interestingly, scores were developed in the previously described videotape study utilizing a method similar to that employed in our research on performance ratings (Borman, 1978a). Dimensions were carefully defined for the judges who had ample opportunity to review the tape along with transcripts of the verbal behavior (just as our judges viewed tapes of the performances and studied verbal transcripts). Although this approach of true-score development is open to criticism, it may come as close as is possible to generating true scores on meaningful dimensions in personality and performance.

Summary of the Impact of IPT and Systematic Distortion on Ratings

Systematic distortion in its "weaker" form (Shweder & D’Andrade, 1980) seems to be an established component of behavior ratings. In the absence of relevant information about target rates or when sufficient time has elapsed between observation and rating to result in significant "memory decay," interdimension correlations of ratings are likely to be biased in the direction of semantic similarity.

Cooper (1981b) provides a review of possible reasons for halo and systematic distortion biases persisting in raters' evaluations of others. First, raters do not attend properly to their "hit rates." That is, we tend to underlearn our prediction errors from experience, and attend more to our successes in making correct perceptual judgments (Einhorn & Hogarth, 1978). Second, when we process interpersonal information, correlated features associated with a stimulus person are more likely to be attended to and remembered than uncorrelated features in the person (Tversky, 1977). And third, raters tend to adopt hypothesis-testing judgment strategies that confirm their prior beliefs about correlations between features (Snyder & Swann, 1978); essentially, we search for confirmatory evidence and then succeed at finding it. Finally, as perceivers, we tend to make initial judgments about others (usually in keeping with our prior beliefs of correlated features) and, importantly, then move to form a memory impression consisting primarily of summary information lacking in detail (Cantor & Mischel, 1979). These cognitive phenomena are likely to contribute to the kinds of biases and distortions we have discussed.

Also, Sawyer's (1966) assertions related to clinical versus statistical prediction have relevance here. He concludes that clinicians are "more likely to contribute . . . (to prediction of behavior) through observation than integration [p. 178]." In other words, the clinician (or rater) is reasonably good at observing and then recording observed behavior but poor at combining and integrating these observations to make summary evaluations of others. Thus, work on halo and systematic distortion along with certain supporting research in social cognition and clinical judgment is very suggestive of significant biases and distortions in ratings and limitations in raters' abilities to make evaluative judgments.

Implications of IPT Research for Performance Measurement

The most direct implication of IPT research for performance rating is that correlations between dimensions, especially in the interpersonal aspects of the job, are probably inflated and conform more to raters' implicit theories of work behavior than to actual behavior. Consider for a moment the raters' task in a
typical organizational setting within the context of the D’Andrade–Shweder research. The rater usually attempts to remember a ratee’s performance over a six-month or even year-long period. He or she may have notes and documents to help remember specific instances of performance, but often the ratings are based on memory of impressions developed over time. On the other hand, in most work settings there is considerable opportunity for the rater to gather performance information on a ratee.¹

My guess is that a fair test of the systemic distortion hypothesis with memory-based performance ratings made by raters on the basis of retrospective examination of performance over time will reveal that this kind of distortion does indeed occur with evaluations of performance. We have seen that performance ratings tend to yield inflated interdimension correlations, and some distortion towards semantic similarity in ratings is also likely to occur in my judgment.

Consideration of halo, systemic distortion, and possible sources of these biases by Cooper (1981b) and Feldman (1981) yield definite implications for the way in which performance information on individuals should be gathered. Research reviewed so far suggests that raters do a poor job of remembering, evaluating, and then combining and integrating behavioral information generated during interpersonal activities. One remedy for this is to take the rater out of the judgment process as much as possible.

Borman (1978a), Cooper (1981b), Feldman (1981), and Landy and Farr (1980) have discussed the performance-rating judgment process, and they provide some suggestions on how to reduce these distortions or biases are likely to occur. Borman (1978a) presented a three-step model of the rating process. The first step involves observing individual work behaviors, the second, making judgments about the effectiveness of each behavior, and the third, weighting and combining these judgments to form a single rating on a dimension. We have seen that it is during the first step in this process that the rating is least sensitive to distortion. These two subsequent steps require the kinds of judgment and integration that are most problematic.

Cooper (1981b) elaborated on Borman’s depiction of the rating process with an 11-step sequential model. He also describes the kinds of random and systematic errors likely to enter into the process at each step. Only the first two steps appear unlikely to generate the types of biases we have discussed: observing ratee actions and encoding these observations into short-term memory (perhaps

¹Cooper (1981a) at least partially addressed the question of systemic distortion in performance ratings by studying the D’Andrade–Shweder paradigm in the job-performance domain, but results were not in keeping with the systemic distortion hypothesis. The conceptual similarity matrix predicted the rated behavior matrix to a degree (r = .55), but the rated behavior–actual behavior matrix correspondence was higher (r = .89). As Cooper (1981a) points out, however, this was probably not an adequate test of systemic distortion in performance rating data because the rated behavior ratings were recorded immediately after viewing each ratee, perhaps before memory decay became a factor.

6. IMPLICATIONS OF PERSONALITY THEORY AND RESEARCH

recording them in some way, as well). Subsequent steps in the model refer to short- and long-term memory decay, retrieval of the (decayed) information from memory, comparing remembered observations and impressions of the ratee with rater and rating-scale standards, and actually making the rating. All of these steps are problematic in the context of our discussion.

Likewise, Feldman (1981) offers a model related to the cognitive tasks a rater must perform in making performance ratings. He or she must attend to relevant information about ratees, organize and store this information for later recall, and integrate the information into a summary judgment (rating). In light of the difficulties previously documented, the integrating and summarizing operations will give the rater the most trouble.

The Landy and Farr (1980) model is much more ambitious and attempts to reflect the entire process of generating and using performance ratings in organizations. Focusing on the rating-judgment process aspect of the model, the observation/storage and retrieval/judgment steps represent cognitive operations in which the rater must engage during the rating process. Clearly, the observation operation is the only one in this sequence requiring minimal integration of performance information and is therefore likely free from serious distortion.

Suggested Performance Measurement Methods
Offered in Light of IPT/Systematic Distortion Findings

What rating methods come closest to dealing successfully with these apparent limits in raters’ ability to make error-free interpersonal judgments? Within the now-familiar behaviorally anchored rating scale (BARS) strategy, the original guidance provided by Smith and Kendall (1963), recently updated by Bernardin and Smith (1981), is in keeping with the view of the rating-judgment process discussed here. Raters should record behavioral examples of employee performance relevant to each dimension. If raters do a good job of sampling the work behavior related to each dimension and provide a rich enough description of each behavior including the context of the action, then other persons knowledgeable about the job can evaluate the effectiveness levels reflected in each behavior, and the means of these effectiveness judgments.

A second approach in keeping with the seeming limits of raters’ evaluative capabilities is behavioral analysis or behavior assessment, applied to the performance-measurement problem by Komaki (Komaki, Collins, & Thoene, 1980) and Cone (1980). A brief description of one Komaki study characterizes this strategy. In her work with the Marine Corps, Komaki studied preventive maintenance jobs and, as part of that work, developed a behavioral observation performance-rating system. The system was designed to evaluate a unit rather than individuals in the unit, but some parts of the rating system are appropriate for measuring individuals’ performance. One dimension, Action Taken, required observers to sample a number of items in need of repair and to evaluate the
preventive-maintenance actions taken using a carefully prepared checklist containing the correct or desired actions. In this manner, highly reliable and objective "ratings" of performance were generated. One might question the proportion of the relevant job performance domain that we might expect to cover using such methods, but for those jobs and dimensions adequately covered, the system certainly avoids difficulties with complex judgment-process steps.

Cone (1980) has also discussed the relevance of behavior assessment for performance measurement. He describes important issues that point up the positive and negative features of using behavior-assessment principles in measuring performance (Cone, 1980; Foster & Cone, 1980). A critical issue is validity of measurement. This refers both to accuracy of the observations with respect to what the researcher is intending to measure and to ecological validity, or how well the measurement system "covers" the constructs of interest. Regarding the first type of validity, as in research in performance ratings, it has often been more convenient to study not validity or accuracy, but some other characteristics of the ratings. Thus, for example, interobserver reliability has been examined, and one of the important findings is that agreement in ratings using even carefully developed behavior-coding systems typically starts out high, but "observer drift" (e.g., Kent & Foster, 1977) reduces this interobserver agreement as raters forget details of the coding system and/or develop idiosyncratic interpretations or elaborations of the system.

Problems with ecological validity have to do with selecting the proper behaviors to assess: behaviors that are relevant to the construct of interest. Major difficulties are likely to occur if the target construct is complex. In addition, there is the issue of which behavioral property to measure (Foster & Cone, 1980). For example, frequency, latency, and duration may be important, as well as qualitative aspects of the behavior. Unfortunately, the familiar fidelity-bandwidth problem is at work here. Very narrow (and in many cases trivial) behavior categories can be measured very precisely and unambiguously, whereas more complex (and often more psychologically interesting) categories are more difficult to measure.

Thus, we have seen that behavior analysis and assessment principles applied to performance measurement are potentially in concert with this view of performance raters' cognitive limitations related to evaluating employees. Using these principles, raters might view large samples of behavior on a job, and form behavioral categories and rules for sorting behaviors into categories. They would then "score" individual job incumbents on each category by observing each one in representative samples of situations and classifying and counting behaviors according to the scoring system. Such behavioral-measurement systems have actually been placed in operation by Komaki and her colleagues (Komaki, Blood, & Holder, 1980; Komaki, Collins, & Thoene, 1980), demonstrating that for at least some components of jobs, this kind of behavioral-measurement method might be very useful.

A previously mentioned reservation regarding behavior-analysis approaches to performance measurement is the likely difficulty in adequately measuring all elements in the performance domain for, especially, complex jobs. Consider, for example, trying to develop a behavioral-measurement system to tap successfully all important performance constructs on a management job. Some parts of such a job may be amenable to the method (e.g., interpersonal effectiveness at meetings with subordinates might be indexed by noting eye-contact behavior in support of subordinate statements, proportion of time subordinates have the floor, etc.), but the coverage of other important performance areas may well be more problematic (e.g., decision making).

Proposal for an Idealized Performance-Measurement System

An adaptation of the BARS approach referred to previously seems to offer clear advantages for an idealized performance-measurement system intended to derive relatively pure performance scores for individuals on a job. The idea of such an idealized system is to derive performance scores that are not dependent on rater perceptual processes we have discussed, but that are based on behavioral data sufficiently rich and detailed to render a complete and fair depiction of job performance. The proposal I discuss is impractical for application to on-going performance-appraisal systems in organizations, although features of the approach can be incorporated. Instead, the intent is to gain a glimpse of the actual structure of work behavior on some job to see how closely it comes to the structure inferred from performance ratings—that is, to obtain an estimate of the degree of distortion ratings provide in our "ratings picture" of job performance.

I believe that the strategy of gathering large and representative samples of ratee-performance examples provides a potentially rich source of highly relevant raw behavior data that can be used in this idealized system. Features of behavioral examples, as formulated by Flanagan (1954) and Smith and Kendall (1963), seem well suited for providing usable raw behavioral units. Actual ratee behavior along with the context in which this behavior occurred are both represented in such examples.

The most important reason for favoring an approach employing behavioral examples for this purpose is that the entire performance domain is likely to be sampled. In our work with behavior scaling (e.g., Borman & Dunnette, 1975), we have been most impressed with the coverage of the performance domain provided by a sampling of behavioral examples generated by persons knowledgeable about the job. The inductive approach of sampling widely each ratee's behavior relevant to each area of performance in a form that includes information on the context appears to offer a potentially excellent behavioral unit for analysis, with considerable advantages over counts of specific behaviors.

Proceeding with this idealized proposal, to obtain an estimate of the actual covariance structure of performance on a job, we could first record over a period of time behavioral examples of actual performance for a number of individuals on a target job. I realize this is a tall order. It would require an extensive time/
situations sampling plan to obtain a representative sample of behavioral examples for each job incumbent. Indeed, I suspect that the only way to generate such samples of behavioral information is to videotape on-going work behavior in a natural setting so that the tapes and transcripts of the verbal behavior can be studied to extract the examples. Perhaps performance in the Center for Creative Leadership’s “Looking Glass” organizational simulation should be videotaped. This simulation is designed to present a diverse sampling of management tasks, it is realistic in portraying situations managers typically face, and it provides a relatively long stream of behavior that would help generate the necessary large number of examples.

The second step in this approach is to content analyze the examples generated to form relatively specific performance dimensions. Next, it would be possible to get persons familiar with the job’s performance requirements to sort each performance example into a dimension and assign an effectiveness level to the behavior. Finally, incumbents could be scored on each dimension by averaging the effectiveness levels (perhaps weighted in some way) for all of their examples on each of the dimensions. With a large enough sample of incumbents, this data matrix can yield a rough estimate of a “true” performance-covariance matrix, and this matrix might be compared to covariance matrices generated from performance ratings assigned using typical performance-appraisal methods.

If performances in the “Looking Glass” or some other simulation were videotaped, we might have subjects viewing these performances and making ratings under varying conditions, and these ratings could then be compared to the target behavior-performance scores. It would be important in such research to follow Feldman’s (1981) advice to give rater subjects other tasks besides their behavior-observation-performance-rating task to increase the realism of the setting for performance appraisal and, along with it, the generalizability of results to actual job applications. In sum, I recognize that this plan is quite cumbersome, but it is offered in the spirit of obtaining experimentally an idea of the “true” covariance structure for a job and providing reasonably realistic stimulus materials to enable study of rating distortions and accuracy in ratings.

A Different Reaction to “Negative” IPT Research Results: Studying the Performance-Rating Process

In this section of the paper, performance-measurement systems that avoid the rating process have been presented, these systems keeping performance scores as much as possible untainted by rater biases that have been documented in IPT and related research. A completely different approach is to actively study the rating process to learn as much as possible about it, the goal being to make use of this knowledge to change performance-appraisal practices.

Indeed, one purpose of this paper is to stimulate research on process by introducing research strategies and findings from work in personality psychology into performance appraisal. Likewise, recent papers in Industrial-Organizational psychology journals have begun to specify how the performance-rating process might be studied: (1) from an attribution perspective (Feldman, 1981; Green & Mitchell, 1979; Stone & Slusher, 1975; (2) in relation to interpersonal constructs and category systems (Borman, 1978a; Feldman, 1981; see also the third section of this paper; (3) with a focus on rater individual differences correlates of rating behavior (Borman, 1979; Schneier, 1977); and (4) within a policy-capturing framework (Zedeck & Kafry, 1977).

Another possible approach to studying the performance-rating process is to study “rating styles”—that is, different strategies for making ratings, preferences for using certain performance cues, methods of combining information to arrive at ratings, and so on. A recent study characterizes this approach and demonstrates how difficult such research may be in comparison to “product” research.

In a doctoral dissertation study, Banks (1979) devised a procedure to identify the behavioral cues used by raters. In particular, Banks’ research examined: (1) interrater agreement in specific cues used to make ratings on individual dimensions; (2) within-rater consistency of certain rating-style variables; and (3) differences in outcome-rating errors (e.g., halo) made by persons classified according to folk concepts about effective rating styles.

In this study, Banks had subjects view each one of six of the 5-to-9-minute manager performances on the Borman (1978a) videotapes and rate each performer on only one of six performance dimensions, such that different dimensions were rated each time. Subjects had before them computer consoles with seven buttons corresponding to seven effectiveness levels (1 = very ineffective; 7 = very effective), and they were instructed to press one of these buttons each time they viewed behavior they thought was relevant to the performance dimension being considered. The particular button pushed (of the seven) indicated the rater’s judgment of the behavior effectiveness. Also, the buttons were attached to a timing device that provided an exact record of where in the tape each button was pressed. Finally, subjects were instructed to provide a brief verbal description of the behavior they were attending to each time they pressed a button.

Results of the study showed first that different raters tend to identify different behaviors as relevant for making evaluations on a dimension. Apparently, the subjects looked for or attended to a substantially separate set of cues when searching for behavior relevant to performance on a dimension. Second, the within-subject, across-trial reliability of certain rating-style variables was high: for number of cues attended to (button presses per tape), the intraclass r = .95, for response latency (time to first button push), r = .77. Also, verbal reports were classified according to: (1) global versus specific; (2) behavioral versus evaluative; and (3) descriptive versus prescriptive (what the rates did versus what they should have done), and within-rater stabilities of the proportion of global and behavioral cues were high (r = .88 and .85, respectively). Finally,
raters were classified as effective or ineffective according to "folk concepts" of
good rating styles—that is, using relatively specific (versus global) information,
employing many (versus few) behavioral cues, and attending to both effective
and ineffective performance within ratee. Raters with relatively effective styles
in this regard provided ratings with lower restriction of range error. In this
research, stability of some of the style variables is especially heartening, suggest-
ing that to some extent "rating styles" in appraising others' performances are
relatively consistent and thus may reflect meaningful individual differences in
behavior related to the performance-rating process.

Summary
Thus, we have seen that work on IPT and related areas in personality and social
psychology have important implications for research and practice in performance
appraisal. We concluded that relatively serious errors, biases, and distortions are
likely to occur in performance ratings as they are usually generated, and that one
way to address this problem is essentially to avoid asking raters to perform
memory and judgment activities on which research suggests they do poorly.
Adaptations of selected performance-measurement methods were offered relative
to this avoidance theme. Also, an idealized study to map the performance space
in a job was presented as a possible way to obtain relatively objective, yet rich
and relevant performance information on individuals in jobs.

In addition, a strategy was briefly presented to study the performance-rating
process in a long-term effort to learn enough about this process to improve the
accuracy of ratings in the future. We are beginning to specify research directions
focusing on the rating process, and an example of research on "rating styles"
showed how difficult yet potentially rewarding such research might be. Of
course, it is also possible that such research will suggest that performance-rating
systems should not rely on rater judgments. Nonetheless, research in this direc-
tion should proceed.

THE TRAIT-SITUATION CONTROVERSY IN
PERSONALITY PSYCHOLOGY AND
CONSISTENCY-VARIABILITY IN EMPLOYEE JOB
PERFORMANCE

The trait-situation "controversy" in personality psychology relates to disagree-
ment about whether stable, enduring traits in people largely dictate behavior, or
whether situations, differing meaningfully in stimulus characteristics, determine
behavior to a great extent. A third interactionist position hypothesizes that different
situations elicit different behaviors in different people. That is, to understand
human behavior, we must attend to both persons and situations.

6. IMPLICATIONS OF PERSONALITY THEORY AND RESEARCH

The Positions

Trait Position. The trait position contends that relatively stable individual
differences in personality have broad meaning for the way persons conduct
themselves in different situations. If we know how a person stands on certain
relevant traits, we can predict how that person will act in various situations.
Selected arguments for this position from Bowers (1973), Epstein (1979),
Stagner (1977), and others are as follows:

1. Many of the studies that demonstrate low across-situation correlations of
traits utilize very different situations, and thus variability in the stimulus character-
istics is greater than what is typically encountered in everyday life (Bowers,
1973).
2. Laboratory studies usually place subjects in situations whereas in real life
we can to some extent choose our environment, which in turn leads to stability in
our behavior (Bowers, 1973; Epstein, 1979).
3. Assessment of personality has shown relatively high temporal stability in
traits, and this suggests, albeit indirectly, that traits are important determiners of
behavior (Bäckman & Magnusson, 1981; Block, 1971; Olweus, 1979).
4. The unit of analysis employed in research favoring a situationist position is
often focused at the individual-behavior level rather than at a level that considers
behavioral tendencies that are likely to be more stable (Bowers, 1973; Epstein,
1979).
5. Traits have been reasonably successful at predicting important outcome
variables, more successful than situationists have claimed (Hogan, DeSoto, &
6. Multitrait-multimethod studies of trait measures have shown relatively
high convergent validity (Block et al., 1979; Stagner, 1977).

As can be seen, some of these points contradict pro-situationist research whereas
other points support the usefulness of trait concepts.

Situationist Position. The situationist position contends that there is little
stability in personality; instead, the situation and its stimulus characteristics
largely dictate a person's behavior. Mischel (1968) has often been cited as the
main proponent of this view, although it should be noted that his position is far
from one of extreme situationism (Mischel, 1979; Mischel & Peake, 1981).
The main prostitutionist arguments are as follows:

1. Behavior related to a trait in one situation rarely correlates highly with
other behavior related to the same trait in other settings (Mischel, 1968).
2. Scores on trait measures correlate minimally with behavior in specific
situations (Mischel, 1968; Vernon, 1964).
3. ANOVA analyses involving person, situation, and person–situation interaction components often show situation (and the person–situation interaction) accounting for considerable variance in behavior (Endler, 1966; Magnusson, 1971).

4. Ratings of others, sometimes used to help support a trait position, reflect more stability in rate behavior than is actually warranted (Jones & Nisbett, 1971; Shwedler, 1975).

Interactionist Position. This position contends that consideration of only the person or only situations is insufficient for fully understanding human conduct. Behavior is a joint function of the person and the situation. What this means is that the rank order of individuals on a trait may be relatively consistent, but only within certain situational constraints. To learn more about human behavior we therefore need to classify situations in some way in addition to classifying behaviors. Trait-oriented psychologists have identified stable clusters of behaviors designated as traits, which are useful when employed in research, but not as much has been accomplished regarding a similar classification of situations (Bem & Funder, 1978).

Unfortunately, the latter classification task is made even more difficult and complex when we consider the impact of persons' perceptions of situations over and above the objective situational characteristics. Perceptions related to situations are very likely to intervene to help determine a person's behavior in a situation, and these perceptions may vary across persons even when the objective situation is held constant. One kind of evidence offered in defense of the interactionist position is results of studies employing ANOVA designs indicating that the person–situation interaction term accounts for more variance than either the person or the situation terms (Bowers, 1973; Endler & Hunt, 1968, 1969). Interpretation of data in these ANOVA studies has been questioned (Golding, 1975), although reanalysis using a generalizability approach confirms the importance of such interactions.

Another view of the person–situation interaction, a view that is different from the ANOVA interaction concept, is that situations and persons are interdependent—that is, relationships between person and environmental variables are reciprocal, and causation works in both directions. Person variables “cause” environment through differing perceptions of that environment, and the environment has a causative effect on person variables in that it affects behavior in meaningful ways. The latter interpretation suggests that to some extent the environment (perceived) is a function of person variables and standing on person variables is a function of the environment (Buss, 1977).

Several writers have pointed to the importance of “getting beyond” the simple debate of which term is most influential: person, situation, or their interaction. In the first place, one could design experiments to show any effect by manipulating the variability in person or situation variables (Bowers, 1973; Epstein, 1979; Golding, 1975; Mischel, 1973). Across a relatively homogeneous set of situations, person variance is likely to be most influential; studying a homogeneous group of subjects in widely divergent situations will “demonstrate” the importance of situations. What is needed in this area is information about how persons and situations interact to determine behavior (cf. Endler, 1973).

The present lines drawn in the controversy appear to be these. Mischel (Mischel & Peake, in press), Fiske (Fiske, 1974), Shwedler (Shwedler & D'Andrade, 1980), and others devoted to what we might call a situationist–interactionist interpretation of personality believe that cross-situational consistency in behavior is not sufficient to warrant the use of broad trait measures in personality psychology. Those psychologists more of the trait persuasion—call them trait–interactionists to reflect their appreciation of “the situation,” though to a lesser degree—believe that traits do serve a useful function in personality theory and research. We cannot resolve the issues here. Both sides agree that more data are needed to shed light on the issues. What we can do is outline a couple of the most intriguing directions research and thinking are taking in this controversy.

Developments in the Trait–Situation Controversy

Epstein (1979, 1980) has presented a seductively simple argument that largely explains away inconsistency in behavior across time and situation. He suggests that unreliability of behavioral measures used in personality research favorable to the situationist position is the main reason for low correlations between behaviors across time and situations. Epstein demonstrates his “solution” by using aggregated, reliable measures of (1) self-reports of emotions (e.g., happy–sad, kind–angry); (2) rated behavior (e.g., number of times subject initiated contact with others); and (3) objective behavior (hours of study). His findings suggest considerable stability in emotions and behavior over time, much greater stability than when single (and unreliable) indices are used. In another study he finds that self-report personality scales correlate moderately well with reliable self-kept records of emotional states and objective events. Importantly, Epstein (1979) argues that cross-situational stability has been demonstrated because reports of the behaviors and emotions were elicited in a range of situations representative of everyday life (they were recorded daily for about two weeks). In sum, Epstein acknowledges the importance of the situation, but believes that response dispositions—that is, relatively broad behavioral traits—can on the average explain much of behavior.

Mischel has commented on Epstein's research, and, predictably, has a different view of this work (Mischel & Peake, 1981). He first acknowledges that temporal stability in behavior within narrow ranges of situations is likely to exist as long as reliable measures of the behaviors are taken. The real issue, he
believes, is the cross-situational consistency of behavior, and his reading of Epstein's data is that they are not impressively supportive. For example, Mischel & Peake (1981) eschew the personality-scale self-rating data offered as support for this kind of stability and point out that correlations between personality scales and reliable objective behavior indices are low.

Further, they present their data from multiple assessments of college students' conscientiousness. Examples of variables tapping this construct are: psychology class attendance, study session attendance, and assignment punctuality. In all, 19 such variables were considered, and aggregated measures of each were intercorrelated. Results indicate that considerable reliability is present for individual measures (mean reliability of the across-time composites = .69), but that the average correlation between indicators is relatively low (r = .20, corrected for attenuation). Thus, the debate continues, but importantly, more data are being generated to help us understand better the psychological significance of consistency and lack of consistency in behavior.

A second development, intriguing in its implications for the trait-situation controversy, involves the search for individual differences in cross-situational consistency. Bem and Allen (1974) asked subjects to rate the consistency of their own behavior related to friendliness and conscientiousness in different situations and found that for those who reported high consistency, behavior across situations related to these traits was in fact relatively stable. Lower stability coefficients were obtained for individuals reporting low consistency. Thus, Bem and Allen (1974) concluded that we may be able -- to predict some of the people some of the time"—namely, those who are more consistent in their behavior patterns related to individual traits.

The self-monitoring individual differences variable (Snyder, 1974) may be important in this regard. Theoretical considerations led Snyder to develop the self-monitoring scale to identify persons tending to monitor closely situational cues in their interpersonal interactions (high on the scale) and persons who are less apt to rely on situational information for cues on how to act during interpersonal dealings (low on the scale). Accordingly, one would predict that persons higher in self-monitoring should be less consistent than those low on the variable in terms of their behavior in different situations. Snyder and Monson (1975) confirmed this when those relatively high in self-monitoring showed more variability in self-reported likelihood estimates of performing different behaviors in different situations.

More recently, Schneiderman (1980) found higher correlations between across-trial variability scores derived from objective indices of behavior (e.g., time of eye contact, time talking in a situation) than from ratings, confirming that intrapersonal variability—consistency has some generality. He also found evidence suggesting that variability may be somewhat content specific. In other words, we may tend to be very consistent in one domain of behavior, but more variable in other behavioral domains.

Implications of the Trait-Situation Controversy

I believe that the trait—situation controversy contains lessons for performance-appraisal research and practice. In this section variability in performance across time and situations is considered and attempts to develop a measure of it are presented.

When we ask a rater to make a rating of an employee on a single dimension of performance, we are really asking (I believe) for an average or modal effectiveness rating of that individual on that category of performance. In other words, the rater is implicitly asked to ignore variability in performance level, and instead somehow average the observed performance over the time period for the appraisal. Yet, clearly, performance levels of individuals on dimensions vary to some extent across situations and time. Just as we have seen that "the situation" is probably an important factor in determining behavior relevant to the personality domain, different situations faced by employees on the job may affect their performance. Also, as the research reviewed suggests, there may be reliable individual differences in the consistency of performance on jobs.

The issue of variability—consistency in performance is of great practical importance for some jobs. For example, the power-plant operator job requires steady performance in the technical competence and decision-making areas of job performance. An operator's performing successfully in those parts of the job "most of the time" is not sufficient. In light of our review of behavioral-consistency issues in the personality domain, it is important to ask how variable performance is likely to be.

Two factors make it likely that work performance is more consistent than behavior in general. First, the number of different situations confronted in a job is typically much less than the total range of situations faced in everyday life. Of course, this depends to some extent on the job. A salesperson responsible for many different kinds of products appropriate for a wide variety of people will have more situations in which to perform than will a factory worker assigned to a single machine. But, in general, the "situation space" at work will form a relatively small subset of a person's total situation space.

In addition, situations faced at work tend to be repetitive, further reducing the variety of situational contexts to be expected on a job. Thus, with a relatively homogeneous set of situations in which to perform, the variability in job performance for individuals is likely to be less than variability of personality-related behavior across all situations confronted by these individuals.

A second reason for expecting reduced levels of variability in performance for individual employees is that relatively stable abilities underlie performance in most jobs. As we have seen, there is disagreement about the stability of behavior dependent on personality-related variables, but there exists much more agree-
ment on the consistency of behavior dependent on abilities, especially intelligence (e.g., Mischel, 1968; Olweus, 1979). It seems clear that performance on jobs is more dependent on abilities than on the behavior personality psychologists have typically studied; although, again, this varies with different types of jobs. That ability and aptitude tests have often been robust predictors of job performance (see especially the validity generalization work: Schmidt & Hunter, 1977; Schmidt, Hunter, & Pearlman, 1981) supports the contention that these kinds of variables underlie much of performance effectiveness. Therefore, we might expect relatively high consistency in work performance because of this dependence on abilities. If this notion has merit, it also follows that on jobs with both technical and interpersonal components, performance on the technical part is likely to be more consistent than performance in the interpersonal aspects.

We have argued that reduced “situation space” and relatively stable abilities, which dictate performance levels to some degree, are two reasons to expect less variability in performance compared to variability and inconsistency in personality-related behavior in everyday life.

Another way to look at the variability in performance question is to consider the well-known model: Performance = Ability × Effort. As we have discussed, abilities are likely to be a stabilizing force in determining performance levels, but the motivation of individuals over time may be more variable and thus contribute to inconsistency in work performance. The main point to be made is that although stability in job effectiveness is likely to be considerable, there is also likely to be some variability in performance over time and in different situations. This kind of variability has been largely ignored, and yet information about distributions of performance will aid in describing a person’s job performance more completely and may help increase our general understanding of the nature of performance on jobs.

Consider, for example, two different performance distributions graphed in Fig. 6.2. Although these hypothetical workers have the same *modal* performance level on this dimension, B operates at close to his or her maximum performance most of the time whereas A operates at close to his or her minimum performance most of the time. Also, B is more consistent than A in performing on this dimension. Such distributions may, in addition, provide clues about the abilities and motivation of employees. A is clearly *capable* of performing at the highest level, but does not do so most of the time. B, on the other hand, is apparently limited in the level of performance possible (at least for the period rated) but works up to his or her level of capability the majority of the time. This may be stretching interpretation of such performance distributions, but the point is that distributions give us more information about the work performance of individual employees than do modal or average ratings.

**Indexing Variability—Consistency in Work Performance**

The most explicit treatment of performance variability in the performance-appraisal literature is provided by Kane (e.g., Kane & Lawler, 1978). Kane (1981) places present methods of eliciting performance information under a “Physical Measurement Model” and points out that the only parameter to be specified in this model is average performance. In his Distributional Measurement Model the standard deviation and skewness become parameters of interest. Kane views variability in effort or motivation and external constraints beyond the performers’ control as two important sources of variation in job performance, sources that should be accounted for in a performance-measurement system. His Performance Distribution Assessment method does just that.

With this method the rater is asked to record for each performance factor the percent time that each level of performance *cannot be attained* because of factors beyond the ratee’s control and then to note the percent time the ratee performed at or above each successive performance level. Kane (1981) provides formulas that allow computation of an average performance score, a consistency in performance score, and what he calls a negative range avoidance score, an index of how successfully the ratee avoids performing poorly. Thus, in Kane’s system, variability in performance is explicitly addressed and indexed. At issue, of course, is the operational usefulness of the system. For example, can raters make reliable judgments of these percents as demanded by the method? Research is overdue to test this system in operation.

I too have been interested in variability in worker performance. I have been particularly interested in examining the interrater reliability of two simple methods of assessing performance variability—consistency. My concern is that judgments of variability in performance might be awkward for raters, foreign to the way they think about performance, and therefore difficult for them to make.

To try out these methods a study of a sample of Navy recruiters was conducted that provided peer and self-ratings of the recruiters in their same stations.
that raters may have difficulties in providing reliable ratings of performance variability—consistency. Future research should evaluate methods of indexing individual employees’ performance distributions.

PERSONAL-CONSTRUCT THEORY

The Theory in Relation to Person Perception

George A. Kelly (1955) developed a sweeping psychological theory about humans and how they function. A full description of his theory is beyond the scope and intention of this paper, but those aspects of the theory that pertain to person perception are important for our purposes. Kelly states in his organization corollary that each person characteristically evolves, for his or her convenience in anticipating events (or other persons' activities), a construction system reflecting his or her personal way of viewing and interpreting those events. This means that individuals develop personal-construct systems, categories that they use to judge events and to make predictions about future events, and that they impose some of these categories on their person perceptions. That is, individuals see other persons through their own “interpersonal filters,” and they interpret what they see according to these target persons’ perceived standing on each of their personal constructs. As Kelly (1955) puts it, “... man creates his own ways of seeing the world in which he lives; the world does not create them for him. He builds constructs and tries them on for size [p. 12].”

This view of personal constructs is in a sense close to the conception of implicit personality theory. Notice that like IPT, the focus is on the perceiver and the way he or she makes judgments about events and other persons. With respect to person perception, personal constructs are likely to affect the experience of observing others, to filter interpersonal information, or even to “bias” the perception of other persons. Of course, Kelly and his followers would not regard these personal constructs as biasing perception. But, the point is that both conceptions relate to interpersonal filters used by perceivers. Interestingly, personal-construct theory, much more than IPT research, examines individual differences in these filters and we explore this facet. Also, personal-construct research focuses on the content of the interpersonal dimensions as well as on their structure, and this creates additional important implications related to formation of performance judgments.

The serious student of personal-construct theory will notice that we refer here to only one part of a person’s personal-construct system—those categories that have as a focus of convenience events that involve interpersonal judgments. For example, we are not concerned with those constructs that might be used to judge such objects as music, art, or writing.
Research on Personal Constructs

Research in personal-construct theory has been stimulated by the development of the Role Construct Repertory Test (RCRT), a protocol that in its most common form asks a subject first to record the names of persons who fit each of a number of roles (e.g., mother, father, best friend, etc.). The respondent is then asked to consider various triads of these role persons, and for each triad to identify an important way in which two of these persons are alike and yet different from the third person. Taken together, the responses constitute the subject's personal constructs. Fig. 6.3 contains a sample RCRT protocol. Two lists of constructs are then presented in Fig. 6.4 to provide a feel for the nature of personal constructs.

Research in personal-construct theory has often employed the RCRT to help subjects generate their personal constructs. This work has proceeded in several directions. Lines of research include: (1) the study of thinking among clinical groups such as thought-disordered schizophrenics (Bannister, 1963; Frith & Lilie, 1972; Widdow, 1976); (2) an exploration of suicidal behavior (Kelly, 1961; Landfield, 1975); (3) the investigation of cognitive complexity (Bieri, 1961; Crockett, 1965; Metcalfe, 1974); (4) a study of the psychotherapy process (Fielding, 1975; Fransella & Joyston-Bechal, 1971); (5) an examination of interpersonal relations (Duck, 1973; Huston, 1974); (6) the study of meaningfulness of interpersonal categories and extreme response style (Cronwell & Caldwell, 1962; Hamilton, 1968; Landfield, 1965; O'Donovan, 1965; and finally (7) the investigation of individual differences in personal constructs (Hamilton, 1970; Kuusinen, 1969; Sechrest, 1968).

To provide a flavor for research in personal-construct psychology, we review a study examining the effect on ratings of the meaningfulness of dimensions employed. Using the previously described Rep Grid, Isaacs (1966) asked college student subjects to provide ten personal constructs related to making judgments about other people. Isaacs then had subjects rate ten acquaintances...
on their own personal constructs as well as on ten Manifest Anxiety Scale items and ten Semantic Differential dimensions. Results showed that ratings on the personal constructs yielded larger variances across the target rates than did ratings on the other dimensions. This has been a consistent finding (e.g., Adams-Webber, 1979; Bonarius, 1970); it appears that individuals prefer to use their own constructs to rate others (Bonarius, 1965; Landfield, 1965), and that they differentiate more finely between persons when employing their own dimensions. Presumably, over time an individual uses a similar set of constructs in making judgments about others, and thus grows used to and comfortable with employing these to evaluate others.\footnote{This is not to say that in very different interpersonal situations (e.g., in a heated argument with two or three others versus initially meeting individuals at a party) a person uses the same or even very similar constructs to make judgments about others. Kelly recognized that the content of personal constructs depends to some extent on the situation the perceiver faces. To a degree, different contexts demand different sets of categories if relevant interpersonal judgments are to be made. However, personal-construct theory posits a reasonably stable core set of interpersonal dimensions an individual uses to make judgments about people within a delimited set of situations.}

An important consideration in personal-construct research (and IPT research, as well) is individual differences in constructs and IPTs. The strategies for studying such individual differences can be divided into four categories: (1) studies of the underlying structure of IPT across all persons; (2) investigations of different "points of view" related to IPT; (3) comparisons of dimensions underlying different individuals' interpersonal categories; and (4) studies related to individual differences in constructs/IPTs using free-response methods.

Structure of IPT/Personal Constructs. The studies reviewed in this section are not usually considered to be related to questions of individual differences in IPT. Yet, conclusions drawn in this line of research imply that few if any individual differences in IPT exist. The Passini and Norman (1966) study reviewed previously is an example of such a study. Conclusions from this and similar studies (e.g., Norman, 1963; Tepes & Christal, 1958, 1961) imply that there exists a "universal IPT," with individual differences nonexistent or at least unimportant. Of course, an approach that uses group data to study individual differences does not really focus on these differences and may well overlook them.

Different "Points of View." Individual differences multidimensional scaling has been used to identify groups of subjects who differ in their patterns of perceived similarities between traits. Essentially, the multidimensional scaling analysis sorts individuals into groups according to similarities in the ways they have made judgments about how traits covary. Individuals clustered into a group are then said to have a certain "point of view," a particular way of viewing trait interrelationships—in short, a single IPT.

In this kind of study, subjects are given traits in some protocol asking them to judge the similarity of each pair of traits. The individual differences multidimensional scaling procedure then clusters subjects as previously described, and often personality and other variables are correlated with the different points of view. Using this paradigm, Pedersen (1965) found three "points of view," one of them associated with authoritarianism. Walters and Jackson (1966) identified two groups of subjects and labeled the critical dimensions that separated these groups as Interpersonal Affectivity and Harmfulness. Finally, Messick, and Kogan (1966) identified a number of "points of view" for both males and females, but relationships between personality variables and dimensions separating the groups were low and difficult to define. Even though nothing definitive has been concluded from such studies, the finding of subgroups at least tends to call into question the universality of IPTs.

Dimensions Underlying Interpersonal Categories. In studies of these dimensions, subjects are given interpersonal categories and are then asked to rate the same stimulus persons (or roles such as insurance salesperson, factory worker) on each of the categories. Correlations between categories are computed for each subject and factor analyses of each subject's correlation matrix are compared. The two most widely cited papers using this approach draw very different conclusions about the extent of individual differences between subjects' underlying categories. Kusinen (1969) compared each subject's factor solution to the group solution (pooled across all individuals in the sample) and concluded that no meaningful individual differences were present. However, Hamilton (1970) faulted Kusinen on his method of assessing factor congruence. He reanalyzed Kusinen's data and supplied data on his own to reassess the magnitude of these individual differences. Applying Tucker's coefficient of congruence (Harman, 1967), a factor similarity index, and making comparisons between individuals, Hamilton found wide individual differences in both samples. He concluded that substantial differences are present in the underlying dimensionality of interpersonal categories for different perceivers.

Free Response Methods. Several psychologists interested in IPT have called for more emphasis on free response in generating subjects' IPTs or personal constructs. For example, Cliff (1973) suggested that when subjects all use the same traits, the IPTs emerging from the rating task are artificially restricted and may unduly reduce many of the individual differences that might appear in a less structured task. Cronbach (1958) too has suggested more use of free-response protocols to elicit individuals' IPT. And, in his review of IPT research, Schneider (1973) concludes, "... so long as such (person perception) research provides perceivers with categories of explanation, it can never answer the question of which categories are 'naturally' used by perceivers [p. 307; parentheses added]."
But, perhaps the strongest proponents of free response in this domain are personal-construct theorists. Indeed, one of the central reasons for the development of the RCRT was to elicit in a clinical setting the patient's personal categories couched in his or her own terms. Kelly (1955) believed that personality assessment must take into account the person's own abstractions and generalizations about him- or herself, even when they are in rough form. Thus, a number of writers have pleaded for more use of free-response methods in describing IPTs or personal-construct systems.

Yet, only a few studies focusing on individual differences have used free-response methods. Sechrest (1968) asked nursing students to form personal constructs and found some overlap in the content of these categories but also substantial individual differences. Rosenberg, Jones, and colleagues (Jones & Rosenberg, 1974; Rosenberg & Sedlak, 1972) have used a free-response method to deduce the IPTs or underlying interpersonal dimensions of subjects. Their method requires subjects to provide words or phrases describing each of a number of persons familiar to them. Thus, both the choice of target persons and the selection of descriptors is left to the subject. Subsequently, the number of times each subject was rated on the same target person in the protocol. Results of multidimensional scaling analyses on similarity data generated from the frequencies of descriptor cooccurrences in individual subjects' responses suggest substantial individual differences in subjects' interpersonal space (Rosenberg, 1977).

Implications for Performance Rating

Recall that a distinction was made previously regarding personal construct systems and IPT. IPT is typically concerned with the way in which dimensions covary in raters' beliefs about behavior and personality. Personal constructs, on the other hand, have to do with both the relationships between interpersonal dimensions and the content of those dimensions.

Importantly, individual differences in dimension content appear to have additional implications for performance rating. Content in personal constructs related to viewing rates' work behavior may to a degree affect what the rater looks for in observing persons at work. If one rater has as an important construct "getting along smoothly with others on the job," and a second rater does not share that construct or anything like it, the first rater may be more likely than the second to focus on work behavior related directly to that aspect of performance.

The relationships of personal constructs to perceptions of work behavior may be akin to what might be called "folk" theories of work performance. In my experience, interviews with persons about work on jobs or even casual conversation with people about their jobs sometimes reveal what appear to be deeply felt and rather idiosyncratic "theories" of job performance. Consider these statements: A sales manager says with conviction, "You know what the key to this (sales) job is? Thinking on your feet with customers." A first-line supervisor speaks, "Show me a person who comes to work on time and I'll show you a good employee," and a manager states, "You gotta be political on this (manager's) job; that's the way to make it." Folk theories like these held by persons responsible for performance appraisal may reflect individual's category systems that help shape judgments about how effectively employees are performing, in part because of what raters attend to in gathering performance information. If we were to use the Rep Grid to elicit personal constructs related to job performance, we might in effect be generating these kinds of folk theories.

Of course, characteristics of the work situation and of the employees themselves will in part dictate what raters observe and process when viewing work behavior. When a salesperson makes the biggest sale in the history of the region, the regional manager rater is highly likely to attend to that piece of performance information no matter what the content of his or her personal constructs might be. Also, other features of the situation that increase the salience of a particular construct will make perceivers' use of that construct more likely (Taylor & Fiske, 1978; Tversky, 1977). An example offered by Feldman (1981) is that race is more likely to be a salient construct when a ratee group has only one black person than when it contains all blacks. However, it is also plausible that perceiver differences exist in personal-construct systems related to observing work performance, and these differences may lead to differences in the work behavior attended to and subsequently in evaluations of performance. I submit that these kinds of personal-construct differences are an important potential source of inter-rater disagreement. Research is needed to: (1) determine if raters actually have and can report meaningful personal constructs related to observing work behavior; (2) evaluate the stability of these constructs in assessing work behavior in different situations and contexts; (3) examine individual differences in such constructs; and (4) assess the impact of these differences on observations of work behavior and ratings of work performance.

General Research Proposal for Investigating Performance-Related Personal-Construct Systems

Perhaps the best way to proceed in this research is to elicit constructs from individuals by asking them to list persons with whom they have worked on the target job and by presenting triads of these stimulus persons in the manner of the Kelly Rep Grid. The researcher then asks each subject to think of how two of the stimulus persons are similar in their work performance and yet different from the third for each of a number of triads, until the subject believes that he or she has provided all important constructs. Similar to what has been done in other work with the Rep Grid, "test-retest" reliability and across-situation stability of the constructs might be studied. Previous research involving interpersonal constructs suggests reasonably good stability of these constructs over time (Bieri & Blacker, 1956; Field & Landfield, 1961; Hamilton, 1970; Kehoe & Reynolds, 1977).
Also, dimensionality of individual subjects’ constructs can be examined by asking them to rate each stimulus person on each construct, then factoring or clustering the constructs. This method has been used in research on cognitive complexity, with some subjects’ ratings yielding complex multiple-factor solutions and others more simple one- or two-factor solutions (Adams-Webber, 1979). It may be, for example, that an overall evaluative component overwhelms any differentiation between performance constructs for some or even a majority of subjects. It seems important to study individual differences in the content of constructs, but this is difficult because there is no good way to generate numerical indices of similarities/differences in content. A simple “trait-implication” method suggested by Borman (1978b) allows such numerical analysis.

This method provides a vehicle that allows each individual subject to define his or her own interpersonal categories according to the perceived similarity between each construct and each of a number of traits. A subject “defining” a category or construct rates the similarity in meaning between his or her category and each one of a sizable number of well-defined trait concepts. This vector of ratings then defines the category numerically. For example, consider an interpersonal construct, “affectionate.” Suppose the subject with this category were given the following five traits and rating instructions:

Make ratings on a nine-point scale using the following guidelines:

9. High standing on my category implies high standing on this trait, and low standing on my category implies low standing on this trait; the two are very highly correlated.
5. High or low standing on my category implies nothing about a persons’ standing on this trait; the two are uncorrelated.
1. High standing on my category implies low standing on this trait and low standing on my category implies high standing on this trait; the two are very highly correlated in a negative direction.

Make your ratings on this nine-point scale for each of your categories against each of five traits we supply here.

<table>
<thead>
<tr>
<th>Category</th>
<th>Trait</th>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affectionate</td>
<td>Likability</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Energy Level</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Originality</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Dominance</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>3</td>
</tr>
</tbody>
</table>

Now consider another subject who has a category that he or she labels “Nice to me.” In response to the same trait-implication task the subject might provide the following ratings:

<table>
<thead>
<tr>
<th>Category</th>
<th>Trait</th>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nice to me</td>
<td>Likability</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Energy Level</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Originality</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Dominance</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>5</td>
</tr>
</tbody>
</table>

Notice what we have accomplished with this procedure. We now have numerical definitions of these categories. Further, the subjects themselves have provided the categories and their own definitions of the categories in the context of more widely used trait concepts, consistent with the important principle of free response. This subject-generated numerical depiction of interpersonal categories paves the way for correlational analyses to index similarity in the content of constructs.

For example, the two sets of ratings just shown are highly related ($r = .77$), and this coefficient provides a useful index of the relationship between the content of these two categories. This method might be used, then, in correlational analyses to examine similarities and differences between different subjects’ work-performance constructs. The most obvious hypothesis related to the latter is that persons with similar personal constructs will tend to agree more closely in their performance ratings than persons with very dissimilar construct content.

But other questions must be investigated first to gain understanding of how individuals’ performance constructs work. Do persons with a construct in common tend to “look for” the same kind of behavior in rates? That is, do they search for performance information related to the content of their constructs? More broadly, do raters attend to, process, and remember information in line with their construct systems and fail to see and/or tend to forget rate behavior that falls outside the domain of their category system? And, as mentioned, there are basic questions about the stability of such constructs across time and situations.

Another possibility is that raters simply lack well-defined construct systems related to judging others’ performances. When evaluating rate effectiveness on the job, raters may use many of the same constructs they employ in making general, personality-related evaluations of others. One interesting implication of this speculation is that raters may not think very “naturally” in performance-category terms. Performance-related rating dimensions might be awkward to use because raters are accustomed to making judgments more in the personality domain. Thus, information they have processed about rates related to personality must be “translated” somehow to form a judgment related to performance. Personality research suggesting that we act as naive trait theorists, typically thinking about others in trait terms (e.g., Ross, 1977), lends some credence to this possibility.
The discussion so far of the impact on performance ratings of personal constructs related to work behavior has in a sense suggested a passive examination of these constructs and their impact. A completely different way to approach this is to try to impose upon raters a single, common personal-construct system that corresponds to the dimensional system they should be using to make performance judgments. With this strategy, considerable time and effort should be devoted to getting raters familiar with the rating dimensions and comfortable in using them within the performance-appraisal system. Raters should be made to accept the dimensional system as reflecting a useful and reasonable way to divide up the performance domain.

In addition, raters might be encouraged to organize data about individual rates according to the dimensional system. As performance information is gathered and recorded (as in diary keeping, for example), it should be tied directly to the dimensions. In other words, raters should be strongly encouraged to look for performance data related to dimensions in the system. Over time, this might essentially create a common personal-construct system related to work behavior in each of the raters using the system.

This "active" notion about personal constructs and performance-appraisal dimension systems has an interesting parallel in person-construct theory approaches to dealing with clinical patients. A basic idea in treatment is to work at altering the patient's construct system to make it more useful to the person for dealing with"real-world" interpersonal situations. Here, we are talking about forming a dimensional system maximally useful from the organizations' point of view, and then attempting to alter raters' construct systems to bring them in line with the organizational system.

Summary

In sum, we have briefly reviewed person-construct theory and research, and then we explored possible effects on performance ratings of raters' construct systems related to evaluating ratee work behavior. Individual differences in such construct systems may provide one important reason for interrater disagreement in ratings. In "looking for different things" regarding ratee work performance, raters may in fact observe and process different samplings of behavior, resulting in ratings based on different performance information. Methods and strategies were offered for studying personal constructs in this domain. In addition, I introduced the notion of altering individuals' personal-construct systems to conform to dimensions with which raters should be concerned in evaluating job incumbents.

ACKNOWLEDGMENT

The author thanks Marvin D. Duquette for reading and commenting on an earlier version of this chapter.
6. IMPLICATIONS OF PERSONALITY THEORY AND RESEARCH


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


