THE EFFECT OF AFFECT IN PERFORMANCE APPRAISAL

by

Robert L. Cardy

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

APPROVED:

J.F. Kehoe, Chairman

H.J. Bernardin

N.R. Feimer

J.E. Danes

J.A. Sgro

June, 1982

Blacksburg, Virginia

ACKNOWLEDGMENTS

I would like to express sincere appreciation to my advisor and committee chairman, Dr. Jerard F. Kehoe, for his skillful guidance throughout the course of this project. I would also like to thank Dr. Kehoe for the continual and patient support he has given to my professional training as well as to this particular project. Sincere thanks is also extended to Drs. H. J. Bernardin, N. R. Feimer, J. A. Sgro, and J. A. Danes for the attention they have given to this project.

Special gratitude is extended to my wife, Laurel, whose encouragement and support provided immeasurable assistance during the course of this project. Finally, I would like to thank my parents, Mr. and Mrs. Ralph L. Cardy, for their continual support of my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Performance appraisal is the evaluation of the performance properties of a person, a description of a ratee's job-relevant strengths and weaknesses (Cascio, 1978). Performance appraisal plays an important role in personnel decisions, and is used as both a predictor and a criterion. For example, performance ratings are used when making promotion decisions and in employee counseling. They are used as criteria when evaluating the effectiveness of training programs and selection instruments. In addition, performance appraisals are used to directly link performance and pay, and thus, hopefully, increase motivation. In sum, performance appraisal is an important component in any personnel system.

Actual measurement of performance is typically based on human judgments rather than objective indices (cf., Landy & Farr, 1976). However, subjective performance ratings are subject to conscious and unconscious biases that limit the validity and utility of performance ratings.

The appraisal literature has tended to approach the problem of bias in ratings through formats and rater training programs aimed at reducing rater errors (cf., Bernardin & Pence, 1980). While there has been a considerable amount of research on rating formats, Landy and Farr (1980) have contended that formats account for only 4-8% of the variance in ratings. Based on this estimate and the contention that formats cannot be further improved until we have a better understanding of the rating process, they suggest a moratorium on format-related research.

Rater training programs have often emphasized educating raters about rating errors, in essence teaching raters to not make ratings that "look like that." However, this approach to training has been seriously questioned. Bernardin and Pence (1980) demonstrated that this type of training does indeed decrease certain types of rating errors but does not increase rating accuracy.

Another perspective in the appraisal literature which seems to be gaining in popularity focuses more on the rater and the rating process than on the format (e.g., Cooper, 1981; Feldman, 1981; Jacobs, Kafry, and Zedeck, 1980; Landy & Farr, 1980). The review of performance rating research by Landy and Farr indicates the distinction between the current and previous research trends.

> The major theme in the research that has been conducted in the area of performance rating has been that variables of major importance can be found in the rating scales themselves. Individual differences in raters were only occasionally investigated. Even when these differences were examined, they tended to be second-level or demographic differences, such as sex or experience, rather than first-level direct influences, such as cognitive operations or feelings toward the stimulus object (p. 96).

This emerging stream of research shifts the focus to the cognitive processing involved in performance rating. Examples of this perspective are the models proposed by Cooper, Feldman, and Landy and Farr. Cooper (1981) emphasizes the role of schematic, while Feldman (1981) emphasizes automatic vs. deliberative processing, and Landy and Farr (1980) focus on the rating context as well as rater characteristics.

The present research derives from this recent perspective which emphasizes rater characteristics and cognitive processing in the study of performance appraisal. Specifically, the present research is an investigation of the effect of affect (feeling towards the stimulus object) in performance appraisal. In addition to this major thrust, the performance level of ratees, selective attention ability of raters, and the memory demand imposed by the rating task are also examined. Each of these factors will be briefly reviewed, and the issue of criteria in performance appraisal research will be discussed.

Affect

Affect, or liking, refers to that psychological experience varying from positive (good) to negative (bad). Liking is a self-referent evaluative response to a stimulus, "represented by the prototype 'I like Joe' (Zajonc, 1980, p. 154). In contrast, nonaffective judgments concern judgments of non self-referent stimulus properties and rely on perceptual, cognitive processes rather than affective processes.

Performance appraisal refers to judgments concerning the performance properties of a person and is considered a nonaffective judgment task. In the context of performance appraisal, a rating judgment is represented by the prototype "Joe's lectures and course procedures are well organized." The major concern of the present research is the effect of liking on performance ratings.

Even though Landy and Farr cited feelings (e.g., liking) towards the stimulus object as a potential direct influence on ratings, there has been no research which has directly investigated the effect of liking on performance appraisal. Brown (1968) and Koltuv (1962) investigated rater-ratee familiarity and found a significant negative relationship between familiarity and halo in multi-dimensional trait ratings. However, familiarity and affect are not equivalent and halo in trait ratings may have very little to do with performance rating accuracy. A recent investigation of peer assessment rating methods included a measure of friendship (Love, 1981). It was found that the relationship of the peer assessments with supervisory ratings was not influenced by the degree of friendship among peers. However, this result does not imply that liking doesn't bias ratings, it just may not differentially bias peer and supervisory ratings. That is, since friendship ratings were collected only from subordinates, it may be that perceived social attractiveness of others is comparable between supervisors and subordinates. In addition, the friendship measure employed by Love is a composite of ratings (friendship, knowledge of ratee, time spent with ratee off and on the job, and liking for ratee). Thus, the variable isn't a direct index of liking but is a measure of an overall general construct of, perhaps, ratee social attractiveness.

Clearly then, the effect of affect on rating accuracy has not been directly investigated in performance appraisal research. In addition, none of the recently proposed models of the appraisal process (DeCotiis & Petit, 1978; Feldman, 1981; Landy & Farr, 1980) have included the variable of affect. The lack of attention to affect is characteristic of other areas of study as well. For example, social psychologists are concerned with affect but few of their studies exa-ine affect per se as the major variable (Zajonc, 1980). While the effects of attitude similarity, propinquity, and physical attractiveness on liking have been investigated (cf., Baron & Byrne, 1979), there seems to be little attention given to liking as an independent variable. Rather, attention has been given to variables such as physical attractiveness stereotypes (e.g., Cash, Gillen, & Burns, 1977; Landy & Sigall, 1974).

Only a few researchers have studied the effect of affect on ratings. In probably the most often cited example, Nisbett and Wilson (1977) investigated the effect of a rater's liking for a ratee on ratee attribute ratings. The ratee in the Nisbett and Wilson study was a college instructor in a videotaped interview. The college instructor was depicted as being warm and friendly or cold and distant by manipulating the content of his answers. Introductory psychology students rated the instructor's likability and the attractiveness of his physical appearance, mannerisms, and accent. The manipulation was found to have a significant effect on liking for the teacher. In addition, subjects in the warm condition rated the instructors physical appearance, mannerisms, and accent as signficicantly more appealing than did subjects in the cold condition. Nisbett and Wilson concluded that the affective influence of a global (liking) evaluation alters evaluations of particular attributes, even when the rater has information sufficient to allow for an independent assessment of these attributes.

The Nisbett and Wilson study indicates the potential importance of liking in appraisal. However, it should be emphasized that the appraisal employed in the study involved ratings of physical appearance, mannerisms, and accent on scales ranging from appealing to irritating. Clearly, these evaluative person attribute ratings imply an affective judgment. Thus, the findings of the Nisbett and Wilson study actually indicate that an overall affective judgment of a stimulus affects affective judgments of the dimensions of that stimulus.

As mentioned previously, there is research which has investigated the effects of variables such as attitude similarity, propinquity, and physical attractiveness on rating. For example, physical attractiveness has been found to influence judgments of essay quality (Landy & Sigall, 1974) and judgments of employment suitability (Cash, Gillen, & Burns, 1977). While this research demonstrates that irrelevant information can bias evaluative judgments, this bias is not necessarily due to affect unless it is further assumed that liking mediates physical attractiveness effects, an issue not addressed in this research. In sum, the research on evaluation of others leaves the question of the effect of liking on nonaffective judgments unanswered.

Other laboratory research findings do imply that affective and nonaffective judgment processes are interrelated and influence each other. Research on time estimation appears to support the old adage that 'time flys when you're having fun.' Thayer and Schiff (1975) found that time estimates were significantly longer when in the presence of a scowling face than when in the presence of a smiling face. The literature on perceptual defense (e.g., Erdelyi, 1974) shows a decrement in recognition speed and accuracy for emotionally charged stimuli and thus supports the position that affective and nonaffective judgment processes are interrelated. The perceptual and memory literature reviewed by Matlin and Stang (1978) also indicates that affect influences recall. For example, Nodine and Korn (1968) demonstrated significantly better recall for pleasant than for unpleasant words with delays of only 9-15 seconds. Based on the above types of findings, Matlin and Stang (1978) have proposed the Pollyanna Principle. The Polyanna Principle states that we are positively biased information processors, that cognitive processes slectively favor processing of pleasant over unpleasant information. Specifically, Matlin and Stang propose that the value and intensity of affect can influence nonaffective judgments. However, this conceptualization and supporting evidence is concerned more with the effects of a non-specific, general affective state rather than the effects of stimulus-specific affect on judgments of the stimulus. This latter relationship is the focus of the present research.

A third type of laboratory research focuses directly on liking. Zajonc's research on exposure effects (Moreland & Zajonc, 1977; Wilson & Zajonc, 1980) indicates that liking precedes recognition. The exposure effect refers to the phenomenon of increased liking for a stimulus object due to repeated exposure (Zajonc, 1968). In the Wilson and Zajonc study, subjects were tachistoscopically presented members of a set of octagons for a duration of only 1-msec. Subjects then made paired comparisons between the octagons they had previously seen and new octagons. For each stimulus pair subjects indicated which octagon they liked better and which one they had been shown previously. The recognition judgments were found to be at a chance level but the affective responses reliably discriminated between old and new stimuli, with old stimuli being liked significantly more often than new stimuli. This type of finding has led Zajonc (1980) to propose that affect precedes recognition. That is, one can reliably report liking for a stimulus before the stimulus features are identified. Zajonc further proposes that this early affective reaction is "capable of influencing the ensuing cognitive process to a significant degree" (1980, p. 154).

Laboratory research involving classification tasks supports the proposition that affect influences certain cognitive processes. Kehoe and his associates (Kehoe, Cardy, & Dobbins, note 1; Cardy, Dobbins, & Kehoe, note 2) have investigated affect and the latency of nonaffective judgments. For example, Cardy, Dobbins, and Kehoe investigated the effects of variation in liking, size, and predacity of animal names on latency of liking, size, and predacity judgments. The most pertinent finding of this research is that irrelevant variation in liking significantly interferes with size and predacity judgments. When the subject's task is to classify each animal as either large or small, the classification judgments take significantly longer when liking for the animals varies across trials than when liking remains constant. Although the exact nature of this interference process is unclear, it appears that subjects cannot help but process the affective value of a stimulus even though it is irrelevant to the task. Since it is just this sort of interference effect that is the primary empirical definition of dimensional integrality (Monahan, note 3), liking effects may be explained as though liking were an integral dimension (e.g., Monahan and Lockhead, 1977) of a stimulus and is not easily separated from other dimensions of a stimulus.

Since dimensional integrality is typically studied in relation to judgment latency rather than accuracy, there is no empirical data on the effect of integrality on the accuracy of judgments about stimulus dimension values. However, there is reason to expect that attention to an irrelevant stimulus dimension will degrade accuracy of judgments of the values of relevant stimulus dimensions. As succinctly pointed out by Goldstein and Blackman (1978), "to the extent that individuals are influenced by irrelevant factors in the stimulus field, the accuracy of their judgments should be adversely affected" (p. 224).

In sum, evidence indicates that affect does influence nonaffective judgments. Affect is a primary and early outcome of cognitive processing (Zajonc, 1980). It also seems to be unavoidable and interferes with judgments of stimulus dimension values. Affect also biases judgments such as stimulus size (e.g., Saugstad & Schioldborg, 1966) and stimulus weight (Dukes & Bevan, 1952).

As indicated by Zajonc (1980), a model of affect and its relation to cognition would be premature at the present time. However, some speculations concerning the process mediating the effects of affect do offer themselves. First, affect may produce effects due to its integral nature, attending to it is unavoidable and it is difficult to separate from other stimulus dimensions. Second, affect may produce effects due to its early and primary nature. That is, affect may operate as a schema which guides (i.e. biases) further information processing and recall.

This schema speculation is consistent with the notion that affect is an early outcome of stimulus processing and with the current beliefs concerning limitations of human information processing (cf., Lord, note 4). Human information processing is selective throughout (Erdelyi, 1974). It may be that earlier outcomes of processing, such as affect, are the basis for selectivity in the later phases of processing and in storage.

It should be noted that the integrality and schema speculations are not mutually exclusive. Integrality is conceived as a perceptual phenomenon involving attention and early stages of perception. The schema construct is conceived as being pervasive throughout the entire information processing system and involves cognitive structures. The major difference seems to be in the relative pervasiveness of the two constructs.

The above discussion indicates that liking for a ratee may affect performance ratings of that ratee. More specifically, based on the general proposition that affect towards a stimulus influences judgments of the dimensional values of that stimulus, it is expected that in the context of performance appraisal liking for a ratee will bias ratings of the performance levels of the ratee.

Before proceeding with specific hypotheses, factors which may influence this potential effect of affect will be briefly discussed. It will be proposed that the performance level of the ratee, the rater's selective attention ability, and the memory demands of the rating task may all be important factors in the relationship between affect and rating accuracy.

Performance

Empirical results have confirmed that ratee performance level is, as it should be, the most important determinant of performance ratings. For example, ratee performance was found by Bigoness (1976) to have the largest effect on performance ratings. And, more recently, DeNisi and Stevens

(1981) found ratee performance level to be the most important determinant of overall (single score) performance ratings. However, it has also been found that level of ratee performance is a determinant of rating accuracy (Gordon, 1970, 1972). Gordon found raters to be more accurate when rating high performing ratees than when rating low performing ratees. Interestingly, Landy and Farr's (1980) suggestion that this phenomenon may be due to a bias to seek less information about people we dislike is very similar to Matlin and Stangs' Polyanna Principle that we are positively biased information processors. Furthermore, these results may be accounted for by the schema representation of liking.

If liking operates as a schema, then the consistency of liking and ratee performance should be the important determinant of rating accuracy. Information that is inconsistent with a schema is less likely to be recalled than information that is consistent with the schema (Cohen, 1981; Howard and Rothbart, 1980).[1] Thus, the consistent conditions should yield the most accurate ratings and the inconsistent the most inaccurate ratings.

[1] It should be noted however, that the opposite effect has been found (Hastie, 1981; Hastie & Kumar, 1979). Hastie has concluded that consistent information is likely to be recalled on recognition tasks but inconsistent information is likely to be recalled on free recall tasks. Research by Crocker, Weber, and Binna (note 5) indicates that inconsistent information is typically discounted, thus increasing recall for the inconsistent information but limiting its impact on judgments.

Selective Attention Ability

If liking biases ratings because it is unavoidably attended to and difficult to separate from relevant dimensions as though it were integral, then the selective attention ability of raters may be an important consideration. As previously indicated, judgment accuracy should be adversely affected to the extent that irrelevant stimulus dimensions are attended to. Raters who are high in selective attention ability should be better able to separate liking from other stimulus dimensions. Thus, performance ratings by raters of high selective attention ability should be less biased by liking for ratees than ratings by raters of low selective attention ability.

It appears that field dependence-independence (FDI) measures provide a suitable means of determining the selective attention ability of raters (Witkin, Oltman, Raskin, & Karp, 1971). The FDI construct represents individual differences in the tendency to perceive and process information in an analytic or global fashion. Field dependent persons tend to deal with information in a global fashion and be unaware of its subtle variations while field independent persons tend to deal with information in an analytic fashion and be aware of its subtle variations (cf., Gruenfeld & Arbuthnot, 1969). Goldstein and Blackman (1978) state that "the person who is high in field articulation" (i.e., field independent) "is selective in attention" (p. 9). Thus, a field independent person can separately attend to the dimensions of a multi-dimensional stimulus.

Selective attention ability influences the effect of irrelevant characteristics, such as affect, on judgments, and influences the ratings of others. In an investigation of the perceptual defense phenomenon, it was found that field independent subjects exhibited significantly less perceptual defense than did field dependent subjects (Minard & Mooney, 1969). These results indicate that subjects of high selective attention ability can better separate emotion from perception than can subjects of low selective attention ability. In addition, selective attention ability of raters appears to be a determinate of the degree of differentiation between ratees (Gruenfeld & Arbuthnot, 1969). Gruenfeld and Arbuthnot found FDI to be significantly related to variability in trait ratings of others. Field independent subjects exhibited more variability in their trait ratings than did field dependent subjects. Based on this finding, Gruenfeld and Arbuthnot propose that field dependent raters are less able to distinguish among the traits and performances of ratees than field independent subjects.

Although not directly concerned with accuracy in performance ratings, the above conceptualization indicates the potential importance of rater selective attention ability. Raters of high selective attention ability should be less influenced by the irrelevant dimension of affect and should be better able to differentiate among ratee performance levels than raters of low selective attention ability. Thus, high selective attention ability raters should provide ratings that are more accurate and less biased by affect than ratings provided by raters of low selective attention ability. However, these expectations are based on the notion that liking is an irrelevant dimension that has an integral nature. If liking operates as a Schema, then selective attention ability should not influence the effect of liking.

Memory

Performance ratings often must represent six to twelve months of performance and are apparently provided without the aid of any record of performance over the time period (Landy & Farr, 1980). Ratings made under such an extreme memory demand would be expected to be less accurate than those made not under a memory demand. Although an accuracy measure wasn't available, Bernardin and Walter (1977) demonstrated that keeping a record of ratee performance incidents decreased rater errors.

Why would a memory demand decrease rating accuracy? One possibility is that we simply lose information over time due to decay. However, while decay may occur, current theorizing on human information processing indicates that it is the streamlined and selective nature of our processing and recall that produces the memory effect. As mentioned previously, our entire information processing mechanism seems selective. What appears to determine this selectivity is the schema we bring to bear on the act of perceiving and information storage (e.g., Neisser, 1967). What we attend to, perceive, and store is biased by the schema we employ in the perceptual act (Neisser, 1967; Schiffrin & Schneider, 1977). In addition our recall appears to be biased in the same fashion (e.g., Srull & Wyer, 1979). Once a stimulus is categorized into a schema, subsequent information is biased towards the category and may actually operate on the category rather than the stimulus. In his review of the literature on schema effects on memory, Hastie (1981) concluded that memory for schema relevant information is better than for schema irrelevant information. The schema approach is a cognitively efficient processing mechanism but can lead to systematic biases in perception and recall. This bias is demonstrated by numerous studies finding recognition for false but schema consistent information (cf., Hastie, 1981). In sum, we construct perceptions and recollections to be consistent with our cognitive structures.

The constructive and selective nature of human information processing should lead to increased bias to recall schema consistent information. It may be that information irrelevant to the judgment task forms the basis of a schema (cf., Lord, note 4; Rush, Phillips, & Lord, 1981). Rush et al. examined the performance cue effect (being told the ratee performed well or poorly) on leadership ratings under immediate and delayed rating conditions. Under the demand of memory, there was a marginal increase in the biasing effect of the performance cue. The constructivism perspective and, to some extent, the Rush et al. study indicate increased bias under memory.

Therefore, it is plausible that ratings provided under memory demands may be more biased by affect towards ratees than ratings provided under no memory demand. From the constructivist perspective, this increased bias would be expected if affect were utilized in stimulus categorization. In addition, empirical evidence indicates that bias due to affect increases over time. For example, on the basis of their review of the literature on affect in learning and memory, Matlin and Stang (1978) conclude that recall is increasingly biased in favor of affective over neutral stimuli. The studies reviewed by Matlin and Stang used free recall and recognition for words considered to be pleasant, neutral, and unpleasant. Thus, the results are not necessarily relevant to liking, but do support the general contention of increasing selectivity based on affect.

In addition to the above basis for expecting memory to influence affect bias, the inescapability of affect (Zajonc, 1980) also supports increased bias under memory. As pointed out by Zajonc,

> "We may completely fail to notice a person's hair color or may hardly remember what it was shortly after meeting the person. But we can seldom escape the reaction that the person impressed us as pleasant or unpleasant, agreeable or disagreeable, as someone to whom we were drawn or someone by whom we were repelled. And these affective reactions—and, more important, the retrieval of affect—occur without effort" (p. 156).

This notion of the inescapability of affect is consistent with the constructivist perspective. While details of a ratee's performance may be forgotten, affect toward the ratee will be inescapably recalled, and perhaps, form the basis from which further judgments are made.

In summary, as either a schema component or an integral feature, liking for a ratee may systematically bias performance ratings and this bias may be influenced by the three factors included in this research: the performance level of the rates; the selective attention ability of raters; and, the memory demand imposed by the rating task.

Appraisal Criteria

There are numerous criteria by which performance ratings are evaluated. The various criteria and their conceptual and operational definitions have been extensively discussed elsewhere (e.g., Saal, Downey, & Lahev. 1980). In sum, most of the criteria are various psychometric characteristics of ratings such as halo, leniency, and central tendency. These characteristics are assumed to be the result of rating inaccuracies or biases and are thus termed constant or rater errors. As pointed out by Murphy and Balzer (note 6), these rater error measures are prescriptive in that ratings are presumed to be inaccurate unless they meet assumptions about the true distributions and intercorrelations among performance measures. Thus, it is presumed that less haloed or lenient ratings are more accurate than more haloed or lenient ratings. Since objective measures of performance are typically unavailable, these rater errors have been used as indirect but surrogate indices of accuracy. However, there has recently been progress in developing a more direct and empirical, rather than prescriptive, criterion for assessing the quality of ratings.

Borman (1977, 1978, 1979) has developed and applied "true scores" for directly assessing rating accuracy. Briefly, Borman's methodology for developing "true scores" involves the following. Scripts for ratees are written which include effectiveness levels or "intended true scores" for each performance dimension. Videotapes of actors following these scripts are then rated by expert raters and the mean expert ratings on the dimensions are taken as the "true scores."

Another method of generating "true scores" has been developed by researchers employing written vignettes of ratees (cf., Bernardin & Pence, 1980). In this approach, a sample of raters rate the effectiveness level of each performance incident. The method involves the retranslation procedure used in BARS development. The retranslation procedure involves three groups of raters. The first group identifies the dimensions of performance. A second group generates critical incidents for each of these performance dimensions. A third group of raters is then given the dimensions and a randomized list of the incidents. These raters sort the incidents into the dimensions which they best represent. A criterion of at least 60% agreement concerning what dimension each incident belongs to is used to select incidents. Any incident over which there was less than the criterion level of agreement is eliminated. After this retranslation phase, a fourth group of raters is given the dimensions and surviving incidents and rate the effectiveness level of each performance incident in the context of the appropriate dimension. Means and standard deviations of these effectiveness ratings are then computed for each incident. If a BARS was being developed, at this point in the method incidents would be selected based on the means and standard deviations to define the points on each dimension rating scale. However, the incidents can also be used to develop a vignette of a ratee. Rather than selecting incidents to define scale values, incidents can be selected to define the performance level of a fictitious ratee across the performance dimensions. Thus, the incidents are used to define the true level of performance of a ratee as well as to define the scale values of rating dimensions. For example, three incidents relevant to each dimension might be selected to form a ratee vignette. These incidents would be presented in a random order in a vignette. The "true scores" on each dimension would be the average of the mean effectiveness ratings for the three incidents representing each dimension. These scores then serve as a criterion against which ratings of the fictitious ratee can be evaluated.

The "true score" approach to assessing rating effectiveness is not beyond dispute. But, when such true scores are available, they are preferable to surrogate accuracy measures such as halo and leniency which may, in fact, reflect reality rather than rater biases (e.g., Schwab, Heneman, & DeCotiis, 1975). Thus, the use of the constant error criteria should, at best, be reserved for situations where "true scores" are unavailable, unless the object of investigation is the relationship of these criteria to accuracy.

There are a variety of accuracy measures (cf., Cronbach, 1955) which could be used to assess performance ratings, just as there are a variety of measures for each of the constant rater errors. As Cronbach (1955) has pointed out, the typical difference score measure of accuracy contains the accuracy components of elevation (overall rating level), differential elevation (discrimination between ratees), stereotype accuracy (discrimination between performance dimensions), and differential accuracy (DA) (discrimination between ratees within dimensions). Of the various accuracy measures, Borman (1977) has argued that DA is the most theoretically and practically meaningful measure for assessing performance ratings. The DA measure indicates the ability of a rater to correctly discriminate between ratees in terms of rank ordering and relative performance effectiveness differences among ratees. Operationally, DA is the average of Z' transformed correlations between ratings and true scores across ratees and within dimensions. DA is high when raters provide ratings which rank order ratees and indicate performance effectiveness differences among ratees similar to the rank ordering and performance differences indicated by the true scores. When the rank orderings and relative performance differences indicated by ratings do not correspond to the rank orderings and relative performance differences indicated by the true scores, DA is low. DA indicates the ability to accurately discriminate among ratees.

While there is a measure of rating accuracy which seems conceptually most appropriate, there are no such clearly preferred measures for assessing the constant rater errors. As indicated by Downey and Saal (note 6), there is some disagreement concerning conceptual definitions for several constant error criteria, and even more disagreement concerning the operational definitions of those criteria. The review by Saal et.al. (1980) indicates at least four operational definitions for halo, three for leniency or severity, and four for central tendency and restriction of range. Research by Murphy and Balzer (note 6) indicates that the alternate measures of halo, leniency, and range restriction are not highly correlated within a criterion.

In addition to the disagreement among alternate rater error measures, there appears to be little relationship between the rater error measures and accuracy. Cooper (1981) has noted that the correlation between halo and accuracy tends to be weak and positive. If halo is indeed a surrogate measure of accuracy, then there should be a strong negative correlation between the measures, the more halo the less accuracy. Borman (1977) found the correlation between halo and accuracy to be between .12 and .18. Research by Murphy and Balzer (note 6) directly investigated the relationship between the rater errors of halo, leniency, and central tendency with rating accuracy across three studies. None of the correlations between accuracy and error measures were significant across all three studies. Only two error and accuracy correlations were significant across two studies, and both were the paradoxical positive relationship between halo and accuracy. However, the paradox actually seems to be a statistical artifact, raters whose halo was closest to the true score estimates of halo were most accurate. Murphy and Balzer conclude that rater error measures are poor indicators of rating accuracy. Further, they contend that there is

> no data which would lead one to state with any confidence that a particular error measure was in fact related to any known measure of accuracy. The burden of proof is clearly upon those who wish to use rater error measures as indirect indicators of accuracy (p. 12).

In sum, there are clear reasons for preferring a direct accuracy measure over rater errors when assessing quality of ratings. First, estimates of rating accuracy are based on direct and empirically derived rather than indirect and prescriptive criteria. Second, there is an operational definition of accuracy which is the most conceptually appropriate, particularly when the concern is with discrimination ability. In contrast, there are a variety of operational definitions of rater errors which exhibit little covariance and no clear reason for preference among them. Third, there is a questionable relationship between the rater errors and accuracy. In situations where true score approximations are available, a direct measure of accuracy should be used. To find an effect on a rater error has, at the present time, dubious value. For example, unless the focus of study is on the relationship of accuracy and rater errors, to find a significant effect on halo, but not a corresponding effect on DA, has little value. DA is the most appropriate criterion for which the indirect rater error measures are not surrogates.

Statement of the Problem

The effect of liking for a stimulus on judgments about attributes of that stimulus has received minimal attention. However, empirical findings and the recent theoretical speculation offered by Zajonc (1980) have indicated the potential importance of affect in cognitive judgments. The present research was an attempt to determine the effect of liking for ratees on the DA of performance appraisal judgments and to identify plausible explanations of this effect, with special consideration given to implications these explanations have for rater training and format development.

Three separate studies were done. The first and second focused primarily on determining what effects affect has on DA and what other variables moderate these effects. The third study was an attempt to examine the plausibility of three possible explanations for the effects of affect observed in the first and second studies.

Study 1

Overview

In an attempt not only to demonstrate but also to explicate the nature of the effect of affect, study 1 included all of the factors previously discussed. Liking for ratees, performance level of ratees, selective attention ability of raters, and the memory demand imposed by the rating task were combined in an orthogonal manner. Equal numbers of male and female subjects served as rater-subjects in the study. Each subject was presented with vignettes of four instructors that included trait descriptors of the instructor-ratees which were utilized to manipulate a rater's liking for the ratees. Six sets of four instructor vignettes were used, each set representing one of the six combinations of instructor likableness (liked, neutral, disliked) and performance (high, low).

The selective attention ability of subjects was measured in a pretest with the Hidden Figures test (Jackson, Messick, & Myers, 1964). Only those subjects scoring in approximately the upper and lower thirds of the distribution of scores on the pretest participated in the rating study. The memory demand of the rating task was manipulated by either allowing subjects to refer back to the vignettes while making ratings or requiring subjects to make ratings without referring back to the vignettes. As a liking manipulation check, subjects were asked to rate their liking for the set of instructors they had rated.

Study 1 Hypotheses

Liking and Performance Ratings

Although research reviewed by Zajonc (1980) and the interference finding by Kehoe and associates suggest that liking may influence performance ratings, liking, or "feelings toward the stimulus object" (Landy & Farr, 1980), has not yet been directly investigated in performance appraisal research or included in models of the appraisal process. However, consideration of relevant research does allow the statement of a number of exploratory hypotheses.

The research discussed by Zajonc indicates that affect precedes cognitive evaluations of a stimulus. This early nature of affect suggests its possible role in schematic processing. That is, an early evaluation may serve as a basis for further processing of the stimulus. In addition, research by Kehoe and associates suggests an integral nature of liking. If either of these propositions is correct, then the salience of liking for ratees should decrease the DA of the ratings of ratee performance properties. If liking is part of the rater's schema, the decrease in DA would be due to the affect based selectivity in the perceptual process. If liking is an integral feature of the ratee, the decrease in DA would be due to the difficulty of separating affect from the performance properties of ratees. In either case, liked and disliked ratees should be discriminated less accurately than affectively neutral ratees. This analysis leads to the first hypotheses.

> Hypothesis 1: Liked and disliked ratees should be discriminated less accurately than affectively neutral ratees.

Ratee Performance and Performance Appraisal:

The most important determinant of performance ratings should be ratee performance level. Consistent with this, the investigations by Bigoness (1976) and Denisi and Stevens (1981) have found ratee performance to be the most important influence on performance ratings. However, research by Gordon (1970, 1972) indicated that high level performers are rated more accurately than low-level performers.

The effect of liking and the accuracy difference between ratings of high and low performers may be due to schematic processing. In particular, low performers may be rated less accurately than high performers because the low performance schema isn't as developed and differentiated as the high performance schema, perhaps due to less experience with persons perceived to be low performers. In addition, liking may operate as, or be a basic component of, a schema. If so, the consistency of performance with liking should be a determinant of rating accuracy. The inconsistent or incongruent situations of likable low performers and dislikable high performers should be rated least accurately, while the consistent situations should be rated most accurately.

Within the integrality conceptualization of affect, rating accuracy is not determined by the consistency of affect and performance but rather by the inseparability of affect from other ratee features which thus obscures differentiation among ratees. If affect operates as an integral dimension, then it should obscure the differentiation of stimulus attributes and the consistency of the dimensional values of those attributes with affect should not influence the effect. Thus, positive or negative affect should obscure the differentiation among ratees regardless of their overall performance level.

Based on this rationale, the following hypotheses were proposed.

> Hypothesis 2(a): Low performers will be rated less accurately than high performers.
>
> Hypothesis 2(b): If the integrality conceptualization of affect is correct, then the effect of liking on rating accuracy should not be dependent upon ratee performance level.
>
> Hypothesis 2(c): If the schema conceptualization of affect is correct, then there should be a liking by performance interaction with consistent liking and performance situations (i.e. likable high performers and dislikable low performers) being rated more accurately than inconsistent situations.

Rater Selective Attention Ability and Performance Appraisal

As previously indicated, selective attention ability is reflected in the level of FDI. High selective attention ability people deal with information in an analytic fashion while low selective attention ability people deal with information in a global fashion. Since high selective attention ability raters are less influenced by irrelevant information, judgments concerning an attribute of a stimulus should be less biased by the values of other attributes of the stimulus for high selective attention ability raters than for low selective attention ability raters.

Specifically, high selective attention raters should be better able to separate the performance of ratees from their liking for ratees. The perceptual defense findings of Minard and Mooney (1969) support the contention that high selective attention ability raters are better able to separate affect from nonaffective processes. In affect decreases rating accuracy due to its integral nature, the high selective attention ability raters should be better able to separate this irrelevant dimension from their judgments of ratee performance and provide more accurate ratings than raters of low selective attention ability. When there is neutral affect for ratees, ratings provided by high and low selective attention ability raters should be of comparable accuracy.

In contrast, if liking influences nonaffective judgments due to a schematic property, then the selective attention of raters should not influence the effect. From the schema perspective, liking is a basis for selectivity of information processing. In this case, affect is not an irrelevant dimension but a basis on which further processing is directed.

Based on this rationale, the following hypotheses were proposed:

> Hypothesis 3(a): Raters of high selective attention ability should provide more accurate ratings than raters of low selective attention ability.
>
> Hypothesis 3(b): If liking operates as an integral dimension, then ratees of high selective attention ability should provide more accurate ratings of liked

and disliked ratees than raters of low selective attention ability.

Hypothesis 3(c): If liking operates as a schema, then rater selective attention ability should not influence the accuracy of ratings of liked and disliked ratees.

Memory and Performance Appraisal

As far back as Ebbinghaus it has been known that the amount of information retained declines over time (cf., Woodworth & Schlosberg, 1954). In the domain of performance appraisal, it has been shown that keeping a physical record of ratee performance incidents results in ratings that are psychometrically superior to ratings based on the retention of ratee performance in memory. While there have been no performance appraisal studies on memory and rating accuracy, the decay or loss of information from memory should produce ratings of ratee performance attributes that are less accurate when ratings are based on recall than when no recall is required.

Aside from a simple loss of information over time, processing and recall appears to be directed toward the schema employed in the processing, storage and recall of information (e.g., Neisser, 1967; Shiffrin & Schneider, 1977; Srull & Wyer, 1979). As discussed previously, affect may be utilized in the schematic and constructive processes. In addition, it has been argued that affect is an inescapable accompaniment of stimulus recollection (Zajonc, 1980). If affect is utilized in schematic processing and is an inescapable accompaniment of stimulus recollection, then memory may influence the degree of inaccuracy introduced by affect. The schema conceptualization of affect indicates that liking for ratees should systematically influence the storage of ratee information towards the appropriate affect-laden categories. Thus, recall of ratee information should be biased by the affect of the category used for storage and by the inescapability of affect when ratee information is recalled. When no recall is required, the additional biases introduced by affect in information storage and recall should not be in operation. Ratings in liked and disliked conditions based on recall of ratee performance should be less accurate than ratings made without a recall demand due to the inescapability of affect and its role in schematic processing.

If affect operates as though it is an integral dimension, then the effect of affect on rating accuracy should not be greater when the rating task imposes a memory demand than when it does not impose a memory demand. From the integrality perspective, affect influences nonaffective judgments because it is a dimension that is difficult to separate from other stimulus dimensions. However, the integral nature of a dimension does not imply its utilization as a basis for storage and recall. The degree to which affect obscures ratee differentiation should not be increased when ratings are based on recalled rather than present ratee information. There is no theoretical consideration or empirical evidence which indicates that degree of integrality is influenced by memory.

Based on this rationale, the following hypotheses were proposed.

Hypothesis 4(a): Performance ratings will be less accurate when based on recall of ratee information than when no recall is required.

Hypothesis 4(b): If affect is a basis for schematic processing and is inescapable, then the accuracy of rating liked and disliked ratees relative to neutral ratees will be less when ratings are based on recall than when no recall is required.

Hypothesis 4(c): If affect acts as an integral dimension, then memory demand will not affect the inaccuracy introduced by liking.

Method

Subjects

The subjects in the present study were 144 male and 144 female introductory psychology students. All subjects received extra credit for participation in the experiment. The subjects were selected for participation in the study on the basis of pretested selective attention scores. A total of 345 males and 239 females were administered the hidden figures test (HFT) in a series of pretests. The HFT is a group administered measure of FDI (Jackson et al., 1964). On the basis of preliminary data from 195 subjects, the test score values of 8 and 14 were selected as lower and upper cutoff points. Thus, only those subjects scoring in the upper and lower thirds of the distribution of HFT scores were offered participation in the present study. Subjects meeting the pretest criteria were contacted by phone and offered the opportunity to participate in the present study for an experimental credit. The average HFT scores for all males and females who participated in the pretest sessions were 12.6 and 11.1, respectively. Subjects scoring in the lower portion of the distribution were considered low selective attention raters while those scoring in the upper portion of the distribution were considered high selective attention raters. The average HFT scores for low and high selective attention males who participated in study 1 were 5.6 and 19.5, respectively. The average HFT scores for low and high selective attention females who participated in study 1 were 5.7 and 17.5 respectively.

Experimental Design

The experimental design was a 3x2x2x2 (liking by ratee performance by rater selective attention ability by memory demand of rating task) between subjects factorial. Twelve subjects (six male and six female) were randomly assigned to each of the 24 treatment combinations.

Instructor Vignettes

Vignettes described the eight hypothetical instructor ratees used in the study. Each vignette consisted of 15 critical incidents that described the instructor's classroom behavior. The classroom behavior incidents were selected from the list of incidents developed by Sauser, Evans, and Champion (note 7). Sauser et al. reported incidents on five dimensions of teacher behavior. The incidents had survived a typical BARS retranslation procedure. The ratee vignettes were formed by selecting three incidents from each of the five dimensions.

The true score for each dimension for each ratee was defined as the mean effectiveness ratings of the three incidents within each dimension, as provided by Sauser et.al.

Ratee performance level. The performance level of ratees was either high or low. A high performing ratee was defined as a vignette with three dimension true scores more than one point above scale midpoint, one dimension true score within one point of scale midpoint, and one dimension true score lower than one point below scale midpoint. A low performing ratee was defined as a vignette with three dimension true scores more than one point below scale midpoint, one dimension true score within one point of scale midpoint, and one dimension true score higher than one point above scale midpoint. Table 1 presents the dimension true scores for all ratee vignettes.

Appendix A contains the list of incidents used in the instructor vignettes. The median effectiveness rating of each incident, the semi-interquartile range of ratings of each incident, and the percentage of subjects giving the modal rating for each incident found in the Sauser et al., study are also presented in Appendix A.

Rating Scales

The rating scales were 11 point scales with behavioral incidents defining the ends and middle of each scale. Each rating form contained five rating scales, one for each of the five performance dimensions. Incidents provided by Sauser

Table 1

Dimension True Scores for Each Ratee

	Low Performers				
	Dimension*				
Ratee	A	B	C	D	E
Morgan	5.36	1.83	9.53	3.06	1.60
Pierce	1.93	9.16	6.10	4.30	2.66
Witkin	9.63	1.73	1.23	5.83	3.83
Ritter	1.53	6.36	3.03	7.90	3.33
Standard Deviation	3.26	3.15	3.16	1.81	0.84

	High Performers				
	Dimension*				
Ratee	A	B	C	D	E
Morgan	5.36	10.10	1.26	7.80	10.46
Pierce	10.03	2.13	6.10	8.30	10.00
Witkin	3.80	8.26	10.43	5.83	8.80
Ritter	8.50	6.36	10.13	2.56	9.96
Standard Deviation	2.47	2.96	3.72	2.26	0.61

*Dimension	Label
A	Relationships with Students
B	Ability to Present the Material
C	Interest in Course and Material
D	Reasonableness of the Workload
E	Fairness of Testing and Grading

et al. with median rating values closest to the value at the middle and ends of the rating scales were used to behaviorally define those scale values. Incidents used to form the rating scales were not included in any of the instructor vignettes. A sample rating form is presented in Appendix B.

Liking manipulation. Liking for ratees was manipulated by describing ratees with likable, dislikable, or neutral trait terms. The terms used were taken from Anderson (1968) who reported likableness ratings for 555 trait terms. Subjects in the Anderson study rated each trait term according to how much they would like a person with that characteristic. However, in addition to engendering differing levels of liking, trait terms may have direct implications for ratee performance on the five rating dimensions.

In order to control for the possible confounding of implications with liking, a total of 300 of Anderson's trait terms were pretested for their implications for performance the five rating dimensions. Sixty subjects rated how sure they were that behavior characteristic of each of the five dimensions would be exhibited by a college professor described with one of the traits. Each subject made these ratings for fifty traits, first on dimension 1, then on dimension 2, and so on. The name of the dimension and behavioral examples indicating low, medium, and high levels of performance on that dimension were presented with the list of 50 traits. For each dimension the list of 50 traits was repeated in a random order. A total of 10 subjects rated the degree of implication for each trait for performance on the five dimensions.

The object of this pretest was to identify trait terms which manipulate liking but do not have implications for performance. If the mean implication rating of a trait was less than the "somewhat-sure" category for <u>each</u> of the five dimensions, that trait term was accepted as a possible term for the liking manipulation. Any effect in study 1 due to the traits selected in this fashion should be due to the different levels of liking they engender, not due to differing implications for performance dimensions.

The directions for the implications pretest and the materials for a set of 50 traits are presented in Appendix C. Appendix D presents the 300 trait terms and the mean implication rating for performance on each of the five performance dimensions.

Of the 68 terms that satisfied the criterion, 13 were likable, 28 were dislikable, and 27 were neutral terms according to Anderson's (1968) liking ratings. These 68 terms were pretested for differences between levels of personal and popular liking they engender. This pretest was carried out to determine whether there were differences in how much an individual would like a person described with a particular trait versus how much most people would like a person described with that trait. It is possible that most people would dislike someone described as "obnoxious" but that some individuals would like a person so described. In order to determine whether there is such an interaction between personal and normative judgments in liking engendered by the traits, a total of 36 subjects were asked to rate the likableness of each of the 68 terms. One group of 18 subjects was asked to rate each term according to how much they, personally, would like a person being described with that term. Another group of 18 subjects was asked to rate each term according to how much most people would like a person being described with that term. A sample of the pretest materials are presented in Appendix E.

A main effect for the personal versus normative orientation manipulation would simply indicate a deflection in liking ratings. For example, personal ratings may be more extreme than popular ratings, but the direction of ratings be consistent in the two situations. A main effect for liking would indicate the effectiveness of the liking manipulation. However, a liking by orientation interaction would indicate that liking is influenced by differences between personal preference and norm preference as well as by trait differences. Such an interaction would indicate that an idiographic rather than nomothetic approach should be taken in the liking manipulation.

The means and standard deviations of ratings given to each of the 68 terms are presented in Appendix F. Means of the ratings for words considered likable, dislikable, and neutral were calculated for each subject and entered as data points in an ANOVA. It was found that the personal and norm conditions did not significantly influence ratings [$F(1,34) = .64, P > .05$] while the likableness of words had a significant main effect

[$F(2,68) = 6.94, p < .01$]. The orientation by liking interaction was not significant [$F(2,68) = 1.4, p > .05$].

On the basis of the above pretests, a total of 40 trait terms were selected for the liking manipulation. The terms selected are presented in Table 2. There are a total of ten likable terms, twelve dislikable terms, and eighteen neutral terms.

A set of six traits was included in each instructor description. Likable instructor descriptions included four liked traits and two neutral traits while dislikable instructors included four disliked traits and two neutral traits. Neutral instructor descriptions included six neutral traits.

Procedure

Subjects were run in groups of approximately twenty individuals. On the basis of HFT scores, selected subjects were contacted by phone and offered credit for participation in the study. Each subject who agreed to participate in the study was randomly assigned to an experimental condition and the subject's name was written on the appropriate set of materials. Subjects first were given directions for the rating task. There were two sets of instructions, one for the memory condition and another set for the no memory condition. The only difference between the two sets of instructions was that in the memory condition subjects were told that they could not look back at the instructor descriptions when rating while subjects in the no memory condition were encouraged to

Table 2

Liking Manipulation Terms*

Like	Dislike	Neutral
courageous	underhanded	impressionable
tactful	untrustworthy	lonely
loyal	self-conceited	solemn
wholesome	boastful	daydreamer
well-mannered	self-centered	lonesome
cheerful	superficial	aggressive
sharpwitted	greedy	bashful
gentle	shallow	unlucky
amusing	liar	restless
clean	dishonest	cunning
	childish	shy
	bragging	daredevil
		quiet
		self-righteous
		ordinary
		emotional
		innocent
		suave

*The average of the implication values across the five rating dimensions for the likable, dislikable, and neutral sets of terms are 2.72, 2.73, and 2.76, respectively.

do so. Memory and no memory condition subjects were always run in separate groups.

All subjects were told that the experiment was concerned with instructor ratings and that their task in this experiment was to rate, as accuratley as possible, the performance of the four instructors described in the vignettes. The rating scales were then described and shown to the subjects. It was explained that behavioral examples of instructor performance, rather than adjectives, were used to indicate the meaning of the numbers on each rating dimension. Subjects were told to familiarize themselves with the rating form prior to reading the instructor descriptions. All subjects were told to read all four instructor descriptions before rating. After completion of the rating task, each subject was given a questionnaire in which an overall rating for the level of liking for the set of four instructors was requested.

Each subject rated four instructors who were all at the same level of liking, liked, disliked, or neutral, and the same level of performance, high or low. The order of the instructor vignettes was determined by a Latin square. Thus, within a condition, each of the four male (or female) subjects received a different order; for each of the remaining two male (or female) subjects in each condition, one of the four orders was assigned to one subject and the inverse of that order was given to the other subject.

Results and Discussion

A DA score was calculated for each rater according to the following procedure. The correlation between ratings and true scores across ratees and within each of the five dimensions was calculated for each rater. Each of these correlations was then transformed to a Z' score and the five Z' scores for each rater were averaged to obtain each rater's DA score. Each DA score reflects the extent to which a subject's ratings differentiated between ratees in the same way that the true scores differentiate between ratees in terms of rank ordering and distance between ratees.

After presenting the results for the order factor and liking manipulation check, the results will be presented in the order in which the hypotheses were presented.

Order of Instructor Vignettes

The order of instructor vignettes for each subject was determined with a Latin square. The Latin square resulted in four unique orders and two redundant but inverted orders for each set of six same-sex subjects within a treatment combination. In order to assess the effect of order, those subjects receiving the redundant orders were not included in the analysis. A one-way analysis of variance (ANOVA) on the four unique orders revealed no significant order effect on DA [$\underline{F}$ (3,188) = .79,$\underline{p}$>.05]. The order factor will be collapsed over in all following analyses.

Liking Manipulation Check

After completion of the rating task each subject was asked to provide a summary rating of liking for the set of four instructors he/she had rated. This rating was entered as a dependent measure in 3x2x2x2x2 (liking x performance x memory x selective attention x rater sex) ANOVA. The results revealed main effects for liking [$\underline{F}$ (2,230) = 4.54, $\underline{p}<.01$], performance [$\underline{F}$ (1,230) = 323.46, $\underline{p}<.01$], and memory [$\underline{F}$ (1,230) = 7.38, $\underline{p}<.01$) on the liking ratings. In addition, the two-way interactions of rater selective attention ability by rater sex [$\underline{F}$ (1,230) = 7.22, $\underline{p}<.01$] and memory by rater sex [$\underline{F}$ (1,230) = 4.12, $\underline{p}<.05$] were both significant sources of variance in the liking ratings. While the liking manipulation significantly influenced self-report of liking, other sources significantly influenced liking ratings. In sum, the liking manipulation was effective, but not the only source of influence on liking. The mean liking ratings for like, disliked, and neutral ratee sets were 3.64, 3.30, and 3.67, respectively.

Liking

Hypothesis 1 predicted that ratings of liked and disliked ratees would be less accurate than ratings of affectively neutral ratees. A 3x2x2x2x2 (liking x performance x memory x selective attention x rater sex) ANOVA on DA revealed the liking manipulation to not be a significant source of variance ($\underline{F}<1.00$). A summary table of the overall analysis is presented in Table 3. The DA means for like, disliked, and neutral

Table 3

Summary of Analysis of Variance on Mean DA

Source of Variation	Sum of Squares	df	Mean Square	F	$\hat{\omega}^2$
LIK (Liking)	0.05	2	0.025	0.12	
SA (Selective Attention)	1.96	1	1.960	9.48*	.03
LIK x SA	0.71	2	0.360	1.71	
MEM (Memory)	23.12	1	23.120	111.64*	.28
LIK x MEM	0.11	2	0.055	0.27	
SA x MEM	0.01	1	0.010	0.03	
LIK x SA x MEM	0.72	2	0.360	1.74	
PERF (Performance)	5.18	1	5.180	25.01*	.08
LIK x PERF	0.24	2	0.120	0.57	
SA x PERF	0.50	1	0.500	2.44	
LIK x SA x PERF	0.34	2	0.170	0.83	
MEM x PERF	0.01	1	0.010	0.04	
LIK x MEM x PERF	0.77	2	0.385	1.85	
SA x MEM x PERF	0.04	1	0.040	0.19	
LIK x SA x MEM x PERF	0.12	2	0.060	0.30	
SEX (Sex of Rater)	0.47	1	0.470	2.26	
LIK x SEX	0.27	2	0.135	0.66	
SA x SEX	0.02	1	0.020	0.11	
LIK x SA x SEX	0.87	2	0.435	2.09	
MEM x SEX	0.07	1	0.070	0.33	
LIK x MEM x SEX	0.19	2	0.095	0.47	
SA x MEM x SEX	0.48	1	0.480	2.31	
LIK x SA x MEM x SEX	0.06	2	0.030	0.14	
PERF x SEX	0.42	1	0.420	2.04	
LIK x PERF x SEX	1.78	2	0.890	4.31*	.05
SA x PERF x SEX	0.003	1	0.003	0.02	
LIK x SA x PERF x SEX	0.57	2	0.285	1.38	
MEM x PERF x SEX	0.01	1	0.010	0.06	
LIK x MEM x PERF x SEX	0.33	2	0.165	0.80	
SA x MEM x PERF x SEX	0.08	1	0.080	0.39	
LIK x SA x MEM x PERF x SEX	0.17	2	0.085	0.42	
Error	49.699	240	0.207		
Total	89.38	287			

*p<.05

ratees were .92, .91, and .93, respectively. The planned comparisons of the neutral condition with the average of the liked and disliked conditions revealed no significant difference (_p_ >.05).

Performance

Hypothesis 2(a) predicted that low performers would be rated less accurately than high performers. The overall analysis revealed that DA was significantly lower for ratings of low performers than for ratings of high performers [_F_(1,240) = 25.01, _p_ <.01). The DA means for low and high performers were .79 and 1.05, respectively.

Hypothesis 2(b) predicted no interaction between liking and performance while hypothesis 2(c) predicted such an interaction. The overall ANOVA revealed that the liking by performance interaction was not significant (F<1.00). Hypothesis 2(c) was further tested by comparing the average of DA in the consistent conditions (likable high and dislikable low performance conditions) with the average of DA in the inconsistent conditions. This comparison revealed the difference in DA to not be significant [_t_(240) = .04, _p_>.05].

The interpretation of the above results is qualified by the significant liking x performance x rater sex interaction [F(2,240) = 4.31, _p_<.05]. The triple interaction is graphically presented in figure 1. Simple effects analyses were performed on the interaction by testing the two-way interactions separately for male and female raters. The analyses revealed that the liking by performance interaction did not significantly affect the DA of ratings made by males [*F*(2,240 = 2.43, *p*>.05] or females [F(2,240) = 2.44,*p*>.05]. The liking main effect was not significant for male or female raters (*p*>.05), while the main effect for ratee p4rformance significantly affected the DA of ratings made by males [*F*(1,240) = 6.38,*p*<.05] and females [*F*(1,240) = 20.66,*p*<.05]. The analyses indicate that liking and performance do not have a joint influence on the DA of ratings made by either males or females.

Simple effects analyses were also performed on the triple interaction by testing the two-way interactions at each level of performance. The analyses revealed that the liking by rater sex interaction significantly affected the DA for low performers [*F*(2,240) = 3.62,*p*<.05] but not the DA for high performers [*F*(2,240) = 1.34, *p*>.05]. The liking main effect was not a significant source of variance in the accuracy of rating low performers (*F*<1.00) or high performers (*F*<1.00). The main effect for rater sex was not significant for low performers (*F*<1.00) but did significantly affect the DA for high performers [*F*(1,240) = 4.29, *p*<.05]. The significant liking by rater sex interaction for ratings of low performers indicates that females are more accurate than males in the inconsistent situation of likable low performing ratees but less accurate than males in the consistent situation of dislikable low performing ratees. While the liking by sex interaction was insignificant for high performance ratings, the direction of the mean difference in rating accuracy between males and females rating dislikable high performers again indicates the accuracy superiority of females in inconsistent rating situations.

Selective Attention

Hypothesis 3(a) predicted that ratings by high selective attention raters would be more accurate than ratings by low selective attention raters. This hypothesis was confirmed. The ANOVA revealed selective attention to be a significant source of variance [$\underline{F}(1,240) = 9.48$, $\underline{p}<.01$], with DA means for high and low selective attention raters of 1.00 and .84, respectively.

Hypothesis 3(b) predicted that raters of high selective attention would rate liked and disliked ratees more accurately than low selective attention raters. Hypothesis 3(c) predicted no difference in the accuracy of rating liked and disliked ratees between high and low selective attention raters. The ANOVA revealed no significant liking by selective attention interaction [$\underline{F}$,$(2,240) = 1.71$, $\underline{p}>.05$]. A planned comparison test revealed no significant accuracy difference in ratings of liked ratees between high and low selective attention raters [$\underline{t}(240) = 1.12$, $\underline{p}>.05$]. However, the comparison test did reveal a significant difference in the predicted direction between high and low selective attention raters' DA means for disliked ratees [$\underline{t}(240) = 3.28$,$\underline{p}<.05$].

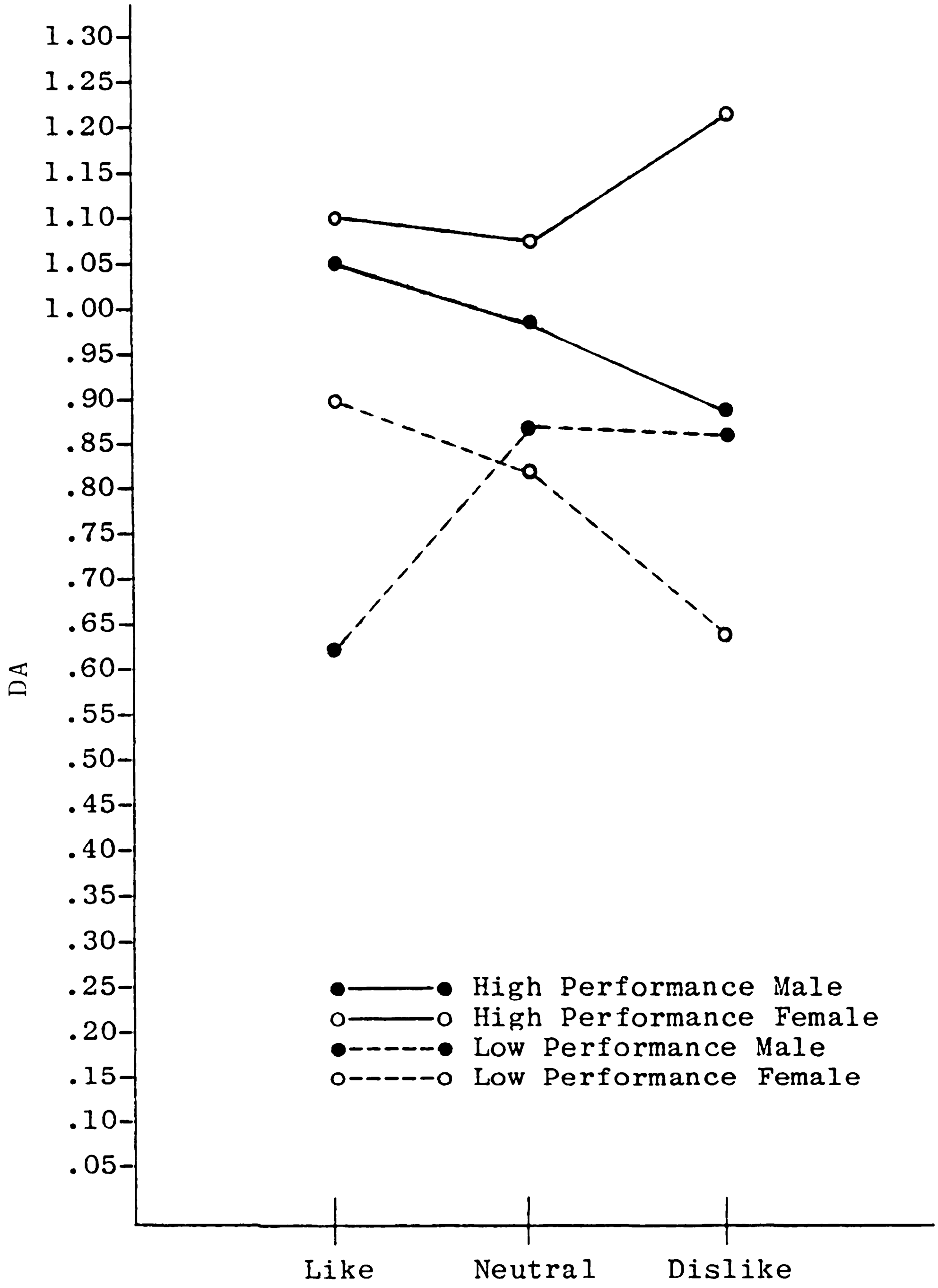

Figure 1: Mean DA as a Function of Ratee Likableness, Ratee Performance, and Rater Sex.

Memory

Hypothesis 4(a) predicted less accurate ratings when based on recall than when not based on recall. The ANOVA revealed that ratings under recall were significantly less accurate than ratings under no recall [$\underline{F}(1,240) = 111.64, \underline{p} < .05$].

Hypothesis 4(b) predicted that, relative to neutral ratees, liked and disliked ratees would be rated even less accurately when ratings were based on recall than when ratings did not require recall. The ANOVA revealed the memory by liking interaction to not have a significant effect on DA ($F < 1.00$).

Hypothesis 4(c) predicted that there would not be a significant memory by liking interaction. The present result is consistent with this prediction. However, hypothesis 4(c) presupposes a liking main effect on accuracy. Thus, the insignificance of this interaction is consistent with the integrality prediction, but the overall pattern of results precludes drawing support for the hypothesis.

The results of study 1 failed to support the main effect for liking predicted in hypothesis 1. DA of ratings did not differ among likable, dislikable, and neutral ratees. Perhaps affect doesn't have a direct influence on performance ratings. However, as discussed previously, the various empirical findings and possible mechanisms through which affect operates indicate that affect should influence nonaffective judgments. There would seem to be little reason to believe that affect should have less influence on performance appraisal judgments than it has been demonstrated to have on other cognitive judgments.

The lack of a liking main effect may be due to the weakness of the liking manipulation. This possibility seems quite probable given the level of strength of association between the liking manipulation and the post hoc liking ratings. Estimates of omega squared values for each of the significant influences on liking ratings are shown in Table 4. As can be seen from inspection of this table, the liking manipulation accounts for only 2.5% of the variance in post hoc liking ratings in contrast to the 54% accounted for by ratee performance level. It is clear that the liking manipulation should be strengthened.

The lack of a direct influence of liking on DA may also be due to the procedure employed in study 1. All of the student raters in study 1 were asked to familiarize themselves with the rating form prior to studying the vignettes. The reason for this procedure was due to the memory manipulation. Subjects in the no memory condition were instructed to look back at the ratee vignettes when rating. In this situation the raters could go back and study the vignettes after "discovering" the rating dimensions. In the memory condition the subjects were not allowed the opportunity to restudy the vignettes. If subjects in the memory condition were not presented with the rating form prior to ratee information, then their ratings would be based on recall as well as being made under the condition of observing the rating format subsequent to observing ratee vignettes. Without presenting the rating

Table 4

Partial Omega Squared Estimates for Significant Sources of Variance In Post Hoc Liking Ratings

Source	$\hat{\omega}^2$
Liking	.025
Performance	.540
Memory	.020
Selective Attention * Sex	.020
Memory * Sex	.010

format prior to ratee vignettes, the memory condition would differ from the no memory condition in the recall requirement and in observing the rating form only after study of ratee vignettes. To avoid this confounding, all subjects were presented the rating form prior to presentation of ratee vignettes. However, the procedure of presenting the rating form prior to ratee vignettes may provide the rater with a context for studying and evaluating the ratees (cf., Bernardin & Walter, 1977). When the rating form is presented prior to ratee observation it may serve as a prime (Wyer & Srull, 1979) and set up a schema for ratee observation in which affect is clearly recognized as irrelevant. Thus, the procedure of presenting the rating form prior to ratee vignettes may have negated the potential main effect of affect.

The performance level of ratees was found to be a relatively important determinant of rating accuracy. As with previous findings (Gordon, 1970; 1972), low performers are rated less accurately than high performers. While the reason for the performance effect is unclear, there are at least two possibilities. First, low performers may be rated less accurately because less information is sought for disliked ratees (Landy & Farr, 1980). The strong relation between ratee performance level and post hoc liking ratings supports the possibility that liking may mediate the performance effect. Another possibility for the performance effect is the difference in schemata for low and high performers. Given that schemata develop from experience (Neisser, 1967), it may be that students have more experience with high performing instructors than with low performing instructors. Thus, various levels at the high end of the performance distribution may be better differentiated than various levels at the low end of the distribution due to a high performance schema being more developed than a low performance schema.

The integrality conceptualization of affect predicted no liking by performance interaction. While this interaction was found to be nonsignificant, the lack of a main effect for liking precludes support for the integrality position being drawn from the null interaction effect.

While the insignificant liking by performance interaction does not support the schema conceptualization of liking, the simple interaction effects analyses of the liking by performance by sex interaction do support the schema possibility. Simple effects analyses of the triple interaction revealed that the consistency of liking and performance information was a significant influence on rating accuracy when rater sex was taken into account. The pattern of results indicates that female raters are more accurate than male raters in inconsistent situations while male raters tend to be more accurate than female raters in consistent situations.

The operation of affect as a schema seems to be compatible with the findings for males but not for females. However, it may be that both liking and performance form the category or schema for processing ratee information and for evaluating ratees. If this is the case, then it appears that females have a better schema for inconsistent affect and performance than do males, while males have a better schema for consistent affect and performance. Theoretical conceptualizations concerning sex roles and stereotypes (e.g., Rosen & Jerdee, 1975; Ashmore & DelBoca, 1979) do not appear to offer further explanation of the present finding. The impact of rater sex on the accuracy of performance ratings is in contrast to the conclusion drawn by Landy and Farr (1980) that there has been no consistent findings on the effect of rater sex on performance ratings. While inspection of the partial omega squared values in Table 3 indicate that the strength of association between the triple interaction and DA is fairly small, it is still of theoretical interest. Rather than offering further speculation on the effect, it would seem best, given the questionable consistency of rater sex effects, to first establish the replicability of the present pattern of results.

As predicted in hypothesis 3(a), high selective attention raters provided ratings that were significantly more accurate than did low selective attention raters. This finding is consistent with the contention that high selective attention raters are better able to deal with differing performance levels across performance dimensions and ratees than are raters of low selective attention. In other words, performance rating is a feature analytic task rather than a wholistic integration task, and those persons whose perceptual style tends to be feature analytic are better at this task than those whose perceptual style tends to be global. In addition to this main effect for selective attention, the significant simple main effect of selective attention on the accuracy of rating dislikable performers provides partial support for hypothesis 3(b). Hypothesis 3(b) predicted that liked and disliked ratees would be rated more accurately by raters of high selective attention than by raters of low selective attention. Planned comparisons did reveal that high selective attention raters rated dislikable performers more accurately than did low selective attention raters. There was no significant difference in rating accuracy between high and low selective attention raters when rating likable performers. The significant comparison between low and high selective attention raters when rating dislikable performers may only reflect the main effect for selective attention, not an increased separability of affect from ratee performance attributes for high selective attention raters. The pattern of results, a significant selective attention main effect and no interaction, is consistent with hypothesis 3(c), the schema conceptualization of affect.

The memory manipulation had the expected main effect on DA. However, the memory by liking interaction predicted in hypothesis 4(b) was not significant. Again, while the insignificance of this interaction was predicted by hypothesis 4(c), it does not support the integrality position since the viability of the hypothesis is dependent upon a main effect for liking.

In sum, the results of study 1 failed to demonstrate a major influence of affect and differentiate between the various conceptualizations of affect. The liking by performance by rater sex interaction indicates that the influence of liking may be dependent upon the levels of the other factors. However, as indicated by the omega squared values, the triple interaction is a relatively unimportant determinant of rating accuracy. The factors of performance, memory, and selective attention were included in study 1 due to their anticipated importance in elucidating the effect of affect. While the results of study 1 do not allow the operation of these factors in this moderating role to be investigated, the direct importance of these factors for rating accuracy was demonstrated.

Study 2

The second study also investigated the effects of liking for ratees on the DA of performance ratings. Based on a consideration of study 1 results, the liking manipulation was strengthened and the order of format presentation was directly investigated. The study also included the factor of ratee performance and was balanced for rater sex.

It was suggested that the lack of effects for liking in study 1 may have been due to the weakness of the liking manipulation. The liking manipulation in study 1 simply consisted of six trait terms included in each instructor vignette. Each subject in study 1 was presented with only one level of liking. In study 2 the salience of the liking manipulation was increased by presenting all three liking levels to each subject.

It was also suggested that the presentation of the rating form prior to ratee vignettes may have produced the lack of liking effects in study 1. Bernardin and Walter (1977) found that the presentation of a rating form prior to ratee observation yielded psychometrically superior ratings than when the rating form was presented subsequent to ratee observation. If this effect is due to a more relevant context or schema being provided by the prior presentation of the evaluation scale, then any potential effects of liking in study 1 may have been overridden by this schema. The order of rating form presentation, before or after ratee observation, may be an important factor when considering the effects of liking on performance ratings.

Study 2 Hypotheses

Liking and Performance Appraisal

Based on the rationale presented for study 1, the following hypothesis was again proposed.

> Hypothesis 1: Liked and disliked ratees will be rated less accurately than affectively neutral ratees.

Ratee Performance and Performance Appraisal

The results of study 1 indicated that low performers are rated less accurately than high performers. Study 2 was designed to replicate this finding.

In addition to a main effect for ratee performance level on rating accuracy, the rationale presented in study 1 indicated that ratee performance level may moderate the effect of liking on DA. This possibility will again be examined in study 2. Briefly stated, it has been proposed that affect may operate as a schema or as an integral dimension. Given the schema conceptualization, affect systematically influences processing and recall and the consistency of performance with affect should be a determinant of rating accuracy. Given the integrality conceptualization, affect obscures the differentiation of stimulus attributes and the consistency of the dimensional values of those attributes with affect should not influence the effect.

The following hypotheses were again proposed.

> Hypothesis 2(b): If the integrality conceptualization of affect is correct, then the effect of affect on rating accuracy should not be dependent upon ratee performance level.
>
> Hypothesis 2(c): If the schema conceptualization of affect is correct, then there should be a liking by performance interaction with consistent liking and performance situations (i.e., likable high performers and dislikable low performers) rated more accurately than inconsistent liking and performance situations (i.e., likable low performers and dislikable high performers).

Order of Rating Format Presentation and Performance Appraisal:

As previously indicated, Bernardin and Walter (1977) found that raters who were familiar with rating dimensions prior to ratee observation committed less halo error than raters who first saw the rating scales subsequent to ratee observation. They interpreted this effect as being due to the context or schema provided by the rating form. If the format effect is mediated by the schema it provides, then raters should be more accurate when presented with the rating format prior to ratee observation than when presented with the format after ratee observation.

The order of format presentation may also influence the effect of affect. In other words, providing a schema for processing information may cause people to overlook information that might otherwise form the basis of a schema. The results of a study by Hamilton (note 9) support this general contention. He found that, given the same stimulus information, subjects who were primed with a gender-based schema were most accurate in recalling gender-related information, whereas subjects who were primed with an academic major based schema were most accurate in recalling major-related information. A similar type of finding has been reported by Hoffman, Mischel and Mazze (1981) for differing observational purposes. Thus, if affect operates as a schema, providing a performance-based schema prior to affect information may reduce or negate any effects of affect. The performance based schema may be recognized as being more relevant to the purpose of ratee observation and dominate or be chosen over affect-based information processing. An affect based schema may operate in processing ratee information when the rating form is presented after ratee observation.

The joint effects of liking and performance may also be dependent upon the order of format presentation. If the performance dimensions schema provided by the rating format precedes ratee observation, then the consistency of affect with ratee performance should be unimportant. However, if the rating format is presented subsequent to ratee observation, then the consistency of liking and performance should be a determinant of rating accuracy.

If affect operates as an integral dimension, then the order of format presentation should not influence the effect of affect. Even when subjects are told the relevant and irrelevant dimensions of an integral stimulus and given practice in separating the dimensions over a number of days, subjects still exhibit a significant interference effect (Monahan, Ingate, Diamond, & Kaniper, note 10). Thus, even when supplied with a schema that indicates what is relevant for the task, an integral dimension maintains its effect. The order of format presentation should not influence the effect of liking if it operates as an integral dimension.

Based on the above rationale, the following hypotheses were proposed.

> Hypothesis 3(a): Performance ratings will be more accurate when the rating format is presented before ratee observation than when presented after ratee observation.
>
> Hypothesis 3(b): If the schema conceptualization is correct, then the effect of liking will depend upon the time of format presentation. The expected liking by format presentation interaction will involve no effect for liking when the format is presented prior to ratee observation but decreased rating accuracy for liked and disliked ratees relative to affectively neutral ratees when the format is presented after ratee observation.
>
> Hypothesis 3(c): The joint effect of liking and ratee performance will depend upon the order of format presentation. When the format is presented prior to ratee observation there should be no effect of liking and performance consistency on rating accuracy. When

the format is presented after ratee observation then
the consistency of liking and performance should be
a determinant of rating accuracy.

Method

Subjects

The subjects in study 2 were 72 male and 72 female introductory psychology students. All subjects received extra credit for participation in the experiment.

Experimental Design

The experimental design was a 3x2x2 (liking x ratee performance x format presentation) between subjects factorial. A total of 12 subjects, 6 male and 6 female, were randomly assigned to each of the 12 treatment combinations.

Instructor Vignettes

Vignettes describing hypothetical instructor ratees were identical to those used in study 1.

Liking manipulation. Liking for ratees was manipulated by describing ratees with the likable,dislikable, and neutral trait terms used in study 1. The same sets of six trait terms were used in the ratee descriptions. However, random samples of the likable, dislikable, and neutral trait terms were presented in the instructions. The instructions included a cover story that indicated that the trait terms were used by students in a previous experiment to describe their instructors. The subjects in the present study were further told that a second group of students had grouped these descriptive terms into three sets and that the four instructor descriptions they would be rating all came from one of these sets. Thus, the different likableness levels were clearly contrasted in the instruction cover story.

Procedure

Subjects were run in groups of approximately 30 individuals. Subjects first were given directions for the rating task. There were two sets of instructions, one for the format before ratee observation condition and another for the format after ratee observation condition. The only difference between the two sets of instructions was that in the format before condition subjects were directed to familiarize themselves with the rating form before studying the vignettes while in the format after condition subjects were directed to not look at the rating form until they had studied the vignette and were ready to begin rating. Ratings were done from recall in both conditions. Format before and format after subjects were run in separate groups.

Except for the presentation of the three liking levels and the instructions for the format before and format after conditions (Appendix G), the procedures of study 2 were identical to those of study 1.

Results and Discussion

Order of Instructor Vignettes

The order of instructor vignettes for each subject was again determined with a Latin square. A one-way analysis of variance on the four unique orders revealed no significant order effect on DA [$\underline{F}(3,92) = .99, \underline{p} > .05$]. The order factor will be collapsed across in all subsequent analyses.

Liking Manipulation Check

As in study 1, after completion of the rating task each subject was asked to provide a summary rating of liking for the set of four instructors/he she had rated. This rating was entered as a dependent measure in a 3x2x2 (liking x performance x format presentation) ANOVA. The analysis revealed a significant main effect for performance [$\underline{F}(1,132) = 102.73$, $\underline{p} < .05$]. There was also a significant format presentation by performance interaction [$\underline{F}(1,132) = 4.64, \underline{p} < .05$]. No other sources had significant effects on the post hoc liking ratings.

The mean liking ratings for liked, disliked, and neutral ratee sets were 3.73, 3.21, and 3.38, respectively. Estimates of partial omega squared values (Keppel, 1973; Maxwell, Camp, & Arvey, 1981) for the liking, performance, and performance by format interaction sources of variance were .03, .41, and .02, respectively. As in study 1, the liking manipulation was effective, but performance was a more important determinant of post hoc liking ratings.

Liking

Hypothesis 1 predicted ratings of liked and disliked ratees to be less accurate than ratings of affectively neutral ratees. A 3x2x2 (liking x performance x format presentation) ANOVA revealed that the liking manipulation did not have a significant main effect on DA ($\underline{F}<1.00$). A summary table of the overall analysis is presented in Table 5. The mean DA of ratings of liked, disliked, and neutral ratees were .49, .49, and .48, respectively.

Performance

The overall analysis supported the predicted effect of ratee performance level on rating accuracy. Ratee performance level was a significant source of variance in rating accuracy [$\underline{F}(1,132)=10.81, \underline{p}<.05$], with DA means for high and low performers of .60 and .37, respectively.

In addition to a main effect for performance, Hypothesis 2(b) predicted no liking by performance interaction while Hypothesis 2(c) did predict this interaction. It was found that the liking by performance interaction was a significant source of variance in DA [$\underline{F}(2,132)=3.85, \underline{p}<.05$]. Thus, the pattern of results do not support Hypothesis 2(b). However, Hypothesis 2(c) predicted that the interaction would reflect increased accuracy when affect and performance were consistent and decreased accuracy when affect and performance were inconsistent. An apriori comparison of the average of DA in consistent conditions with the average of DA in inconsistent

Table 5

Summary of Analysis of Variance on DA

Source of Variation	Sum of Squares	df	Mean Square	F	$\hat{\omega}^2$
Liking (LIK)	0.003	2	0.002	0.01	---
Performance (PERF)	1.981	1	1.847	10.81*	.06
LIK x PERF	1.412	2	0.706	3.85*	.04
Format	2.334	1	2.101	12.74*	.07
LIK x FORMAT	0.180	2	0.090	0.49	---
PERF x FORMAT	0.209	1	0.250	1.14	---
LIK x PERF x FORMAT	1.302	2	0.651	3.55*	.03
Error	24.186	132	0.183		
Total	31.607	143			

*p<.05

conditions revealed that the difference was not significant [$t(132) = 1.31, p < .05$]. A Tukey's post hoc multiple comparison test for unconfounded means (Linton & Gallo, 1975) indicated that DA was significantly higher in the liked high performance condition than in the liked low performance condition [HSD $(5,132) = .337, p < .05$]. No other unconfounded mean comparisons were significant. Thus, the accuracy of rating did not differ across liked, neutral, and disliked ratees at either performance level, and high performers were rated more accurately than low performers only when they were liked.

The interaction provides only partial support for hypothesis 2(c). DA did not significantly differ across consistent and inconsistent situations. Only the comparison of liked high performers (a consistent condition) with liked low performers (an inconsistent condition) revealed a significant accuracy difference in favor of consistency. The liking by performance interaction is graphically presented in Figure 2. Visual inspection of the interaction indicates that the overall pattern is not highly supportive of Hypothesis 2(c).

Format Presentation

The significant main effect for format presentation [$F(1,132) = 12.74, p < .05$] indicates significantly more accurate ratings when the rating format was presented before the ratee vignettes than when the format was presented after the ratee vignettes. The DA means in the format before and after conditions were .61 and .36, respectively.

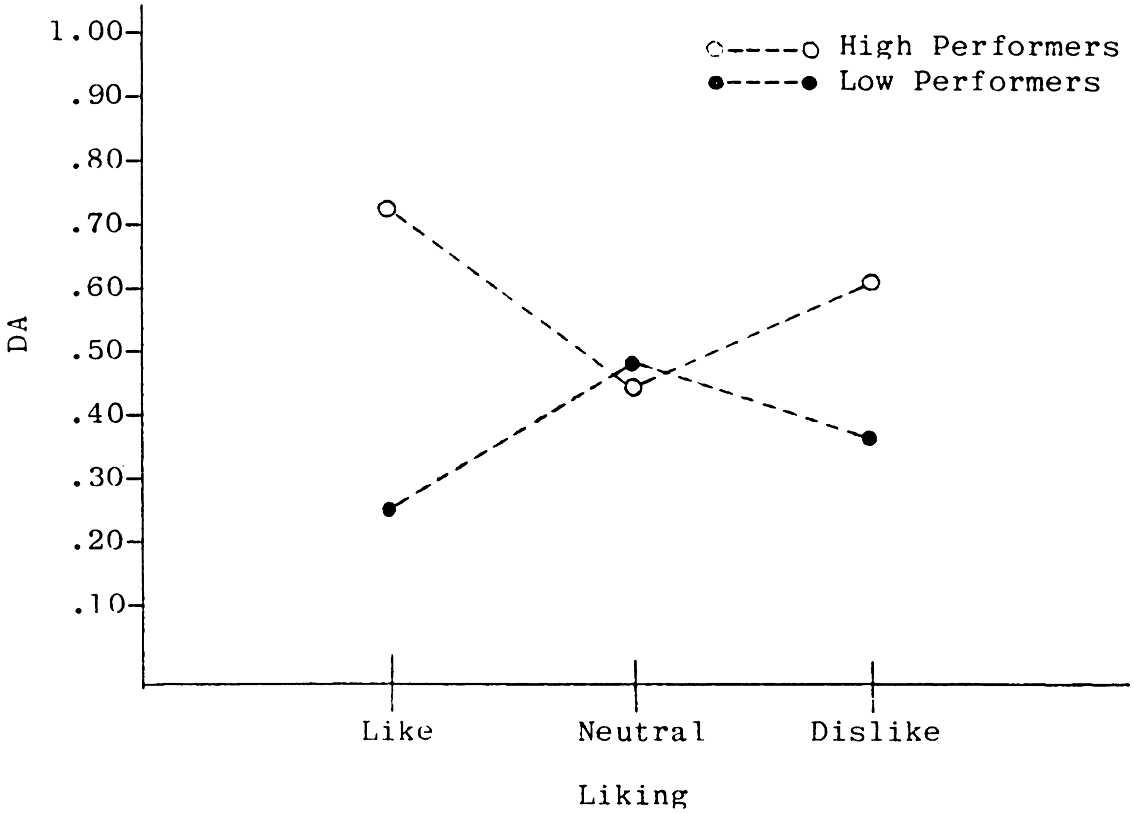

Figure 2. Mean DA as a Function of Liking and Ratee Performance Level.

Hypothesis 3(b) predicted a liking by format presentation interaction, with no effect for liking in the format before condition but decreased rating accuracy for liked and disliked ratees relative to neutral ratees in the format after condition. The overall analysis revealed the liking by format interaction to not be significant ($F<1.00$), and a priori comparisons of liked and disliked with neutral conditions revealed no significant differences in DA in either the format before or format after condition ($\underline{p}_s>.05$).

Hypothesis 3(c) predicted a liking by performance by format presentation interaction, with the consistency of liking and performance being a determinant of rating accuracy only in the format after condition. The overall analysis revealed the triple interaction to be a significant source of variance in rating accuracy [$\underline{F}(2,132) = 3.55, \underline{p}<.05$]. The interaction is graphically represented in Figure 3. A visual inspection of the figure indicates that liking and performance have a joint influence on rating accuracy only in the format after condition. Simple effects analyses of the two way interactions within format presentation conditions support this conclusion. The analyses revealed that the liking by performance interaction significantly affected DA in the format after condition [$\underline{F}(1,132) = 9.49, \underline{p}<.05$] but not significant in the format before condition [$F(1,132) = 2.46, \underline{p}>.05$].

Hypothesis 3(c) further proposed that when the format was presented after, conditions in which liking and performance are consistent would be rated more accurately than conditions

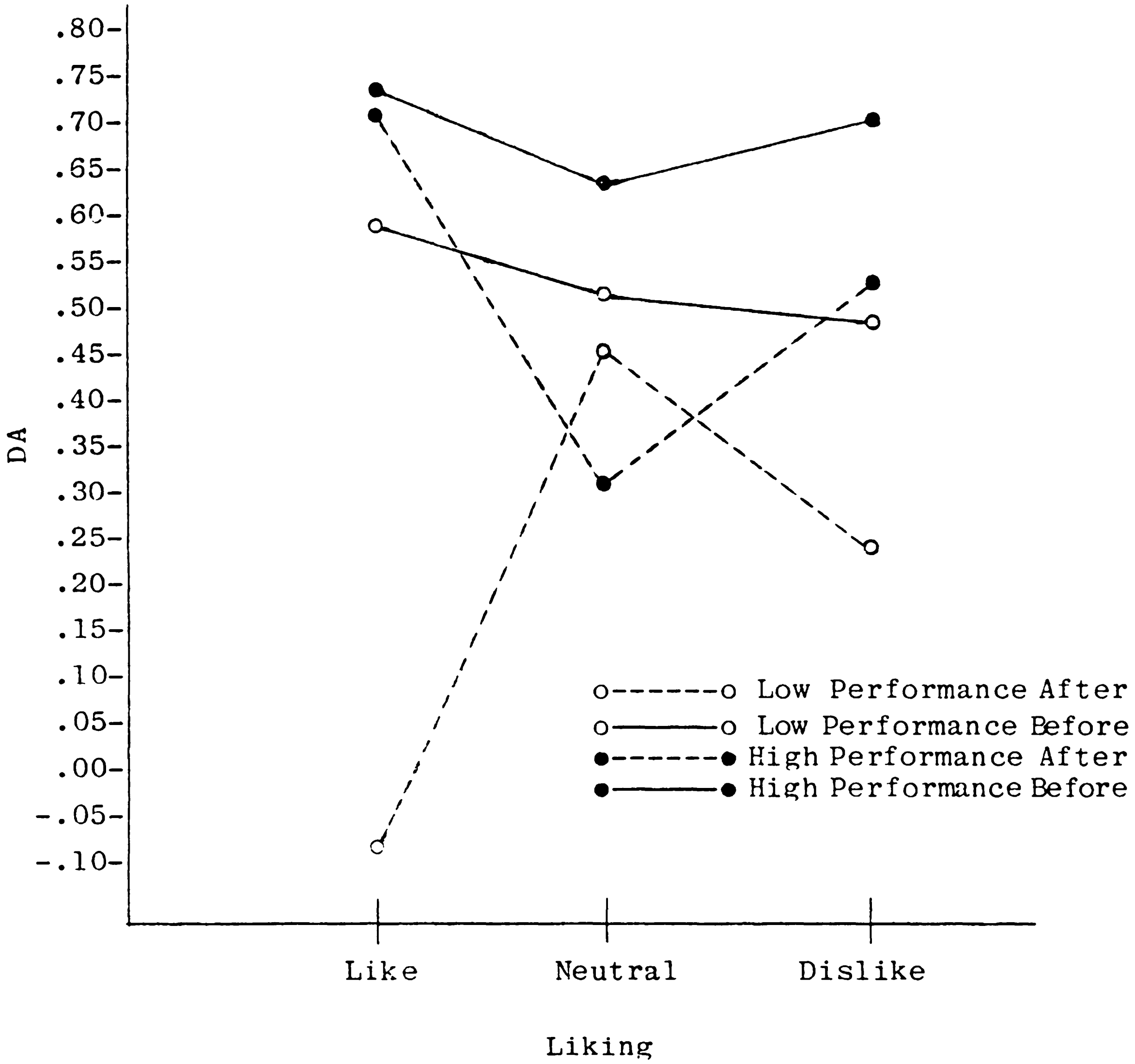

Figure 3. Mean DA as a Function of Liking, Performance, and Format Presentation

in which liking and performance are inconsistent. An a priori comparison of the average DA in consistent conditions with the average DA in inconsistent conditions revealed DA to be, on the average, higher in consistent than in inconsistent conditions when the format was presented after ratee vignettes [$t(132) = 2.92, p < .05$]. While the significance of this comparison supports Hypothesis 3(c), it does not fully portray the nature of the two way interaction under the format after condition.

A Tukey post hoc multiple comparison test for unconfounded means was used to compare the DA means involved in the format after simple interaction. This analysis revealed that DA was significantly different between high and low performance conditions only for liked ratees. In addition, it was also found that DA was significantly higher for ratings of liked ratees than for ratings of neutral ratees at both performance levels. No other unconfounded mean comparisons were significant. Thus, DA was found not to differ significantly between high and low performance conditions for neutral or disliked ratees nor between neutral and disliked conditions at either performance level. While visual inspection of the interaction indicates that, in the format after condition, affect (both liking and disliking) increased the accuracy of ratings of high performers but decreased the accuracy of ratings of low performers, this simple visual inspection is misleading. The multiple comparison test revealed DA between liked and neutral conditions to be significantly different at both performance levels while DA between disliked and neutral conditions did not differ significantly at either performance level.

The significant triple interaction qualifies the interpretations of the performance, liking by performance, and format effects. For example, while it is generally the case that high performers are rated more accurately than low performers, a more accurate statement is that the effect of ratee performance level on DA is dependent upon liking for ratees and the order of format presentation. The most succinct and accurate description of the effects found in study 2 is the triple interaction. The interaction indicates that when the rating form is presented before ratee observation, liking and ratee performance have no significant effects on rating accuracy. However, when the rating form is presented after ratee observation, liking and ratee performance significantly affect rating accuracy. Results of analyses of the simple liking by performance interaction effect indicate that performance affects the accuracy of rating only when affect is positive. In addition, positive and neutral affect for ratees differentially affects the accuracy of rating high and low performers. That is, neutral high performers are rated less accurately than liked high performers, but neutral low performers are rated more accurately than liked low performers.

The significant triple interaction effect and, in the format after condition, the higher average accuracy in consistent liking and performance conditions than in inconsistent liking and performance conditions supports Hypothesis 3(c). However, consideration of the pattern of results indicates that the consistency effect is produced by the accuracy difference between liked high and low performance conditions. Thus, the schema consistency position receives only partial support when the overall pattern of results is considered.

In sum, the results of study 2 do not support the integrality conceptualization of affect and offer only partial support for the schema conceptualization. It was argued that if affect operates as an integral dimension, then the time of format presentation and ratee performance level should not influence the effect of affect on appraisal accuracy. However, it was found that both factors interact with the effect of liking on accuracy. Thus the integrality position received no support from the results of study 2.

From the schema conceptualization of affect it was predicted that both format presentation and ratee performance level would influence the effect of affect. More specifically, it was expected that liking would influence rating accuracy only in the format after condition and that the consistency of ratee performance with liking would be a determinant of accuracy. Analyses of the significant triple interaction revealed the predicted importance of the order of format presentation and of the consistency of liking and performance. However, the overall pattern of the interaction and further analyses revealed the consistency finding to be due to the accuracy difference between the likable low and high performance conditions. Although found not to involve significant comparisons, the interaction indicated disliking to increase accuracy of high performer ratings but decrease accuracy of low performer ratings, just the opposite of what was expected from the schema viewpoint. Thus, the pattern of results offer only partial support for the schema viewpoint.

A reconsideration of the schema perspective may offer a possible explanation for the results of study 2. As discussed previously, a schema is a generic knowledge structure. Thus, a schema is based on prior experience with the class of objects, situations, or categories in question. It was proposed that affect may operate as a schema. Based on this possibility, it was proposed that liking and disliking may be cognitive categories that guide raters' processing and recall of ratee performance information. Given the current theorizing and empirical findings on schematic processing (e.g., Ashmore & DelBoca, 1979; Hastie, 1981), it was expected that the consistency of performance with affect would be an important determinant of rating accuracy. However, it may very well be that both affect and performance form a cognitive category or schema. That is, there may not be schemata based only on affect, rather, there may be cognitive categories that include affect. For example, Fiske and Beattie (note 11) contend that liking for a stimulus is an important factor in the schema for the stimulus. In support of this contention, they found that the degree to which characteristics of a person match an old-flame stereotype, a schema based on past romantic partners, is a determinant of affect for the person. The degree of match was also found to be a determinant of the setting, task or social, preferred for meeting a person. Fiske and Beattie conclude that schemata can include feelings as well as nonaffective features.

Given the possibility that cognitive categories may include affect, the cues of likableness and performance may activate an affect-laden schema relevant to the general performance level being observed and processed. For example, in the present domain, the affect and performance cues may activate schemata such as a liked high performing instructor or a disliked low performing instructor.

If the above analysis is correct, then the extent of prior experience with the various liking and performance combinations should be the determinant of accuracy, not the consistency of the cues within a particular combination. The extent of prior experience, the complexity of the schema, may directly affect information processing (Neisser, 1967) and has been proposed to be a determinant of performance rating accuracy (Bernardin, Cardy, & Abbott, note 8). The basic notion here is that less prior experience means a less defined schema which results in a lower level of discriminability of stimuli within the domain.

The above reconsideration of the schema perspective leads to the possibility that the results of study 2 are due to differences in schema complexity among the various liking and performance combinations. More specifically, it may be that student raters have more experience with liked high performing instructors than with liked low performing instructors. Thus, they have a better schema for the former condition and are more accurate rating liked high performers than liked low performers. In addition, an affectively neutral high performer may be a less typical situation than an affectively neutral low performer. Thus, affectively neutral high performers would be rated less accurately and affectively neutral low performers more accurately than liked high and low performers, respectively. The entire pattern of the simple effect interaction may be explained from this perspective if it is more common to have definitive affect towards high performing instructors but more common to have less definitive or neutral affect towards low performing instructors.

Another possible explanation for the results of study 2 is the operation of discounting. Discounting is the reduction of stimulus weight or importance in information processing (e.g., Kaplan, 1973). It may be that low performance tends to be discounted, particularly when there is positive affect for the low performer. In contrast, it may be that high performance tends not to be discounted, particularly when there is positive affect for the high performer. At the low performance level, the decreased accuracy when rating liked rather than neutral ratees would be due to the decreased importance of performance in the liked condition. At the high performance level, the decreased accuracy when rating neutral rather than liked ratees would be due to the decreased importance of performance in the neutral condition.

A third possible explanation is a positively biased recall. As discussed previously, we may be positively biased processors, a phenomenon termed the Polyanna Principle by Matlin and Stang (1978). Perhaps this bias is accentuated by positive affect. If this is the case, then the bias would be more representative of high performers than low performers, resulting in the accuracy difference in rating high and low performing liked ratees observed in study 2. The difference in accuracy of ratings of liked and neutral ratees may also be due to the differential representativeness of the positive bias.

Study 3

The third study again investigated the effects of liking and ratee performance on the DA of performance ratings. Based on a consideration of study 2 results, schematic, discounting, and recall processes were also investigated.

It was suggested that the accuracy difference between ratings of liked high and low performers may be due to schema, discounting, or recall differences between the two conditions. Liked low performers may be rated less accurately than liked high performers due to a less accurate schema for the former than for the latter condition. The accuracy differency may also be due to the discounting of low performance information. In addition, the accuracy difference may be due to a bias to recall positive information.

Study 3 Hypotheses

DA

Based on the findings of the previous study, the following hypothesis was proposed.

> Hypothesis 1: Liked low performers will be rated less accurately than liked high performers.

Schema

Consideration of the results of study 2 indicated that differences in schemata may have produced the accuracy differences observed in study 2. It was proposed that, perhaps due to an experiential factor, the schemata for the various liking and performance combination conditions may differ in their degree of development and accuracy. A more accurate schema should allow for more accurate ratings than would a less accurate schema (cf., Bernardin, Cardy, & Abbott, note 8). In addition, schema-based ratings should be made more confidently when the schema is more developed.

Based on this rationale, the following hypothesis was proposed.

> Hypothesis 2: The goodness-of-fit of rater's schemata for liked and disliked, high and low performers will correlate with DA.

Discounting

Consideration of the results of study 2 also indicated that discounting may have produced the differences in rating accuracy. Specifically, it was proposed that low performance information tends to be discounted, particularly when there is positive affect for the low performer. However, high performance information tends not to be discounted, particularly when there is affect that is consistent with high performance. Thus, the differences in rating accuracy may be due to differences in the importance placed on performance information.

Based on this rationale, the following hypothesis was proposed.

> Hypothesis 3: The importance placed on performance information will correlate with DA.

Recall

That recall bias may have produced the accuracy differences in study 2 was also considered a possibility. It was speculated that a positive recall bias may be accentuated by positive affect. If so, the bias would result in recall that was more representative of high performers than low performers, resulting in the observed differences in rating accuracy.

Based on the above rationale, the following hypothesis was proposed.

> Hypothesis 4: The tendency to recall positive information will be correlated with DA.

Method

Subjects

The subjects in study 3 were 4'
ductory psychology students. All s
credit for participation in the exp

Experimental Design

The experimental design was a 2x2 (liking by ratee performance) between subjects factorial. A total of 20 subjects, 10 male and 10 female, were randomly assigned to each of the four treatment combinations. Two female subjects who were randomly assigned to additional inverted vignette orders were deleted from analyses.

Instructor Vignettes and Liking Manipulation

Vignettes describing hypothetical instructor ratees were identical to those used in the previous studies. The liking manipulation was identical to the study 2 liking manipulation. However, only two levels of likability were included in study 3, likable and dislikable.

Procedure

Subjects were run in groups of approximately 20 individuals. Subjects participated in two sessions that were one week apart. Schema and discounting information was collected in the first session while performance ratings and recall information were collected in the second session.

Session 1. In session 1 all subjects were given two tasks. Materials for the first task consisted of a general description of an instructor and a list of 20 teacher behavior incidents. The general description included three trait terms and indicated that the instructor received teacher evaluations that were typically either above or below average. The three trait terms were either liked or disliked traits that were used in the instructor vignettes to manipulate liking. Each subject received a general description that was representative of one of the four liking and performance treatment combinations. The list of incidents contained 10 high and 10 low performance incidents that were presented in random order. There were two high and two low performance incidents representative of each of the five teacher behavior dimensions.

Subjects were given a total of seven minutes to indicate, for each of the 20 behavioral incidents, whether or not they would expect the instructor to exhibit that behavior and to indicate their sureness about each expectancy. Expectancies for high performance incidents were scored as correct when they were expected in the high performance condition and when not expected in the low performance condition. Expectancies for low performance incidents were scored as correct when they were expected in the low performance condition and when not expected in the high performance condition. The number of correct expectancies and the average sureness rating were the two measures derived from performance on this task. Since a fundamental effect of schemata is to provide expectancies (e.g., Neisser, 1967), either the accuracy of these expectations or their strength should reflect the development of the schema within each of the four liking and performance treatment combinations.

The second task in the first session consisted of four brief instructor descriptions, each containing a trait term and five teacher behavior incidents. Each ratee description was representative of the particular liking and performance condition to which the subject had been assigned. Thus, each description contained a likable (or dislikable) trait term and three high (or low), one neutral, and one low (or high) level behavioral incidents.

Subjects were asked to make an overall teaching effectiveness rating of each instructor description and to rate the importance of each of the six bits of information in forming the overall evaluation. The average importance rating assigned to performance information items was calculated for each subject. This measure should reflect the importance or weight given to performance information in each of the four liking and performance combination conditions.

Session 2. The second session was held one week after the first session. During the second session, each subject received the set of four instructor vignettes from the condition to which the subject had been assigned. The rating condition was identical to the format after condition in study 2. Order of instructor vignettes was again controlled for with a Latin square.

After completing the rating task, all subjects received a second set of materials consisting of a distractor task and a recall test. In the distractor task subjects were asked to rate 27 non-teaching behavioral incidents in terms of how much they would like or dislike a person exhibiting each behavior. The recall test consisted of 28 items, 16 behavioral incidents and 12 trait terms. Of the 16 behavioral incidents, 8 had been included in the instructor vignettes and 8 were not included in the vignettes and were opposite of the performance level of the vignettes. Of the 8 that were in the vignettes, 4 of the incidents were opposite and 4 were the same level of performance of the vignettes. For example, if the subject was in a high performance condition the recall test would include 8 old incidents that could be correctly recognized, four high performance and four low performance incidents. The test would also include eight new low performance incidents that could not be correctly recognized. These were not in the instructor vignettes.

Of the 12 trait terms included in the test, 6 had been in the vignettes and 6 had not been included in the vignettes and were opposite to the liking level of the vignettes. Of the 6 "correct" trait terms, 3 were neutral and 3 were of the same liking level as the instructor vignettes.

Subjects were asked to indicate for each item whether or not they recalled that item as being in any of the four instructor descriptions. The number of "incorrect," opposite level, behavioral incidents recalled as being included in the instructor descriptions was determined for each subject. This measure indicates a bias in the recollection of performance information. For example, a high score on this measure would indicate a bias to recall performance information opposite of the performance level of the instructor vignettes.

Results and Discussion

Order of Instructor Vignettes

The order of instructor vignettes for each subject was again determined with a Latin square. A one-way analysis of variance on the four unique orders revealed a significant order effect on DA [$F(3,60) = 3.54, p < .05$]. Since order is a balanced factor in the design, the order factor will be collapsed across in all subsequent analyses.

DA

Hypothesis 1 predicted ratings of liked high performers to be more accurate than ratings of liked low performers. A 2x2 (liking x performance) ANOVA revealed a significant main effect for ratee performance level on DA [$F(1,76) = 9.71, p < .05$]. A summary table of the overall analysis is presented in Table 6. The mean DA of ratings of liked high and low performers were .48 and .21, respectively. The mean DA of

Table 6

Summary of Analysis of Variance on DA

Source of Variations	Sum of Squares	df	Mean Square	F	$\hat{\omega}^2$
Liking (Lik)	0.02	1	0.02	0.12	---
Performance (Perf)	1.72	1	1.72	9.71*	0.098
Lik x Perf	0.01	1	0.008	0.05	---
Error	13.47	76	0.177		
Total	15.22	79			

*p<.05

ratings of disliked high and low performers were .47 and .16, respectively. An apriori _t_ test revealed that DA was significantly higher for ratings of liked high performers than for ratings of liked low performers [$\underline{t}(76) = 2.9, \underline{p} < .05$].

Schema, Discounting, and Recall

Hypotheses 2, 3, and 4 predicted that the schema, discounting, and recall bias measures would be determinants of rating accuracy. In order to test these hypotheses, DA was regressed on the schema accuracy, schema confidence, discounting, and recall bias measures. The simple correlations among these variables are presented in Table 7. As can be seen in Table 8, the regression analysis revealed recall bias to be the only significant predictor of DA ($\underline{b} = -.10$), $\underline{F}(1,70) = 5.29, \underline{p} < .05$. The results support hypothesis 4, but not hypotheses 2 and 3.

The results failed to support the schema explanation for rating accuracy. Neither the accuracy of behavioral expectations nor the confidence in these expectancies was significantly related to DA. However, it is interesting to note that when entered as a dependent measure in a 2x2 (liking x performance) ANOVA, the schema accuracy measure is significantly influenced by the liking by performance interaction effect [$\underline{F}(1,76) = 160, \underline{p} < .05$]. Summary tables of the separate univariate ANOVA's of the schema, discounting, and recall dependent variables are presented in Table 9. As inspection of this table reveals, the significant liking by performance

Table 7

Correlations Among Variables

Variables	1	2	3	4	5
1. DA	--				
2. Schema Accuracy	-.06	--			
3. Schema Sureness	.122	.127	--		
4. Discounting	.108	.094	.393*	--	
5. Recall Bias	-.246*	-.07	-.03	.076	--

*p<.05

Table 8

Summary of Regression Analysis of DA

Source	Sum of Squares	df	b	F
Schema Accuracy	0.21	1	-.008	1.14
Schema Sureness	0.01	1	.017	0.08
Discounting	0.18	1	.05	0.97
Recall Bias	0.97	1	-.10	5.29*
Error	12.84	70		
Total	14.21	74		

*p<.05

Note: Of the 80 observations, five were deleted due to missing values.

Table 9

Summary of Analyses of Variance on the Schema, Discounting, and Recall Measures

Source of Variation		Sum of Squares	df	Mean Square	F
		Schema Accuracy			
Liking (Lik)		13.61	1	13.61	0.85
Performance (Perf)		21.01	1	21.01	1.32
Lik x Perf		2565.11	1	2565.11	160.99*
Error		1210.95	76	15.93	
	Total	3810.68	79		
		Schema Sureness			
Lik		0.22	1	0.22	0.28
Perf		4.46	1	4.46	5.72*
Lik x Perf		1.01	1	1.01	1.29
Error		58.94	76	0.78	
	Total	64.63	79		
		Discounting			
Lik		3.89	1	3.89	3.91
Perf		0.05	1	0.05	0.05
Lik x Perf		1.37	1	1.37	1.37
Error		75.65	76	0.995	
	Total	80.96	79		
		Recall Bias			
Lik		1.25	1	1.25	0.29
Perf		18.05	1	18.05	4.21*
Lik x Perf		1.80	1	1.80	0.42
Error		326.10	76	4.29	
	Total	347.20	79		

*$p < .05$

interaction effect indicates that behavioral expectations are most accurate when liking and performance are consistent but least accurate when liking and performance are inconsistent. As has been previously discussed, a schema should be more accurate when information is consistent than when it is inconsistent, and the present result supports this contention. However, the pattern of schema accuracy scores is not comparable to the pattern of DA across the four liking and performance combination conditions. As the regression analysis revealed, schema accuracy isn't related to DA.

The schema confidence and discounting measure also were found not to be significantly related to DA. Thus, the confidence in behavioral expectations or the importance placed on performance information cannot explain the accuracy of ratings across the liking and performance combination conditions.

The recall bias measure was found to be significantly related to DA. Recall bias means in the liked and disliked high performance conditions were 1.60 and 1.65, respectively. Recall bias means in the liked and disliked low performance conditions were 2.85 and 2.30, respectively. Consideration of these means indicates that subjects in low performance conditions were more likely to incorrectly recognize high performance incidents than were subjects in high performance conditions likely to incorrectly recognize low performance incidents. The regression analysis revealed this positive recall bias to be the only significant predictor of DA.

In sum, the results indicate support for hypothesis 4, that the bias to recall positive information is a determinant of rating accuracy.

Discussion

The present series of studies was an attempt to demonstrate and explicate the effect of affect on the accuracy of performance ratings. The results of the first two studies demonstrate that liking influences the accuracy of performance ratings. The results of study 1 indicated that the effect of liking on rating accuracy depends upon the performance level of the ratees and the sex of the rater. The results of study 2 indicated that the effect of liking depends upon the performance level of the ratees and the order of format presentation.

While these findings provide a demonstration of its effect, the investigations still leave the nature of the effect of affect inconclusive. The liking triple interaction in study 1 was not predicted, but post hoc consideration of the interaction indicated that sex differences in the schema generated by various liking and performance combinations may account for the interaction. The overall pattern of the results of study 1 failed to provide clear support for either the schema or intergrality conceptualization of affect. The triple interaction in study 2 was predicted from a schema conceptualization of affect, but analysis of the pattern of this interaction provided only partial support for the schema conceptualization. A reconsideration of the schema perspective provided a post hoc explanation of the pattern on the interaction. In sum, the results of studies 1 and 2 offered no support for the integrality conceptualization of affect and only partial support for a schema conceptualization.

Study 3 was an attempt to test potential explanations for the differences in rating accuracy observed in study 2. It was postulated that schematic, discounting, or recall bias processes may account for the accuracy of ratings in the various liking and performance combination conditions, and thus serve to explicate the nature of the effect of affect. Recall bias was found to be the only measure significantly related to DA. Positively biased recall would be expected from the Polyanna Principle (Matlin & Stang, 1978) since it directly proposes that stimulus information tends to be more positive at recall than it was at input. However, the Polyanna Principle doesn't provide an explanation for positive bias, it simply states its existence. The mechanism offered by Matlin and Stang to explain positive bias is the presence of positively biased selectivity throughout the information processing system. Consideration of selectivity in information processing may offer a potential explanation for the positively biased recall observed in study 3.

Evidence indicates that material is attended to, stored, and recalled in terms of cognitive categories of schemata (e.g., Cantor & Mischel, 1977; Hastie, 1981; Shiffrin & Schnieder, 1977; Srull & Wyer, 1979). That recall can be systematically biased by the cognitive category or schema employed in processing stimulus information has been demonstrated by Cantor and Mischel. These investigators found that describing a stimulus person as introvert (or extrovert) led subjects to incorrectly identify introvert (or extrovert) related items as being in the stimulus person description. Apparently, the introvert and extrovert attributes activated different schemata. The schemata employed in processing the stimulus led to a bias to incorrectly recognize information representative of the schema as being included in the stimulus. It has also been demonstrated that this bias can lead to incorrect recognition of either positive or negative information and to positively or negatively biased stimulus evaluation (Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979). For example, Higgins et al. exposed subjects to either positive or negative traits prior to the presentation of a stimulus person description. Even though the traits were supposedly part of an unrelated study, it was found that subjects' interpretation and recall of the stimulus person description were systematically influenced by exposure to the traits. This type of finding indicates that activating or priming a general schema can influence the processing and recall of subsequent information.

The above discussion indicates that recall tends to be biased towards information that is representative of the activated schema and that this bias can either be positive or negative. Thus, it appears that selectivity in information processing is due to the schema employed in the processing and that positively biased selectivity is a special case of schematic processing. In other words, schematic processing can explain the Polyanna Principle and the selectivity mechanism, and further, it indicates that the bias can operate in either a positive or negative direction.

The schema perspective offers a potential explanation of the recall bias finding in study 3. Perhaps student subjects tend to have a positive schema, or stereotype, for instructors. If this is the case, then asking student subjects to rate instructors may activate this general schema. Regardless of the particular liking and performance information, the general schema may predominate processing and recall. Thus, a general and positive schema for instructors may have produced the positive recall bias.

The possibility of a positive schema for instructors is indicated by the finding that the average teacher evaluation rating is above the average rating scale value (cf., Matlin & Stang, 1978). While speculative, it may be that students tend to have favorable experiences with instructors. A positive schema for instructors would be expected to come to represent this experience.

In contrast to the earlier proposed operation of a schema based on liking and performance information, the possibility of a preexisting positive schema for instructors indicates that the consistency of liking and performance information should not be a determinant of processing and recall. The proposed positive schema for instructors is a schema that subjects would bring with them to the rating task, not one built upon the stimulus information. This is not to say that the liking and performance information does not play a role in processing. What may be important about the liking and performance information is its match with the positive instructor schema. The closer the match between the instructor description and the positive instructor schema, the greater may be the reliance upon the schema for processing and recall of the instructor information. However, this positive schema would be more representative of high performing than of low performing instructors. Thus, high performers may have been rated more accurately than low performers since high performance instructor descriptions most closely match the schema that tends to be employed when rating instructors. In addition, as indicated by the results of study 2, positive affect may increase the reliance upon the positive instructor schema for processing and recall of the instructor information. Likable high performers may have been rated more accurately than neutral high performers due to positive affect producing increased reliance on the positive and representative instructor schema. Likable low performers may have been rated least accurately due to positive affect increasing reliance on the positive but unrepresentative instructor schema.

The possibility of a general schema producing the recall bias and, in turn, the effect on rating accuracy is somewhat similar to the proposed operation of a general leadership impression (e.g., Rush, Phillips, & Lord, 1981). Rush et al. supported the possibility that a general impression of the amount of leadership exhibited by a leader is an important determinant of leadership ratings. In the present context it is suggested that a general impression or stereotype of instructors produced the positive recall bias. That is, it is proposed that the general impression of a class of stimulus objects, rather than impressions of objects within the class, may have produced the recall bias. The above difference may be due to instructor schemata being more homogenous than leadership schemata.

It should be pointed out that study 3 failed to find liking involved in any significant effects on rating accuracy. Thus, although recall bias was found to be significantly related to DA, it may serve to explain the effect of ratee performance level on DA but can only be tentatively offered as an explanation for the effect of liking on DA since liking effects were nonexistent. That is, in order to accept the possibility of recall bias as an explanation for the effects of liking on DA, it must be assumed that had there been a liking effect on DA there would have been a corresponding effect on recall bias. However, this assumption may be incorrect and recall bias serve only to explain the effect of ratee performance level on DA while another variable is responsible for the effect of liking on DA.

In addition to the above concern, it should be noted that, although a significant predictor, recall bias accounts for a relatively small portion of the variance in DA. Thus, there is a great deal of variance in DA left unexplained by the recall bias measure. The small proportion of variance accounted for may simply be due to the unreliability of the measure or may indicate that recall bias is a significant predictor but not a very good explanation of DA.

In sum, the results of the series of studies do not directly support the schema or integrality perspectives and offer tenuous support for the recall bias explanation. Regardless of the process or mechanism that mediates the effect of liking, the present investigation is an initial demonstration of the effect of liking on rating accuracy. That the liking effects were relatively unimportant according to omega squared estimates should not be taken as indicating the triviality of liking in performance appraisals. On the contrary, the effect of liking may be magnified when the performance of ratees isn't clearly distinguished from affect and the affect towards ratees isn't artificially induced. While it is an empirical question, there is reason to expect liking to have a greater effect on rating accuracy in "real world" performance appraisals. For example, liking would most likely be stronger and less differentiated from ratee performance in applied than in laboratory appraisal situations. Research directly investigating the effect of liking in applied settings is called for.

The results of the present investigation should not be taken as evidence contrary to or incompatible with the formulation of affect based on reaction time studies. The schema conceptualization of affect was derived from Zajonc's work supporting the temporal precedence of affect over non-affective judgments (e.g., Zajonc, 1980). The integrality coneptualization of affect was derived from the interference findings of Kehoe and his associates (Cardy, Dobbins, & Kehoe, note 2; Kehoe, Cardy, & Dobbins, note 1). Both the temporal precedence and interference effects were found in studies of judgment latency not judgment accuracy. In contrast, the present investigation was concerned with judgment accuracy, the accuracy of discrimination among ratees. Thus, the present findings indicate a limitation of the generality of the early and interfering conceptualizations of the nature of affect, perhaps due to different cognitive processes reflected by judgment latency and accuracy. That judgment latency and accuracy have an unclear relationship is supported by a study directly investigating the influence of affect on the two measures (Dobbins, note 12). An early and interfering nature of affect may be the best explanation for the effects of liking on judgment latency, but does not capture the effects of liking on judgment accuracy.

These studies and results have exclusively focused on the relationship between liking and discrimination accuracy. However, it is possible that the DA results were artificially produced by an elevation effect of liking which reduces the variability of ratings which, in turn, reduces the DA correlations. Recall that the DA measure is the average within dimension correlation between ratings and true scores. Ratings that discriminate among ratees within dimensions comparable to the true score discrimination among ratees yield a high DA score. A ceiling or floor effect could artifactually produce an effect on DA. That is, decreased DA correlations may actually be the result of ceiling or floor effects induced by elevation shifts rather than decreased discriminability per se. However, consideration of the variances of ratings in study 2 indicates that the effects of liking and performance on DA were not produced by a range restriction effect. If range restriction mediated the DA effects then, in the format after condition of study 2, the higher DA for liked high performers than for liked low performers should correspond to higher rating variance for liked high than for liked low performers. However, the variance of ratings of liked high and low performers in the format after condition were 6.19 and 7.71, respectively. In addition, the higher DA for neutral than for liked low performers in the format after condition should correspond to higher rating variance in the former than in the latter condition. The variance of ratings in these neutral and liked low performer conditions were 6.21 and 7.71, respectively, the opposite of what would be expected from the range restriction possibility. Thus, the effects on DA appear to reflect the accuracy of discriminability among ratees rather than the operation of range restriction.

There is also the possibility that differences in true score variability may have produced higher DA for high than for low performers. However, inspection of Table 1 indicates that the standard deviations of true scores tend to favor discriminability of low performers. True score differences between pairs of ratees may also influence DA. For example, on dimension A, there are a pair of true scores which are closer together in the low performance condition than are any pair of true scores in the high performance condition. This difference may favor accurate discrimination among high performers on dimension A. Taking the results of study 2 as an example, DA in the low performance condition on dimensions A through E were .59, .23, 1.06, .17, and -.20, respectively. The corresponding DA means in the high performance condition were .31, .63,.1.33, .62, and .14, respectively. Differences in the spread of true scores between the high and low performance conditions do not correspond to this pattern of DA means. In sum, characteristics of the true scores do not appear to have artificially produced the performance effect.

While the pattern of results across the studies do not provide clear support for an explanation of the effect of liking, the results do have applied implications. The results of study 1 indicate that the sex of the rater may be an important consideration, with females being more accurate raters when their liking for ratees is inconsistent with the performance level of ratees. The results of study 2 indicate that presenting raters with the rating dimensions prior to ratee observation provides ratings that are most accurate and least influenced by factors irrelevant to the appraisals. While Bernardin and Walter (1977) demonstrated the reduction of leniency and halo errors as a function of administration of the rating form prior to ratee observation, the present study extends the influence of the order of format presentation to include rating accuracy. In addition, format presentation in the present study preceded evaluation by only a short period of time, perhaps only 20 minutes, while format presentation in the Bernardin and Walter study preceded evaluation by 10 weeks. Another study (Bernardin, Cardy, & Abbott, note 8) also found significantly reduced halo due to 20 minutes of prior familiarization with the rating format. Subjects in the format before conditions of the present investigation were told only to familiarize themselves with the rating format prior to study of the vignettes. While time spent in familiarization was not recorded, subjects certainly spent less than five minutes familiarizing themselves with the rating form before studying the vignettes. Even though very brief relative to the manipulation in other studies, familiarization of raters with the rating form prior to ratee observation still significantly increased the accuracy of performance ratings. The effectiveness of this manipulation indicates that the very simple and brief procedure of familiarizing raters with the rating form prior to ratee observation may increase the accuracy of performance ratings in applied settings.

The results of the present investigation also indicate that ratee performance level, particularly when ratees are liked, significantly influences the accuracy of performance ratings. The general rating accuracy advantage enjoyed by high performers replicates the findings of Gordon (1970; 1972). The performance effect on rating accuracy appears not to be due to the likableness of high and low performers as suggested by Landy and Farr (1980) but may be due to a bias to recall positive information.

The consistently observed accuracy advantage of high performers does have applied implications. Increasing ratee motivation by linking performance with outcomes such as pay and identifying training needs are two of the many purposes served by performance appraisal (Cascio, 1978). The relatively inaccurate ratings of low performers indicates that low performers are least likely to have their performance directly linked with outcomes. But low performers are precisely those ratees for whom motivation should be of utmost concern. In addition, the inaccuracy of rating low performers also indicates that appraisal based training given to low performers will be less likely to meet specific needs than the training given to high performers. In sum, the relatively inaccurate ratings of low performers indicates less than maximal human resource management.

While found not to influence the effect of liking, the selective attention ability of raters and the memory demand imposed by the rating task were both found to significantly influence rating accuracy. Subjects who were better able to find figures embedded within a configuration were found to be more accurate raters than subjects who were less able to find figures embedded within a configuration. Thus, raters who have a feature analytic style discriminate more accurately among ratees than raters who have a wholistic perceptual style. Given the feature analytic nature of the performance rating task, the identification of performance on each dimension within and across ratees, those raters whose perceptual style more closely match the type of information processing demanded by the task are more accurate raters.

While the direct applied implications of this finding would be to either select high selective attention ability persons for the position of rater or train raters to include a high level of selective attention ability, neither of these possibilities is practically viable. The role of rater is typically carried out by the supervisor (Heneman, Schwab, Fossum, & Dyer, 1980), and it is doubtful that selective attention ability would ever be used as a supervisory level selection criterion. The trainability of a perceptual style is also questionable (Witkin, Oltman, Raskin, & Karp, 1971). However, as suggested by Landy and Farr (1980), consideration of cognitive processes may lead to effective changes in rating formats. In terms of selective attention, it may be that a rating format which signly focuses attention on each rating dimension would increase the feature analytic processing of raters. For example, a rating format which has only one dimension per page may increase the accuracy of ratings in this fashion.

The effect of the memory manipulation on rating accuracy indicates that ratings are more accurate when the rater can rely on a physical record of ratee performance incidents than when the rater must rely on recall for ratee performance incidents. While this finding is not surprising, the present study is the first empirical demonstration of the effect of recall on the accuracy of performance ratings. The straightforward applied implication of this finding is that performance ratings would be more accurate, and thus have more utility, if they were based on a representative record of ratee performance incidents rather than a recall of performance exhibited by ratees. That performance appraisals typically are meant to represent performance exhibited across a fairly long period of time is indicated by a recent survey of organizations (Bureau of National Affairs, 1975) that indicated the majority of appraisals are made on an annual basis by the ratee's immediate supervisor. If the supervisor does not maintain a record of ratee performance incidents across this time period, the typical appraisal situation asks the supervisors to evaluate subordinate performance based on incidents which may have been observed months or even a year ago. In contrast, student raters in the recall condition of the present investigation were allowed to evaluate ratees immediately following study of the vignettes. Even in this situation where the time since ratee observation was negligible, the ratings based on recall were significantly less accurate than those made without a recall requirement. In the typical appraisal situation, the length of time over which performance incidents are expected to be recalled would only be expected to accentuate the inaccuracy of ratings made when a physical record of ratee performance incidents is not maintained.

The results of the memory manipulation suggest that a record of ratee performance incidents should be kept and used as the basis for appraisal. While it may be thought that this suggestion is overzealous and would be impractical in applied situations, the results of the only published study involving the recording of performance incidents by non-student raters (Flanagan & Burns, 1957) indicates this is not the case. General Motors foremen were found to have positive reactions toward the process of daily recording of observation of ratee behavior and reported the procedure as taking less than five minutes a day. Thus the payoff in increased accuracy and utility of appraisals may very likely more than offset the costs associated with the procedure of recording ratee performance incidents. In addition, based on the Flanagan and Burns findings, implementation and maintenance of the recording procedure may be less difficult than might be imagined.

The present investigation involved students rating written vignettes. Thus, the applied value of the investigation rests upon the generalizability of the findings to non-student raters and "real people" ratees. The generalizability of performance appraisal findings across different populations of raters is questionable. For example, Bernardin, Cardy, and Abbott (note 8) found different patterns of findings across different samples (Psychology students, Business students, and ASPA members) of raters. In addition, the generalizability of findings with "paper people" has been seriously questioned by Gorman, Clover, and Doherty (1978).

The generalizability issue is an empirical question which calls for field replication of the present findings. The applied importance of the present findings should, however, not be prematurely discounted due to questionable generalizability of laboratory findings. Dipboye and Flanagan (1979; 1981) have argued in the support of the generalizability of laboratory research. In a critical discussion of the generalizability of laboratory findings, Wendelken and Inn (1981) point out that process generality can provide an important source of external validity. In the present context, if cognitive processing characteristics, such as schematic and recall bias processes, were responsible for the effects then similar effects would be expected in different contexts, such as non-student raters and "real people" ratees. Thus, while the representativeness of the laboratory studies may be questionable, the cognitive operations underlying the effects are not only theoretically important but provide for the relevance of the findings for applied settings.

REFERENCE NOTES

1. Kehoe, J.F., Cardy, R.L., and Dobbins, G.H. Testing models of preferential choice with measures of choice time. Paper presented at the meeting of the American Psychological Association, Montreal, September 1980.

2. Cardy, R.L., Dobbins, G.H., and Kehoe, J.F. The effects of affect and cognition on choice time. Paper presented at the meeting of the Virginia Psychological Association, Richmond, April 1981.

3. Monahan, J.S. Letter discriminability and feature salience. Paper presented at the symposium: Recent advances in the psychophysics of image evaluation, Rochester, N.Y., October 1977.

4. Lord, R.G. Heuristic social information processing and its implications for behavioral measurement: An Example based on leadership categorization. Paper presented at the meeting of the American Psychological Association, Los Angeles, August 1981.

5. Crocker, J., Weber, R., and Binns, D. Confirming and disconfirming information in stereotyping. Paper presented at the meeting of the American Psychological Association, Los Angeles, August 1981.

6. Murphy, K.R., and Balzer, W.K. Rater errors and rating accuracy. Manuscript submitted for publication, 1982.

7. Downey, R.G. and Saal, F.E. Evaluating human judgment techniques. Paper presented at the meeting of the American Psychological Association, Toronto, August 1978.

8. Sauser, W.I., Evans, K.L., and Champion, L.C.H. Two-Hundred and fifty scaled incidents of college classroom teaching behavior. Paper presented at the meeting

9. Bernardin, H.J., Cardy, R.L., and Abbott, J.C. The effects of individual performance schemata, familiarization with the rating scales and rater motivation on rating effectiveness. Paper accepted for presentation at the meeting of the Academy of Management, New York, August 1982.

10. Hamilton, D.L. Stereotype activation and social information processing. Paper presented at the meeting of the American Psychological Association, Los Angeles, August 1981.

11. Monahan, J.S., Ingate, M., Diamond, N.C., and Kaniper, R. Holistic processing of letters: *The effects of orientation and name designation*. Unpublished manuscript, Central Michigan University, 1979.

12. Fiske, S.T. and Beattie, A.E. *Affect and stereotypic information: Schema-triggered affect in the initiation of close relationships*. Paper presented at the meeting of the American Psychological Association, Los Angeles, August 1981.

13. Dobbins, G.H. *The effect of leader performance and leader likableness upon ratings of leader behavior*. Master's thesis, Virginia Polytechnic Institute and State University, 1981.

REFERENCES

Anderson, N.H. Likableness ratings of 555 personality-trait words. Journal of Personality and Social Psychology, 1968, 9, 272-279.

Ashmore, R.D. and Del Boca, F.K. Sex stereotypes and implicit personality theory: Toward a cognitive-social psychological conceptualization. Sex Roles, 1979, 5, 219-248.

Bernardin, H.J., and Pence, E.C. Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 1980, 65, 60-66.

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

Bigoness, W.J. Effects of applicant's sex, race, and performance on employer's performance ratings: Some additional findings. Journal of Applied Psychology, 1976, 61, 80-84.

Borman, W.C. Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W.C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.

Borman, W.C. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 1979, 64, 410-421.

Brown, E.M. Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 1968, 52, 195-199.

Baron, R.A. and Byrne, D. Exploring Social Psychology. Boston: Allyn and Bacon, Inc., 1979.

Bureau of National Affairs, Employee performance: Evaluation and control. Personnel Policies Forum, Survey 108, 1975.

Cantor, N.E., and Mischel, W. Traits as prototypes: Effects on recognition memory. Journal of Personality and Social Psychology, 1977, 35, 38-48.

Cascio, W.F. Applied psychology in personnel management. Reston, VA: Prentice-Hall, 1978.

Cash, T.F., Gillen, B., and Burns, D.S. Sexism and "beautyism" in personnel consultant decision making. Journal of Applied Psychology, 1977, 62, 301-310.

Cohen, C. Person categories and social perception: Testing some boundaries of the processing effects of prior knowledge. Journal of Personality and Social Psychology, 1981, 40, 441-452.

Cooper, W.H. Ubiquitous halo: Sources, solutions, and a paradox. Psychological Bulletin, 1981.

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

DeCotiss, T., and Petit, A. The performance appraisal process: A model and some testable propositions. Academy of Management Review, 1978, 21, 635-646.

DeNisi, A.S., and Stevens, G.E. Profiles of performance, performance evaluations, and personnel decisions. Academy of Management Journal, 1981, 24, 592-602.

Dipboye, R.L., and Flanagan, M.F. Research settings in industrial-organizational psychology: Are findings in the field more generalizable than in the laboratory? American Psychologist, 1979, 34, 141-150.

Dukes, W.F., and Bevan, W., Jr. Accentuation and response variability in the perception of personally relevant objects. Journal of Personality, 1952, 20, 457-465.

Erdelyi, M.H. A new look at the New Look: Perceptual defense and vigilance. Psychological Review, 1974, 81, 1-25.

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

Flanagan, J.C., and Burns, R.K. The employee performance record: A new appraisal and development tool. Harvard Business Review, 1957, September-October, 95-102.

Flanagan, M.F., and Dipboye, R.L. Research settings in industrial and organizational psychology: Facts, fallacies, and the future. Personnel Psychology, 1981, 34, 37-47.

Goldstein, K.M., and Blackman, S. Cognitive style. New York: John Wiley and Sons, 1978.

Gordon, M.E. The effect of the correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance*, 1970, *5*, 366-377.

Gordon, M.E. An examination of the relationship between the accuracy and favorability of ratings. *Journal of Applied Psychology*, 1972, *56*, 49-53.

Gorman, C.D., Clover, W.H., and Doherty, M. Can we learn anything about interviewing real people from "interviews of paper people"? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance*, 1978, *22*, 165-192.

Gruenfeld, L., and Arbuthnot, J. Field independence as a conceptual framework for prediction of variability in ratings of others. *Perceptual and Motor Skills*, 1969, *28*, 31-44.

Hastie, R. Schematic principles in human memory. In E.T. Higgins, C.P. Erman, and M.P. Zanna (Eds.), *Social cognition: The Ontario symposium on personality and Social Psychology*, 1981, *37*, 25-38.

Hastie, R., and Kumar, P.A. Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 1979, *37*, 25-38.

Higgins, E.T., Rholes, W.S., and Jones, C.R. Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 1977, *13*, 141-154.

Hoffman, C., Mischel, W., and Mazze, K. The role of purpose in the organization of information about behavior: Trait-based versus goal based categories in person cognition. *Journal of Personality and Social Psychology*, 1981, *40*, 211-225.

Howard, J.W., and Rothbart, M. Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, 1980, *38*, 301-310.

Jackson, D.N., Messick, S., and Myers, C.T. Evaluation of group and individual forms of embedded-figures measures of field-independence. *Educational and Psychological Measurement*, 1964, *24*, 177-192.

Jacobs, R., Kafry, D., and Zedeck, S. Expectations of behaviorally anchored rating scales. *Personnel Psychology*, 1980, *33*, 595-640.

Kaplan, M.F. Stimulus inconsistency and response dispositions in forming judgments of other persons. Journal of Personality and Social Psychology, 1973, 25, 58-64.

Keppel, G. Design and Analysis: A Researcher's Handbook. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Koltuv, B.B. Some characteristics of intrajudge trait intercorrelations. Psychological Monographs, 1962, 76, (Whole No. 552).

Landy, F.J., and Farr, J.L. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Landy, D., Sigall, H. Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. Journal of Personality and Social Psychology, 1974, 29, 299-304.

Linton, M., and Gallo, P.S., Jr. The practical statistician: Simplified handbook of statistics. Monterey: Brooks/Cole, 1975.

Love, K.G. Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. Journal of Applied Psychology, 1981, 66, 451-547.

Matlin, M.W., and Stang, D.J. The polyanna principle: Selectivity in language, memory, and thought. Cambridge: Schenkmann, 1978.

Maxwell, S.E., Camp, C.J., and Arvey, R.D. Measures of strength of association: A comparative examination. Journal of Applied Psychology, 1981, 66, 525-534.

Minard, J.G., and Mooney, W. Psychological differentiation and perceptual defense: Studies of the separation of perception from emotion. Journal of Abnormal Psychology, 1969, 74, 131-139.

Monahan, J.S., and Lockhead, G.R. Identification of integral stimuli. Journal of Experimental Psychology: General, 1977, 106, 94-110.

Moreland, R.L., and Zajonc, R.B. Is stimulus recognition a necessary condition for the occurrence of exposure effects? Journal of Personality and Social Psychology, 1977, 35, 191-199.

Neisser, V. Cognitive Psychology. Englewood Cliffs, NJ: Prentice-Hall, 1967.

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

Rosen, B., and Jerdee, T.H. The psychological basis for sex-role stereotypes: A note on Terborg and Ilgen's conclusions. *Organizational Behavior and Human Performance*, 1975, *14*, 151-153.

Rush, M.C., Phillips, J.S., and Lord, R.G. Effects of a temporal delay in rating on leader behavior descriptons: A laboratory investigation. *Journal of Applied Psychology*, 1981, *66*, 442-450.

Saal, F.E., Downey, R.G., and Lahey, M.A. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 1980, *88*, 413-428.

Saugstad, P., and Schioldborg, P. Value and size perception, *Scandinavian Journal of Psychology*, 1966, *7*, 102-114.

Schwab, D.P., Heneman, H., and DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 1975, *28*, 549-562.

Shiffrin, R.M., and Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 1977, *84*, 127-190.

Srull, T.K., and Wyer, R.S., Jr. The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Applied Psychology*, 1981, *66*, 149-158.

Thayer, S., and Schiff, W. Eye-contact, facial expression and the experience of time. *Journal of Social Psychology*, 1975, *95*, 117-124.

Wendelken, D.J., and Inn, A. Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology*, 1981, *66*, 149-158.

Wilson, W.R.K. and Zajonc, R.B. Affective discrimination of stimuli that cannot be recognized. *Science*, 1980, *207*, 557-558.

Witkin, H.A., Oltman, P.K., Raskin, E., and Karp, S.A. Manual for Embedded Figures Test, Children's Embedded Figures Test, and Group Embedded Figures Test. Palo Alto, Calif.: Consulting Psychologists Press, 1971.

Woodworth, R.S., and Schlosberg, H. Experimental Psychology (revised ed.). New York: Holt, Rhinehart and Winston, 1963.

Zajonc, R.B. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology Monograph, 1968, 9.

Zajonc, R.B. Feeling and thinking: Preferences need no inferences. American Psychologist, 1980, 35, 151-175.

APPENDIX A

Instructor Vignette Incidents

Ratee	Incident	M	Q	P*
	Dimension A: Relationships with Students Low Performers			
Morgan	This professor saw students in his off-campus office only.	3.6	2.7	99
Morgan	This professor leaves promptly after giving his lecture.	4.1	3.0	72
Morgan	This professor required his students to visit him at least once in his office to discuss the course.	8.4	2.7	95
Pierce	This professor refused to help students outside of class because of his "demanding" schedule.	1.4	1.2	95
Pierce	This professor was seldom available during his posted office hours.	2.1	1.6	94
Pierce	This professor sets aside only one hour per week for office hours.	2.3	2.0	97
Witkin	This professor stands in the hallway before and after class so that students can ask him questions in an informal atmosphere.	9.4	2.7	98
Witkin	This professor announced his office hours so that students could see him if they needed to.	9.6	2.6	97
Witkin	This professor tried to learn all his students' names.	9.9	2.2	99
Ritter	This professor criticized his class for the way they evaluated him.	1.4	1.3	98
Ritter	This professor scheduled his office hours to conflict with his own class and expressed displeasure in phone calls and afternoon appointments.	1.6	1.2	97
Ritter	This professor tells students not to come to his office unless they have a conflict with the final exam.	1.6	1.3	96

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension A: Relationships with Students High Performers			
Morgan	This professor saw students in his off-campus office only.	3.6	2.7	99
Morgan	This professor leaves promptly after giving his lecture.	4.1	3.0	72
Morgan	This professor required his students to visit him at least once in his office to discuss the course.	8.4	2.7	95
Pierce	This professor counseled students regarding their careers and the job market.	9.7	2.0	97
Pierce	This professor passed out a mimeographed sheet giving his office hours and office telephone number.	9.8	2.1	99
Pierce	This professor made it a point to know every student's name.	10.6	1.5	100
Witkin	This professor posted office hours but made students wait until he could find time to see them.	2.8	2.2	97
Witkin	This professor gives students his office number but does not make them feel welcome.	2.9	2.0	100
Witkin	This professor will see students in his office only if they make appointments.	5.7	1.9	99
Ritter	This professor will see students in his office only if they make appointments.	5.7	1.9	99
Ritter	This professor compensates for limited office hours by offering his time before and after class every day.	9.5	2.9	99
Ritter	When a student was obviously having problems in the course, this professor suggested that they set up an appointment for some extra help.	10.3	1.5	96

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension B: Ability to Present the Material Low Performers			
Morgan	This professor explained things as though he were talking to a class of PhDs.	1.6	1.3	94
Morgan	This professor would often leave out steps when working problems on the board and was unable to tell the students how he reached the solutions.	1.7	1.7	92
Morgan	This professor uses long, involved examples which confuse the class.	2.2	1.6	96
Pierce	This professor uses handouts and the overhead projector to present material.	8.7	2.9	97
Pierce	This professor presents information in brief, easy-to-follow written outline form.	9.0	2.3	96
Pierce	This professor speaks distinctively and uses good grammar.	9.8	2.2	97
Witkin	The information in this professor's lectures conflicted badly with the information in the book, resulting in total confusion.	1.4	1.2	90
Witkin	This professor's lectures are boring and unorganized.	1.8	1.4	92
Witkin	This professor lectured in a very disorganized manner, jumping from topic to topic with no apparent connection.	2.0	1.6	97
Ritter	This professor gave details about the material but never elaborated beyond them.	5.4	2.2	69
Ritter	This professor covered material in class that had already been presented in lab.	6.1	2.4	86
Ritter	This professor always kept his classroom presentations specific and to the point.	7.6	2.9	96

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension B: Ability to Present the Material High Performers			
Morgan	This professor states the objective of each lecture and presents the material in a logically ordered sequence.	9.9	1.7	100
Morgan	This professor gave notes in a very well organized outline form.	10.1	1.9	100
Morgan	This professor used good teaching aids, was articulate, and stressed important points in class.	10.3	1.4	81
Pierce	This professor lectures way above the heads of his students.	1.8	1.6	92
Pierce	This professor constantly interrupted his lectures to rummage through his briefcase for missing papers.	2.3	1.4	94
Pierce	This professor rattled off studies, definitions, and concepts but never tied them together.	2.3	1.9	92
Witkin	This professor always kept his classroom presentations specific and to the point.	7.6	2.9	96
Witkin	This professor uses gestures and theatrical movements when lecturing to keep the students' attention.	8.3	3.0	85
Witkin	This professor passes around books and pictures relating to the class material.	8.9	2.5	66
Ritter	This professor gave details about the material but never elaborated beyond them.	5.4	2.2	69
Ritter	This professor covered material in class that had already been presented in lab.	6.1	2.4	86
Ritter	This professor always kept his classroom presentations specific and to the point.	7.6	2.9	96

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension C: Interest in Course and Material Low Performers			
Morgan	This professor described his own fascination with the material he was covering.	9.3	2.7	89
Morgan	This professor brought in a current article about the course material approximately every two weeks.	9.5	1.9	88
Morgan	This professor brought in up-to-date material and gave the students interesting tid-bits related to the subject.	9.8	1.8	80
Pierce	This professor comes to class and says, "Well, here we are so I might as well lecture on something."	3.3	2.7	80
Pierce	This professor would sometimes get so involved in the subject matter that he would forget to stop lecturing when the class period was over.	6.2	3.0	68
Pierce	This professor provides time during class to talk about current issues.	8.8	2.3	64
Witkin	This professor told his students that he was totally disinterested in teaching and felt it was a waste of his time.	1.1	0.6	89
Witkin	This professor said he was teaching just to earn a paycheck.	1.3	1.2	89
Witkin	This professor told his students that he did not like teaching the class.	1.3	1.2	88
Ritter	This professor has to refer to his notes before answering any questions from students.	2.2	1.8	67
Ritter	This professor was never on time for class.	3.3	2.4	68
Ritter	This professor seldom adds anything to his lectures.	3.6	2.1	82

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension C: Interest in Course and Material High Performance			
Morgan	This professor belittled the class material and described the course as a waste of time.	1.2	0.9	90
Morgan	This professor actually tells the class that he hates the subject matter.	1.3	1.2	95
Morgan	This professor said he was teaching just to earn a paycheck.	1.3	1.2	89
Pierce	This professor comes to class and says, "Well, here we are so I might as well lecture on something."	3.3	2.7	80
Pierce	This professor would sometimes get so involved in the subject matter that he would forget to stop lecturing when the class period was over.	6.2	3.0	68
Pierce	This professor provides time during class to talk about current issues.	8.8	2.3	64
Witkin	This professor travels in order to see and hear things about his profession which he then shares with his students.	10.2	1.7	93
Witkin	This professor, when confounded by a student's question, spent several hours of his own time that same afternoon researching material for an answer.	10.5	1.4	87
Witkin	This professor gets excited about what he is teaching and conveys this enthusiasm to his students.	10.6	1.2	77
Ritter	This professor brought in up-to-date material and gave the students interesting tid-bits related to the subject.	9.8	1.8	80
Ritter	This professor has visited the places and done the things he talks about in class and describes his personal experiences to the students.	10.2	1.8	75
Ritter	Whenever this professor did not know the answer to a student's question, he would look it up and bring it in the next day.	10.4	1.7	75

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension D: Reasonableness of the Workload Low Performers			
Morgan	This professor would not assign work for several days, then would give a heavy assignment for a single night.	2.4	2.2	100
Morgan	This professor requires a typewritten lab report every week in addition to the regular course work.	3.4	2.5	99
Morgan	This professor gives more notes in one hour than most do in two.	3.4	2.9	66
Pierce	This professor sometimes assigned two chapters for one night's assignment.	2.9	1.9	99
Pierce	This professor assigns two chapters and one or two stories to read and summarize each week in addition to class exercises.	3.9	2.8	96
Pierce	This professor assigns ten homework problems per week.	6.1	1.1	96
Witkin	This professor gave an extremely heavy assignment one week, then slacked off for a week or so before giving another assignment.	5.2	2.1	96
Witkin	This professor assigned about 50 pages of reading per week.	5.9	2.6	99
Witkin	This professor would adjust the homework assignments to suit the wishes of the class.	6.4	2.8	85
Ritter	This professor assigned a four-to-five page typewritten paper and specified the format and style in which it was to be written.	7.2	2.9	93
Ritter	This professor assigns no more than two chapters of reading per week.	7.5	3.0	100
Ritter	This professor gave rest periods each week in which no homework was assigned.	9.0	2.5	96

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension D: Reasonableness of the Workload High Performers			
Morgan	This professor gave daily reading assignments and an outline of references to use during the quarter.	7.3	2.5	76
Morgan	This professor assigns either one chapter or two essays (never both) to be reach each week.	7.9	3.0	99
Morgan	This professor assigns homework a few times a week but not every day.	8.2	3.0	99
Pierce	This professor makes optional outside reading assignments.	7.5	2.7	74
Pierce	This professor gives short reading assignments.	8.4	2.8	98
Pierce	This professor assigned reasonable amounts of homework every other day.	9.0	2.3	99
Witkin	This professor gave an extremely heavy assignment one week, then slacked off for a week or so before giving another assignment.	5.2	2.1	96
Witkin	This professor assigned about 50 pages of reading per week.	5.9	2.6	99
Witkin	This professor would adjust the homework assignments to suit the wishes of the class.	6.4	2.8	85
Ritter	This professor assigns more homework for a three-hour course than most do for a five-hour course.	1.6	1.4	99
Ritter	This professor assigned a lot of reserve library reading without enough time for all students to see the material.	1.9	1.3	93
Ritter	This professor requires a lot of memorization for his class.	4.2	2.5	93

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension E: Fairness of Testing and Grading Low Performers			
Morgan	This professor tells his students to study one thing, then tests on something else.	1.2	0.9	99
Morgan	This professor never stated his grading procedures.	1.4	1.1	97
Morgan	This professor's tests are ambiguous and much too long.	2.2	1.7	95
Pierce	This professor's multiple-choice items usually included more than one possible correct answer.	2.3	2.8	99
Pierce	This professor tested over material he did not cover.	2.4	2.7	96
Pierce	This professor asks picky test questions about details.	3.3	2.1	97
Witkin	This professor changed grading procedures in the middle of the quarter.	2.3	2.7	97
Witkin	This professor assigned general problems in class, then gave specific problems on tests.	3.6	3.0	92
Witkin	This professor marks off for poor class attendance.	5.6	2.5	73
Ritter	This professor refused to scale the test grades even when the entire class did poorly.	2.1	2.4	98
Ritter	This professor does not curve grades even if the average score is in the 50's or 60's.	2.1	2.6	99
Ritter	This professor gives hard tests which require the students to study a lot.	5.8	1.8	72

APPENDIX A (Continued)

Ratee	Incident	M	Q	P*
	Dimension E: Fairness of Testing and Grading High Performers			
Morgan	This professor told his students how much each test and project was worth toward the final grade.	10.3	1.9	91
Morgan	This professor gives his students enough time to complete his tests.	10.5	1.4	96
Morgan	This professor's test questions are to the point and are easy to understand.	10.6	1.2	93
Pierce	This professor gave tests weekly and allowed students to drop their two lowest grades.	9.8	2.2	95
Pierce	This professor's tests covered only what he told his students would be on them.	10.0	2.0	93
Pierce	This professor pointed out the types of problems he would include on each test and held to his word.	10.2	1.5	93
Witkin	This professor's tests have a variety of item formats including multiple-choice, fill-in-the blank, essay, and true-false questions.	8.3	2.9	99
Witkin	This professor's tests have a lot of questions so that you can miss one and not worry about failing.	8.9	2.4	98
Witkin	This professor told how many points each essay question was worth at the beginning of each of his tests.	9.2	2.5	99
Ritter	This professor allows students to question his grading after the tests are handed back.	9.7	1.8	90
Ritter	This professor outlines the type of questions to be included on tests, along with the credit values for each type of question.	10.0	1.8	90
Ritter	This professor dropped the lowest two of six equally-weighted questions.	10.2	2.8	96

APPENDIX A (Continued)

*All statistics calculated with $\underline{n} = 100$

M = Median Rating
Q = Semi-interquartile range of ratings given to the incident.
P = Percentage of subjects placing the incident in the modal category.

APPENDIX B

Rating Form

Professor Name J. Morgan
ID Number ______________

Dimension 1: Relationships with Students

1	2	3	4	5	6	7	8	9	10	11
	criticizes students for asking questions in class				will see students in office only if they make appointments				very helpful and supportive freely offers assistance	

Dimension 2: Ability to Present the Material

1	2	3	4	5	6	7	8	9	10	11
	confused and disjointed presentation			relies on notes and gives no elaboration on them					clear, concise, and organized presentation; used variety of presentation methods	

Dimension 3: Interest in Course and Material

1	2	3	4	5	6	7	8	9	10	11
	claims disinterest and dislike of course and material				keeps up with latest developments but doesn't include them in lectures				knows material so well that he is able to answer all questions	

Dimension 4: Reasonableness of the Workload

1	2	3	4	5	6	7	8	9	10	11
	workload so heavy that few students pass				nightly homework assignments			evenly distributes workload across quarter		

Dimension 5: Fairness of Testing and Grading

1	2	3	4	5	6	7	8	9	10	11
	refuses to change grades even if a mistake had been made				reasonable but sometimes tricky test questions doesn't curve grades unless class does extremely badly				test questions concise and easily understood; changes grades if shown mistake	

APPENDIX C

Materials Sample for the Implications Pretest

Directions

Knowing a particular characteristic of a person can sometimes imply to us other things about that person. For example, knowing that a person is intelligent would typically imply something about that person's SAT scores, namely, high scores would be expected. In other cases knowing some characteristic of a person does not at all imply behavior on a dimension. For example, knowing that someone is affectionate probably does not tell you anything about that person's SAT scores.

Your task in this experiment is to rate the degree to which a personal characteristic tells you something about the person's behavior. You will be given a list of 50 words, each one describing a personal characteristic. For each word you are to rate the extent to which knowing that characteristic of a college professor tells you something about that professor's behavior on each one of 5 dimensions. That is, does knowing that a teacher has a particular characteristic or quality tell you what kind of behavior the teacher would exhibit?

In this set of materials you will find five pages, one page for each of five dimensions of teacher behavior. Also, there are 5 computer op-scan sheets, one for each of the five pages. Please be sure that you have the op-scan sheet that goes with the page and dimension you're rating. The dimension for each op-scan sheet is written on the top of the op-scan sheet.

The rating scale to use for all of your ratings is the following 10 point scale.

1 2 3 4 5 6 7 8 9 10

I am unsure about the behavior that would be exhibited by a teacher with this characteristic

I am somewhat sure about the behavior that would be exhibited by a teacher with this characteristic

I am very sure about the behavior that would be exhibited by a teacher with this characteristic

Your rating should be based on the answer to the question, "Given the characteristic being considered, how sure am I about the kind of behavior such a teacher would exhibit?". If you are unsure which of the listed behaviors the teacher would exhibit, then your rating should be a low number such as "1" or "2". On the other hand, your rating should be in the "sure" direction of the scale, a high number, to the extent that one kind of behavior seems to you more likely than the others given the particular characteristic being considered.

Each of the following pages describes a particular dimension and 3 types of behavior for that dimension and also repeats the rating scale. For each of the 50 characteristics listed, use the 10 point rating scale to rate how sure you are about the behavior that would be exhibited by a teacher with that characteristic.

Remember, if you are unsure about which behavior to expect use a <u>low</u> number. If you are <u>sure</u> about which behavior to expect use a <u>high</u> number.

If you have any questions please ask the experimenter now.

Dimension 1: Relationships with students

low level	medium level	high level
criticizes students for asking questions in class	will see students in office only if they make appointments	very helpful and supportive; freely offers assistance

Please rate each characteristic on the following scale.

1 2 3 4 5 6 7 8 9 10

I am unsure about the behavior that would be exhibited by a teacher with this characteristic	I am somewhat sure about the behavior that would be exhibited by a teacher with this characteristic	I am very sure about the behavior that would be exhibited by a teacher with this characteristic

REMEMBER TO RATE EACH CHARACTERISTIC

1. ungracious
2. sympathetic
3. unreasonable
4. gentle
5. deliberate
6. unlucky
7. irrational
8. original
9. outstanding
10. self-centered
11. messy
12. aggressive
13. unpleasing
14. intellectual
15. bashful
16. cheerful
17. uninteresting
18. experienced
19. superficial
20. self-concerned
21. witty
22. greedy
23. tolerant
24. finicky
25. clean-cut
26. observant
27. incompetent
28. hypochondriac
29. heartless
30. snobbish
31. inoffensive
32. bold
33. unkind
34. self-reliant
35. appreciative
36. unethical
37. sharpwitted
38. crude
39. kind-hearted
40. offensive
41. amusing
42. cold
43. sportsmanlike
44. unreliable
45. painstaking
46. irritable
47. cooperative
48. shallow
49. truthful
50. level-headed

APPENDIX D

Mean Implications of 300 Trait Terms for Performance on the Five Teacher Behavior Dimensions

		Dimension				
Term	Trait	A	B	C	D	E
(231)	Reserved	4.7	3.7	3.3	2.1	2.3
(63)	Good-Tempered	6.6	3.4	3.6	4.4	5.4
(442)	Unruly	5.8	3.5	1.5	3.1	2.0
(112)	Inventive	2.0	6.2	5.1	4.0	2.6
(549)	Obnoxious	6.8	3.3	2.2	3.8	6.4
(18)	Kind	6.4	2.7	1.8	6.0	4.56
(105)	Energetic	4.9	5.7	5.9	5.6	3.2
(443)	Fault-Finding	5.9	3.5	3.2	5.7	6.6
(110)	Independent	2.4	4.3	4.6	3.5	4.2
(84)	Courageous*	1.9	2.3	3.3	2.4	2.1
(246)	Perfectionist	3.9	7.3	7.1	6.2	6.6
(117)	Attentive	5.2	5.8	5.0	3.8	4.8
(448)	Anti-Social	5.3	2.44	2.1	3.2	2.4
(61)	Respectful	5.0	2.7	2.7	4.7	6.6
(107)	Self-Contolled	3.5	4.5	2.6	3.1	3.0
(277)	Impressionable*	2.8	4.0	2.9	4.2	4.0
(9)	Open-Minded	5.1	5.9	5.8	4.9	7.4
(453)	Foolish	1.6	3.7	3.0	3.6	2.8
(456)	Negligent	5.4	5.9	6.2	4.5	4.7
(44)	Tactful*	3.6	4.2	4.0	3.3	2.8
(120)	Purposeful	4.3	6.5	6.0	5.6	4.8
(71)	Kindly	6.1	2.0	1.7	4.4	4.9
(70)	Clearheaded	3.8	6.8	5.3	5.9	4.8
(500)	Prejudiced	6.4	1.3	2.2	2.9	5.1
(65)	Conscientious	5.5	6.3	5.6	6.5	6.9
(525)	Ultra-Critical	7.2	3.2	3.4	4.9	6.7
(522)	Unfriendly	6.1	2.8	2.9	4.7	5.5
(454)	Troublesome	4.9	2.6	3.5	4.3	4.0
(28)	Responsible	5.7	7.1	5.7	6.5	7.2
(86)	Productive	4.2	6.4	6.7	6.3	5.4
(3)	Understanding	5.8	4.1	3.3	7.4	6.5
(258)	Unpredictable	4.4	5.1	4.2	4.9	4.1
(4)	Loyal*	3.3	2.1	2.6	3.2	3.2
(92)	Neat	1.4	4.6	2.5	2.4	2.1
(296)	Lonely*	3.5	1.4	1.9	3.2	2.0
(541)	Spiteful	5.5	2.9	1.6	6.3	6.5
(528)	Underhanded*	3.5	2.3	2.2	3.5	4.4
(75)	Perceptive	4.1	5.7	5.9	5.3	4.7
(109)	Active	4.4	4.8	5.6	4.3	3.1
(115)	Cordial	5.8	2.9	2.4	2.4	3.6
(452)	Careless	4.7	6.0	4.2	3.6	4.3

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(462) Touchy	4.9	2.8	2.6	2.3	5.2
(68) Good	5.3	4.3	5.8	5.4	5.6
(95) Prompt	3.8	4.9	4.1	3.8	2.2
(259) Solemn*	3.5	2.4	2.4	3.6	2.9
(469) Lazy	5.5	6.4	6.3	5.9	3.7
(113) Wholesome*	2.5	2.2	1.5	3.6	1.9
(288) Daydreamer*	3.4	3.9	3.0	2.7	1.8
(523) Hostile	6.6	2.2	2.4	5.2	5.8
(10) Thoughtful	7.2	3.9	4.7	7.1	5.9
(276) Unsophisticated	3.8	5.0	3.8	2.1	3.4
(83) Capable	4.1	6.9	6.6	4.3	5.1
(464) Tiresome*	1.7	3.4	2.6	2.2	2.4
(264) Emotional*	2.7	3.0	2.4	3.4	3.6
(499) Irresponsible	4.7	5.9	4.6	3.6	5.0
(546) Deceitful	4.9	1.8	2.0	4.0	6.3
(280) Skeptical	3.3	3.9	3.9	2.1	5.3
(513) Unforgiving	5.3	2.3	1.1	4.0	5.4
(491) Discourteous	3.9	2.4	2.3	4.6	4.4
(248) Excitable*	2.1	3.3	2.4	2.5	2.9
(252) Impulsive	2.9	4.1	3.2	5.0	4.5
(445) Misfit*	2.0	3.0	1.8	1.3	1.9
(255) Conservative	3.8	4.0	5.0	5.4	5.0
(23) Interesting	4.2	6.8	5.6	3.7	3.2
(519) Unkindly	4.9	3.1	2.8	5.5	4.4
(503) Unpleasant	4.7	3.1	1.2	3.9	3.5
(43) Quick-Witted	4.2	5.6	5.7	4.0	4.4
(87) Progressive	3.2	5.5	6.2	5.1	5.2
(484) Meddlesome*	2.3	1.0	2.2	1.4	3.1
(482) Grouchy	4.9	2.6	2.0	4.2	4.8
(85) Constructive	5.4	6.9	6.2	5.3	5.2
(2) Honest	5.4	2.8	3.9	3.6	7.4
(3) Prudent*	3.6	2.8	2.8	3.3	4.4
(234) Unconventional	2.8	4.5	3.5	3.9	3.4
(510) Abusive	4.0	1.5	0.9	5.2	4.5
(26) Honorable	5.0	3.7	4.5	4.8	7.3
(490) Cowardly*	2.8	3.4	2.1	1.2	3.2
(547) Dishonorable	3.2	1.9	2.6	2.7	5.3
(493) Childish*	2.5	2.8	2.8	1.7	2.9
(240) Innocent*	2.2	1.3	2.6	2.0	3.6
(263) Discriminating*	4.0	2.5	1.8	3.1	5.4
(501) Bragging*	2.4	2.5	2.5	2.4	2.6
(73) Patient	7.2	4.9	5.3	5.6	5.7
(232) Persistent	5.4	5.2	5.7	6.6	3.5
(537) Thoughtless	4.7	3.9	3.9	6.2	5.3
(238) Suave*	2.7	3.8	2.8	3.0	3.9

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(275) Theatrical	3.0	4.7	3.3	1.9	1.5
(542) Insulting	4.7	2.0	1.2	2.2	1.9
(233) Meticulous	3.6	5.6	4.0	4.7	3.4
(36) Reasonable	5.0	5.0	5.0	7.5	6.8
(251) Extravagant	2.4	4.6	4.0	2.6	2.3
(481) Deceptive	3.0	2.6	2.4	3.4	5.0
(40) Clever	3.8	7.0	6.8	4.9	4.5
(287) Wordy	2.8	5.2	4.0	2.1	2.7
(103) Generous	7.0	3.8	3.8	6.6	6.3
(50) Forgiving	6.0	2.7	1.5	4.2	5.6
(478) Irritating	4.8	2.7	1.8	4.1	4.6
(498) Unfair	5.5	3.5	3.4	6.6	6.5
(485) Uncivil	4.0	3.6	2.3	5.3	4.5
(474) Dull	2.4	5.7	3.6	2.2	1.5
(487) Unsportsmanlike	3.6	1.8	1.5	4.8	4.6
(45) Helpful	7.5	6.4	7.1	6.1	7.0
(254) Changeable	2.7	4.8	4.1	4.5	6.6
(290) Materialistic*	2.7	4.4	2.1	2.1	3.0
(13) Good Natured	6.4	4.8	2.2	4.7	5.5
(30) Trustful	4.7	3.6	4.0	4.2	4.1
(294) Opinionated	4.1	2.9	4.9	3.9	5.4
(82) Versatile	4.8	7.1	5.6	4.7	5.4
(78) Well-Mannered*	4.2	2.7	2.4	2.5	3.8
(35) Educated	3.1	5.2	6.6	5.5	3.2
(459) Gloomy*	3.7	3.5	4.3	3.4	3.2
(271) Choosy	4.4	4.7	3.8	4.8	4.7
(550) Untruthful*	2.8	2.9	3.2	2.9	3.6
(512) Intolerant	6.2	3.8	3.7	4.8	5.5
(545) Untrustworthy*	1.9	2.8	3.3	2.8	3.8
(273) Naive*	1.8	2.9	2.8	3.5	3.4
(31) Warm-Hearted	5.3	2.5	1.8	4.5	6.1
(67) Alert	3.9	5.5	6.6	4.6	4.3
(243) Methodical	4.0	5.9	3.7	4.8	3.6
(262) Average	2.4	2.7	3.1	3.9	4.6
(59) Ambitious	4.7	5.5	6.0	4.9	3.2
(53) Polite	4.6	2.9	2.7	3.4	3.4
(37) Companionable	5.7	3.9	3.4	4.2	4.5
(66) Resourceful	4.2	6.8	6.7	4.7	4.1
(38) Likable	5.0	3.9	3.2	5.1	4.8
(12) Considerate	6.0	3.7	3.9	6.8	7.0
(532) Selfish	4.0	4.5	3.2	2.9	3.7
(472) Aimless	2.8	6.0	5.8	3.8	3.9
(80) Ethical	3.1	3.9	2.6	4.7	5.4
(41) Pleasant	4.6	3.1	3.3	4.8	5.9
(106) High-Spirited	5.0	5.1	5.6	4.4	3.3

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(441) Resentful	5.0	2.8	4.3	4.2	4.2
(72) Admirable	4.7	3.3	4.7	4.1	4.6
(508) Humorless	2.6	4.7	2.4	4.4	4.4
(450) Stingy	4.4	2.2	3.5	2.7	5.4
(24) Unselfish	5.0	3.8	3.7	3.5	5.4
(119) Frank	5.0	4.7	3.8	3.4	3.8
(272) Self-Possessed*	3.7	3.5	4.1	3.2	3.0
(269) Lonesome*	2.9	3.0	1.9	2.3	0.9
(298) Unsystematic	2.8	6.0	4.8	4.0	4.0
(16) Warm	6.0	3.9	1.5	4.3	5.8
(60) Bright	4.7	5.6	6.3	4.8	4.7
(91) Lively	4.8	5.3	4.8	3.6	3.2
(279) Strict	5.7	4.7	2.2	5.7	5.2
(497) Hard-Hearted	6.2	2.9	3.5	6.0	5.5
(496) Self-Conceited*	3.7	4.0	2.5	3.7	2.4
(458) Profane	2.1	2.0	2.5	2.6	2.3
(473) Boastful*	2.1	3.6	3.4	3.4	2.3
(74) Talented	4.1	5.9	5.6	2.7	3.3
(250) Prideful	3.3	4.7	4.9	2.3	4.5
(455) Ungracious	5.4	2.4	3.1	4.2	5.9
(517) Self-Centered*	4.1	1.5	2.1	2.4	4.1
(444) Messy	1.6	5.0	1.6	1.5	1.7
(253) Aggressive*	3.5	3.0	4.0	3.6	3.0
(489) Unpleasing	4.0	4.5	4.0	5.8	6.4
(81) Intellectual	2.4	4.3	4.8	2.1	3.8
(266) Bashful*	3.2	4.0	2.2	1.6	1.4
(29) Cheerful*	3.9	3.2	3.8	2.6	3.6
(446) Uninteresting	3.3	6.9	6.4	2.5	3.5
(116) Experienced	3.6	6.3	5.7	4.7	5.6
(494) Superficial*	2.5	2.7	3.3	1.9	2.2
(104) Sympathetic	6.5	2.5	2.3	6.2	6.8
(267) Self-Concerned	4.5	2.7	4.1	3.9	4.6
(69) Witty	2.2	4.5	3.1	2.5	2.2
(540) Greedy*	2.3	0.9	1.7	2.7	4.3
(100) Tolerant	6.0	3.1	3.6	5.6	4.5
(440) Finicky	3.7	5.5	4.6	4.4	6.0
(102) Clean-Cut*	0.9	1.3	1.7	1.8	0.7
(89) Observant	4.3	3.5	4.6	5.1	4.8
(492) Incompetent	5.9	6.7	7.0	3.8	4.9
(477) Hypochondriac*	1.8	1.4	2.1	1.6	1.7
(535) Heartless	6.3	2.3	2.6	8.1	7.3
(516) Unreasonable	6.5	2.6	4.6	8.5	7.9
(518) Snobbish	4.4	1.7	1.1	2.5	4.5
(241) Inoffensive*	3.2	2.1	2.4	1.9	2.9
(237) Bold*	3.6	2.9	1.4	1.8	3.3

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(544) Unkind	7.0	2.3	2.8	7.6	7.2
(99) Self-Reliant	3.2	4.7	4.6	3.4	3.5
(46) Appreciative*	2.9	2.5	3.5	4.3	3.5
(515) Unethical	4.2	2.7	2.8	5.8	5.4
(57) Sharp-Witted*	3.3	4.1	2.2	1.1	2.3
(506) Crude*	4.1	2.7	1.6	2.1	3.7
(20) Kind-Hearted	6.1	3.2	1.2	6.9	6.6
(33) Gentle*	2.9	0.7	1.5	4.2	3.5
(526) Offensive	6.2	3.0	1.7	2.7	5.6
(101) Amusing*	1.6	3.9	2.7	0.9	2.7
(486) Cold	6.2	4.0	2.2	6.1	5.2
(77) Sportsmanlike	2.7	1.8	1.8	5.1	5.5
(504) Unreliable	4.2	6.1	6.5	4.8	4.4
(236) Painstaking	4.4	5.9	5.4	4.4	5.3
(449) Irritable	6.5	3.7	2.8	5.1	4.9
(79) Cooperative	5.8	2.8	4.0	5.3	6.2
(480) Shallow*	3.6	2.5	4.1	2.6	2.4
(5) Truthful	1.7	2.6	3.2	1.5	5.9
(235) Deliberate	2.3	4.8	3.8	4.9	3.2
(52) Level-Headed	3.5	5.8	3.4	6.0	6.2
(265) Unlucky*	0.3	1.0	0.7	1.1	1.3
(463) Irrational	5.5	4.6	3.8	6.0	6.4
(54) Original	2.4	6.8	5.1	3.5	2.9
(48) Outstanding	5.4	6.6	6.4	4.7	5.3
(505) Impolite	7.1	2.3	2.6	4.4	6.6
(242) Shrewd	4.4	4.2	5.2	5.1	5.6
(488) Bossy	5.8	3.1	2.9	6.2	6.9
(527) Belligerent	5.3	2.1	3.5	4.1	6.3
(51) Enthusiastic	7.3	6.6	7.6	5.4	4.7
(257) Hesitant	2.9	4.4	5.1	4.0	4.9
(475) Gossipy*	1.4	1.4	2.0	0.4	1.4
(1) Sincere	6.3	3.2	4.6	5.2	7.0
(465) Disobedient	2.7	2.5	3.0	1.7	4.6
(90) Ingenious	3.7	6.6	6.8	4.3	4.7
(47) Imaginative	3.3	7.6	7.5	4.4	5.0
(282) Forceful	4.7	5.8	5.5	6.8	6.7
(270) Restless	4.2	3.3	4.4	2.6	3.3
(483) Egotistical	5.5	4.1	3.5	4.3	7.4
(62) Efficient	4.8	8.5	7.8	7.7	7.4
(22) Clean*	2.3	3.1	1.9	1.4	1.0
(64) Grateful*	3.7	1.1	3.6	2.0	4.4
(297) Dependent	2.7	6.5	5.2	3.1	2.6
(292) Rebellious	3.6	2.4	2.6	4.5	5.7
(21) Happy	4.0	2.8	5.1	3.2	3.9
(509) Quarrelsome	5.6	2.7	2.4	3.8	6.4
(533) Narrow-Minded	7.1	5.3	6.3	6.6	8.4

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(88) Individualistic	5.1	5.4	5.5	4.6	6.3
(502) Jealous*	3.1	0.4	1.5	1.9	3.6
(521) Ill-Tempered	6.7	3.5	2.5	5.4	7.2
(555) Liar*	2.5	1.8	3.1	1.2	2.6
(507) Nosey*	1.6	1.0	2.7	0.6	1.4
(551) Dishonest*	2.7	0.8	3.1	1.6	4.0
(461) Disagreeable	6.4	3.1	5.7	4.3	7.1
(283) Cunning*	2.5	4.0	3.6	3.6	4.3
(284) Inexperienced	3.3	6.6	5.0	4.2	3.7
(39) Trusting	5.0	1.1	2.3	2.6	5.6
(514) Boring	2.5	6.7	5.1	2.8	2.7
(32) Broad-Minded	5.8	6.4	7.2	6.0	7.2
(244) Nonchalant	4.3	5.4	5.5	3.0	4.4
(447) Scornful	6.2	2.1	3.4	4.7	5.9
(538) Rude	7.1	3.3	3.0	5.0	6.2
(543) Insincere	5.6	3.6	2.6	3.3	6.0
(249) Outspoken	4.0	5.5	5.4	4.4	5.8
(256) Shy*	2.3	3.3	2.5	0.5	1.1
(460) Helpless	2.8	5.1	4.1	1.8	3.4
(97) Sensible	5.5	7.1	5.8	7.8	7.4
(466) Complaining	6.6	3.5	6.2	4.2	5.4
(468) Vain	4.5	2.5	2.4	3.0	6.0
(19) Friendly	7.2	4.1	4.4	4.9	6.2
(108) Tender	4.9	0.8	2.4	3.1	5.0
(49) Self-Disciplined	3.3	7.4	6.9	6.9	4.3
(534) Vulgar	5.2	1.7	2.7	0.9	0.9
(286) Daredevil*	1.2	2.3	1.9	0.8	2.0
(96) Accurate	2.8	8.5	8.2	5.3	5.3
(111) Respectable	5.4	5.1	6.0	6.2	6.6
(17) Earnest	6.0	4.4	5.4	4.1	5.1
(8) Dependable	6.1	5.2	6.2	5.8	5.9
(293) Eccentric*	3.0	3.0	3.3	4.0	4.4
(268) Authoritative	4.6	5.1	4.6	5.0	5.0
(291) Self-Satisfied	4.9	4.9	3.9	2.3	3.9
(114) Congenial	4.5	4.7	4.4	3.6	3.9
(15) Mature	5.3	5.3	6.2	6.0	5.7
(529) Annoying	6.3	2.1	2.5	3.4	5.0
(536) Insolent*	3.5	2.4	3.6	2.3	2.6
(25) Good-Humored	5.1	6.0	5.4	5.0	5.2
(6) Trustworthy	4.7	4.0	4.5	5.1	6.8
(530) Disrespectful	5.7	3.2	4.2	3.7	4.0
(470) Unappreciative	5.6	2.9	3.6	4.6	4.8
(476) Unappealing*	4.4	2.9	2.5	3.6	4.1
(239) Cautious	3.6	5.8	4.2	4.9	5.4
(247) Forward	3.9	4.1	5.0	4.0	5.0
(98) Creative	4.3	8.0	5.7	4.7	4.9

APPENDIX D (Continued)

	Dimension				
Term Trait	A	B	C	D	E
(467) Lifeless	3.5	3.9	5.6	4.5	2.8
(295) Stern	4.0	3.3	4.1	5.7	5.0
(58) Well-Read	5.2	6.1	6.9	4.9	4.7
(251) Quiet*	3.1	2.2	3.8	2.3	3.3
(76) Spirited	5.8	5.8	6.0	5.0	5.8
(471) Maladjusted*	3.0	3.3	3.9	3.6	3.3
(261) Self-Righteous*	3.3	3.9	2.8	2.6	3.1
(531) Loud-Mouthed	5.3	2.9	3.7	3.4	2.8
(285) Un-Methodical	3.9	5.2	4.3	4.1	4.7
(524) Dislikable	4.3	3.9	2.8	3.1	5.5
(539) Conceited	4.8	2.8	2.8	2.9	2.7
(479) Petty*	2.9	1.6	3.2	2.7	3.3
(274) Opportunist	5.4	4.2	5.2	5.1	5.2
(34) Well-Spoken	5.6	7.3	6.5	5.5	4.4
(511) Distrustful	4.0	2.4	4.5	3.6	4.5
(42) Courteous	6.4	4.7	4.3	4.9	7.2
(553) Mean	6.1	2.1	3.7	3.5	4.5
(97) Sensible	7.3	7.4	6.7	7.5	7.0
(245) Self-Contented	5.3	3.4	4.2	4.4	4.1
(94) Logical	5.8	6.6	5.1	6.6	7.9
(520) Ill-Mannered	5.5	2.1	2.8	4.0	4.6
(11) Wise	5.6	7.6	6.4	5.7	7.5
(495) Ungrateful	3.3	2.6	2.8	3.7	4.8
(289) Conventional	4.9	4.9	5.6	5.0	5.8
(278) Ordinary*	3.6	3.1	3.5	3.9	3.0
(27) Humorous	4.3	6.3	4.6	4.6	5.0
(118) Cultured	3.5	5.3	5.5	5.4	4.4
(93) Punctual	4.7	2.2	4.1	4.4	4.5
(451) Tactless*	3.1	4.3	3.4	3.2	3.3
(260) Blunt	3.7	4.7	4.5	5.0	3.2
(554) Phony	4.7	3.7	3.8	3.6	3.7
(457) Wishy-Washy	2.3	4.0	3.3	4.5	3.4

*Mean implication rating less than scale midpoint for each of the five dimensions.

APPENDIX E

Materials for the Personal vs. Normative Pretest

Please rate each of the following words using the scale below.

1 least likable	2	3	4	5	6	7 most likable

Rate each word according to how much you, personally, would like a person being described with that word.

1. ordinary
2. petty
3. boastful
4. well-mannered
5. tactful
6. daredevil
7. suave
8. self-centered
9. meddlesome
10. profane
11. hypochondriac
12. quiet
13. underhanded
14. maladjusted
15. tactless
16. shy
17. prudent
18. sharpwitted
19. lonesome
20. loyal
21. gloomy
22. greedy
23. self-conceited
24. childish
25. nosey
26. bold
27. inoffensive
28. cunning
29. bashful
30. courageous

APPENDIX F

Means and Standard Deviations of Personal and Normative Liking Ratings

Trait Term	Mean Liking Rating		Standard Deviation	
	Personal	Popular	Personal	Popular
1. Ordinary	3.33	2.83	1.03	0.99
2. Petty	1.50	0.94	1.15	0.64
3. Boastful	1.22	0.94	1.06	0.94
4. Well-Mannered	4.78	4.94	0.94	0.87
5. Tactful	4.55	4.55	0.92	1.29
6. Daredevil	3.28	3.33	1.07	1.33
7. Suave	3.06	3.39	1.39	1.38
8. Self-Centered	0.72	0.61	1.17	0.98
9. Meddlesome	1.45	0.56	0.92	0.62
10. Profane	1.45	1.39	1.09	1.04
11. Hypochondriac	1.17	1.67	0.85	1.33
12. Quiet	3.39	3.33	0.98	0.84
13. Underhanded	1.33	0.83	1.28	0.79
14. Maladjusted	1.50	1.28	1.10	0.96
15. Tactless	1.39	1.06	1.04	0.80
16. Shy	3.56	3.17	1.10	0.71
17. Prudent	2.39	2.50	1.24	1.04
18. Sharpwitted	4.34	4.72	0.97	0.75
19. Lonesome	3.06	2.94	1.00	1.06
20. Loyal	5.28	5.11	0.96	0.90
21. Gloomy	1.17	1.33	0.92	0.77
22. Greedy	0.72	0.89	0.75	1.23
23. Self-Conceited	0.67	0.50	0.91	0.92
24. Childish	1.39	1.11	0.78	0.68
25. Nosey	1.39	1.17	1.04	1.20
26. Bold	3.78	3.28	1.06	1.13
27. Inoffensive	4.00	3.67	0.90	0.97
28. Cunning	3.28	3.28	1.41	1.13
29. Bashful	3.11	3.17	0.90	0.99
30. Courageous	4.56	4.67	0.86	0.97
31. Unlucky	2.89	3.00	0.47	1.08
32. Impressionable	3.28	3.39	0.89	1.24
33. Cheerful	5.44	5.22	0.70	0.88
34. Emotional	3.50	3.28	1.62	1.18
35. Self-Righteous	2.83	2.50	1.29	1.58
36. Unappealing	1.50	1.33	1.09	1.03
37. Misfit	1.72	1.72	1.23	1.27
38. Lonely	2.94	2.89	1.06	0.96
39. Excitable	4.22	3.44	1.26	0.92
40. Untrustworthy	0.44	0.39	0.62	0.61
41. Clean	4.83	4.61	0.99	0.85
42. Clean-cut	4.00	3.89	1.08	0.96
43. Innocent	3.39	3.28	0.98	0.89

APPENDIX F (Continued)

Trait Term	Mean Liking Rating		Standard Deviation	
	Personal	Popular	Personal	Popular
44. Eccentric	3.06	2.28	0.94	1.32
45. Materialistic	2.00	1.83	1.33	1.15
46. Foolish	1.72	1.61	0.83	0.92
47. Amusing	4.94	4.88	0.73	1.06
48. Wholesome	4.61	4.00	0.85	1.27
49. Cowardly	1.50	1.39	1.09	0.92
50. Superficial	1.00	1.28	1.03	1.18
51. Shallow	1.00	1.06	0.77	0.64
52. Liar	0.28	0.39	0.46	1.19
53. Tiresome	1.16	1.39	0.71	0.98
54. Jealous	1.39	2.11	0.85	1.32
55. Solemn	3.22	2.83	1.00	0.99
56. Grateful	4.39	4.17	1.42	0.79
57. Aggressive	3.17	3.61	1.09	1.24
58. Bragging	0.94	1.22	1.00	1.11
59. Insolent	1.83	1.61	1.25	1.09
60. Appreciative	2.17	1.00	1.42	1.24
61. Self-Possessed	2.17	1.00	1.42	1.24
62. Gossipy	1.22	1.33	1.11	1.03
63. Crude	1.22	0.94	1.06	1.26
64. Daydreamer	3.22	3.11	1.06	0.68
65. Dishonest	0.056	0.50	0.24	1.20
66. Gentle	5.11	4.56	0.76	1.04
67. Naive	2.44	2.89	1.25	1.18
68. Restless	3.11	2.89	0.96	1.08

APPENDIX G

Format and Liking Manipulation Instructions

Directions for Rating J. Morgan, R. Pierce, V. Witkin, and G. Ritter

In this set of materials you will find descriptions of four instructors and a set of rating forms, one form for each of the four instructors. The set of four instructors you have received was randomly selected from three sets of instructors.

The instructors were divided into three sets on the basis of descriptions of these instructors given by their students. In preliminary research, a group of students was asked to provide general descriptions of their instructors. Descriptions of twelve different instructors were provided by this group of students. A second group of students was given complete lists of the terms used by the first group of students to describe each instructor. The second group of students was asked to place the instructors into categories on the basis of these descriptors. It was found that most of the students in this second group divided the instructors into three sets. The three sets formed and some examples of the terms used to describe the instructors in each set are presented below.

	Examples of Descriptive Terms
SET 1	amusing, loyal, well-mannered, sharpwitted
SET 2	liar, underhanded, self-conceited, untrustworthy
SET 3	aggressive, solemn, restless, impressionable

The set of four instructors you have received is one of the above sets.

First, look over the rating forms and familiarize yourself with the five rating dimensions and their anchors. After you've familiarized yourself with the rating forms, set them aside and carefully read through each of the four instructor descriptions. After reading these descriptions, turn back to the rating forms and fill in your ID number on each form. Next, beginning with the first rating form circle the number on each dimension which corresponds to your evaluation of theat instructor on each dimension. When you finish the first form, go on to the second and so on until you complete all four forms. Make your ratings as accurately as you can. Do not look back at the descriptions when rating the instructors. After you are familiar with the rating forms you should read all four instructor descriptions with the rating dimensions in mind. After you have read all four descriptions, rate all four instructors without looking back at the descriptions.

If you have any questions, please ask the experimenter now.

The three page vita has been removed from the scanned document. Page 1 of 3

The three page vita has been removed from the scanned document. Page 3 of 3

THE EFFECT OF AFFECT IN PERFORMANCE APPRAISAL

by

Robert L. Cardy

(ABSTRACT)

The present investigation included three studies of the effect of liking upon the differential accuracy of performance ratings of instructors. Likableness and performance level were manipulated within vignettes. Liking was manipulated with trait terms that were found in a pretest to have little implication for performance on the rating dimensions. Instructor performance level was manipulated with incidents of teacher behaviors.

The first study investigated the effects of ratee likableness (liked vs. disliked vs. neutral) and ratee performance (high vs. low) as well as the effects of rater sex, rater selective attention ability (high vs. low), and the memory demand of the rating task (memory vs. no memory) on the differential accuracy of ratings. A total of 288 subjects, 144 males and 144 females, rated the performance of four instructor vignettes.

The differential accuracy of ratings was analyzed as a 3x2x2x2x2 between-subjects design. Differential accuracy was found to be significantly influenced by the performance level of raters, the selective attention of raters, and the memory demand of the rating task. The factors of likableness, ratee performance, and rater sex jointly influenced differential accuracy.

The second study investigated the effect of order of format presentation (before vs. after ratee observation) as well as ratee likableness and performance on rating accuracy. A 3x2x2 (liking x performance x format) between-subject ANOVA on the differential accuracy of ratings provided by 144 male and female subjects revealed the three factors to have a joint influence on rating accuracy. Ratings were more accurate in the format before than in the format after condition with liking and performance having an interactive effect only in the after condition.

Study 3 investigated the operation of schemata, discounting, and recall bias as well as liking and performance on rating accuracy. A regression analysis revealed the recall bias measure to be the only significant predictor of rating accuracy.

It was concluded that the pattern of results across the three studies did not offer direct support for either an integrality or schema conceptualization of affect. Consideration of schematic processing offered a potential explanation of the recall bias finding. Applied implications of the results are discussed.