Curtin University of Technology

Quality About RPHRM Current Issue Submissions Archive

RESEARCH AND PRACTICE IN HUMAN RESOURCE MANAGEMENT

SHRI Powering Human Capital

Highlight, copy & paste to cite:

Jawahar, I. M. & Stone, T. H., (1997). Appraisal Purpose Versus Perceived Consequences: The Effects of Appraisal Purpose, Perceived Consequences, and Rater Self Monitoring on Leniency of Ratings and Decisions, *Research and Practice in Human Resource Management*, 5(1), 33-54.

Volume 5: Issue 1

Editorial

Regular Papers

Practitioner Focus

Reviews

Appraisal Purpose Versus Perceived Consequences: The Effects of Appraisal Purpose, Perceived Consequences, and Rater Self Monitoring on Leniency of Ratings and Decisions

I. M. Jawahar & Thomas H. Stone

ABSTRACT

Almost 46 years ago, Taylor and Wherry (1951) hypothesized that performance appraisal ratings obtained for administrative purposes, such as pay raises or promotions, would be more lenient than ratings obtained for research, feedback, or employee development purposes. However, research on appraisal purpose has yielded inconsistent results. In this study, we offer and test two explanations. First, since purpose and consequences are naturally confounded, we argue that purpose effects reported in some studies (in contrast to those that failed to find purpose effects) may be due to the variation of consequences across appraisal purposes examined. Second, the inconsistent results may be due to the failure of previous studies to consider rater individual differences. These two possibilities were investigated in a factorially designed experiment in which 240 subjects who varied in level of self monitoring evaluated subordinates for one of the following purposes: administrative purpose with severe consequences, administrative purpose with less severe consequences, and training purposes. Results support the proposed role of perceived consequences by showing that self-monitoring interacts with consequences to influence leniency of ratings and personnel decisions.

INTRODUCTION

Several models of the performance appraisal process have proposethat leniency or accuracy of performance appraisal ratings may depend on the purpose for which those ratings are obtained (e., DeCotiis & Petit, 1978; DeNisi, Cafferty & Meglino, 1984). Performance appraisal ratings are used for several purposes including administrative purposes (e.g., allocating pay increases, promotions, making " retention " decisions), research purposes (e.g., validating selection tools), and employee development purposes (e.g., providing feedback, identifying training needs). Since performance ratings are used for several purposes (e.g., Cleveland, Murphy & Williams, 1989), one objective of

many empirical studies has been to investigate the influence of various appraisal purposes on characteristics of performance ratings such as leniency and accuracy. Most of this research has focused on Taylor and Wherry's (1951) hypothesis which argued that ratings obtained for administrative purposes are likely to be more lenient than those obtained for research, feedback or development purposes.

Early studies investigating Taylor and Wherry's (1951) appraisal purpose hypothesis also examined if the appraisal purpose effect depended on the rating scales used to obtain those ratings. For example, in a military setting, Taylor and Wherry (1951) investigated the resistance of graphic rating scales and forced choice scales to leniency of ratings provided for either "research purpose only" or for making "administrative decisions," such as promotion or demotion of ratees. They reported that ratings provided for administrative purposes were more lenient than those provided for research purposes. In addition, the graphic rating scale was found to be more susceptible to the purpose effect than the forced choice scale. Driscoll and Goodwin (1979) found that teacher ratings were more lenient when students were led to believe that those ratings would be used for making administrative decisions. In another study, Aleamoni and Hexner (1980) found that ratings provided for "salary and promotion" purposes were significantly more lenient than those generated under standard teacher evaluation instructions. In a field study, Bernardin and Orban (1990) examined the influence of three variables including appraisal purpose on leniency of ratings. In this study, thirty-two sergeants from two large municipal police departments evaluated sixty-five rookie patrol officers. As predicted, Bernardin and Orban found that ratings obtained for promotion purposes were more lenient than those obtained for feedback purposes. However, in this study, appraisal purpose was naturally confounded with departments, as one department used ratings for promotion purposes whereas the other used ratings for feedback purposes only.

Several studies have also found that appraisal purpose does not influence leniency of ratings. For instance, in a field study conducted with military personnel, Berkshire and Highland (1953) reported that ratings obtained for administrative purposes did not differ from those obtained for research purposes, regardless of whether those ratings were obtained on graphic rating or forced choice scales. Similarly, in a set of field studies, Hollander (1957, 1965) reported that the reliability and validity of peer-nominations of naval officers obtained for administrative purposes did not significantly differ from those obtained for research purposes. In another study, Centra (1976) also found no differences between teaching evaluations provided by students under administrative (tenure, salary, promotions) and feedback (improvement) conditions. Similarly, Gmelch and Glasman (1977) found no differences between ratings obtained for promotion versus feedback purposes.

Rather than using performance ratings, Zedeck and Cascio (1982) used personnel decisions to test for purpose effects. That is, after reviewing performance information of ratees, subjects directly proceeded to make administrative (e.g., pay increase, promotion or retention) or employee development (e.g., recommend ratee for training) decisions, without first evaluating performance of ratees relative to their job duties. In Zedeck and Cascio's study, 130 undergraduate students made decisions for 33 ratees described as supermarket checkers for one of the purposes of merit raise, development, or retention. Zedeck and Cascio used a seven point scale, with scale points 1, 4, and 7 anchored as follows for the three purposes: development (1 = strong need for development, 4 = some development needed, and 7 = no further development needed now); merit raise (1 = no raise recommended, 4 = average raise recommended, and 7 = highest raise recommended); and retention (1 = do not retain, 4 = neutral regarding retention, and 7 = strongly recommend retention). Using standard deviations of the subjects' responses as data points, Zedeck and Cascio concluded that decisions differ as a function of appraisal purpose with the difference being strongest between decisions made for merit raise and those made for either development or retention purposes. Bernardin and Cooke (1992) replicated the Zedeck and Cascio study, and the effect of purpose was also tested with identical anchors across rating purposes (1 strongly oppose personnel decision, 4 = neutral regarding personnel decision, and 7 = strongly support personnel decision). Bernardin and Cooke found a significant effect for purpose as well as for rating format,

but contrary to the results of Zedeck and Cascio, the greatest distinctions were not between merit raise and other purposes; rather, they were between retention and the other purposes. Murphy, Balzer, Kellam, and Armstrong (1984) found no difference between administrative and research ratings provided by students who rated videotaped lectures delivered by graduate students. In another study, McIntyre, Smith, and Hassett (1984) used videotaped performances of male drama students acting as lecturers and compared ratings provided by undergraduate students for the purposes of hiring, feedback and research. Since no analysis accounted for more than five percent of the total variance in accuracy, McIntyre, et al. concluded that the effect of perceived purpose may be weak. Alternatively, in a recent field study, Harris, Smith, and Champagne (1995) found ratings of 223 production employees obtained for administrative purposes to be significantly more lenient than those obtained for research purposes.

Thus, while studies by Taylor and Wherry (1951), Sharon (1970), Sharon and Bartlett (1969), Driscoll and Goodwin (1979), Aleamoni and Hexner (1980), Bernardin and Orban (1990) and others (e.g., Bernardin & Cooke, 1992; Harris, Smith, & Champagne, 1995; Kirkpatrick, Ewen, Barrett & Katzell, 1968; Meyer, Kay & French, 1965; Reilly & Balzer, 1988; Zedeck & Cascio, 1982) found a significant relationship between appraisal purpose and leniency/severity of ratings, several others (Berkshire & Highland, 1953; Bernardin, Abbott & Cooper, 1985; Borrensen, 1967; Centra, 1976; Gmelch & Glasman, 1977; Hollander, 1957, 1965; Meier & Feldhusen, 1979; McIntyre, Smith & Hassett, 1984) did not.

Synthesis and Critique

Research on appraisal purpose has generated contradictory results such that the relationship between appraisal purpose and leniency/severity as well as accuracy of ratings is clear. As a boundary variable, appraisal purpose has the potential to limit the external validity of performance appraisal research as performance ratings obtained for research purposes may be more lenient or severe than those obtained for administrative purposes. Consequently, suggestions based on ratings obtained for research purposes are likely to be of little value to practitioners who typically obtain them for making several important administrative and personnel decisions. Thus, this inconsistency has theoretical importance as well as practical relevance; and therefore, needs to be addressed.

While there may be many explanations for the inconsistent results reported for appraisal purpose, in this study, we offer and test two. First, Taylor and Wherry ' s hypothesis or the appraisal purpose hypothesis is based on the premise that raters bias ratings obtained for some purposes versus others as consequences of ratings collected for some purposes are more severe than those collected for other purposes. Thus, purpose effects are likely when consequences of ratings collected for one purpose are more (or less) severe than those collected for other purposes. However, previous studies testing for purpose effects did not check to ascertain if in fact raters ' perceptions of consequences of ratings collected for different purposes actually varied. Therefore, it is possible that in studies that reported purpose effects, raters would have perceived consequences of varying intensities across different appraisal purposes whereas in studies that failed to find purpose effects, raters ' perceptions of consequences of ratings collected for one purpose versus another would not have varied significantly.

To test this possibility, we manipulated appraisal purpose such that consequences would vary with manipulations of appraisal purpose by creating two administrative purpose conditions identical in all aspects except for consequences (see appraisal purpose manipulation). By creating the impression that funds for pay increase purpose were plentiful, consequences (of performance ratings) in one administrative purpose condition were made severe (henceforth administrative purpose with severe consequences condition). In the other administrative purpose condition, consequences were made less severe (henceforth administrative purpose with less severe consequences condition) by creating the impression that funds for pay increase purpose were non existent. Unlike previous studies, we measured raters ' perceptions of the extent to which consequences varied across appraisal purpoe conditions (administrative purpose with severe

consequences, administrative purpose with less severe consequences, and training purpose). We expect ratings (hypothesis 1) and decisions (hypothesis 2) provided in " administrative purpose with severe consequences " condition to be more lenient than those provided in " administrative purpose with less severe consequences " and " training purpose " conditions. Additionally, we expect larger " effect (size) " when ratings in " administrative purpose with severe consequences " condition are compared with those in the " training.purpose " condition than when ratings in " administrative purpose with less severe consequences " condition are conipared with those in the " training purpose " condition.

A second explanation, for the contradictory results reported for appraisal purpose may be the failure to identify individual differences with potential to moderate the influence of appraisal purpose and or consequences on leniency of ratings. A recent field investigation by Kane, Bernardin, Villanova, and Peyrefitte (1995) highlights the usefulness of individual differences for explaining inconsistent results. Kane, eta]. (1995) in three separate field studies found leniency to be a relatively stable response tendency and concluded that leniency could be predicted using measures of individual differences. Indeed, given equal situational parameters, difference in the tendency to rate leniently could reflect personality or information processing differences among raters (Wright & Mischel, 1987). While exceptions exist (Bernardin & Orban, 1990; Dobbins, Cardy & Truxillo, 1988, 1986), most prior investigations of purpose effects have ignored individual differences among raters, perhaps, on the assumption that all raters would bias ratings for some purposes versus others.

We contend that all raters may not have the ability as well as the motivation to provide ratings in anticipation of perceived consequences of those ratings for themselves as well as for the ratees. Indeed, research on self-monitoring conducted with college students as well as working adults suggests that, in contrast to low self-monitors, high self-monitors behave in anticipation of consequences of those behaviors (see Caidwell & O ' Reilly, 1982a,b; Fandt & Ferris, 1990; Snyder, 1987, 1979).

Self-monitoring theory and research suggests that high self-monitors are adept at deciphering cues in the social environment and are capable of tailoring their behaviors to fit the social context. In contrast, the behaviors of low self-monitors reflect their feelings and attitudes without regard to the situational or interpersonal consequences of those behaviors (Ajzen, Tiniko, & White, 1982; Snyder, 1979). Research has shown that in comparison to low self-monitors, high self- monitors monitor their behavioral choices on the basis of situational consequences (Snyder & Monson, 1975), manage impressions (CaIdwell & O ' Reilly, 1989a; Fandt & Ferris, 1990), present themselves in socially desirable ways (Lippa, 1978), adapt more effectively to different situations (Snyder, 1979); and engage in socially appropriate behavior to gain approval and minimize disapproval (Snyder & Cantor, 1980).

One example of the differences in self-monitoring dispositions of high and low self-monitors is illustrated in a study conducted by Caldwell & O ' Reilly (1 982b). In this study involving field representatives, Caldwell and O ' Reilly found performance on boundary-spanning jobs to be a function of self-monitoring. Since boundary spanning jobs require attention to cues in the environment, interpretation of those cues and appropriate responses, high self-monitors who behave in anticipation of consequences of those behaviors out-performed low self-monitors who behave without regard to social and interpersonal consequences. In another study, anticipated future interaction induced the " situationally guided " high self-monitors to become even more attentive to situational cues when deciding how to act, while prompting low self-monitors to rely even less on situational cues and more on personal thoughts and evaluations (Shaffer, Ogden, & Vu, 1987). That high self-monitors monitor consequences of behavioral choices before choosing a particular course of action is illustrated in another field experiment. White and Gèrstein (1987) found that when social rewards were contingent upon helping others, high self-monitors, in contrast to low self-monitors, were more likely to help others, but when consequences were less severe high self-monitors were less likely to help. Moreover, when high self-monitors perceived significant consequences (helping- social approval, not helping-social disapproval) they were twice

as likely to help than when consequences of helping were less severe.

Taken together, self-monitoring theory and research suggests that high self- monitors monitor their behavioral choices on the basis of situational consequences and behave in anticipation of consequences of those behaviors. Consequently, in a performance appraisal situation, we may expect raters who are high self-monitors to consider consequences of their ratings and rate in anticipation of consequences of those ratings. Among the three purposes examined in this study (administrative purpose with severe consequences, administrative purpose with less severe consequences, and training purpose), consequences are likely to be highest in " administrative purpose with severe consequences " condition and least in the " training purpose " condition. Since high self-monitors behave in anticipation of consequences and are motivated by the desire to gain social approval (and avoid disapproval), we expect high self-monitors to be lenient when ratings are collected for the former purpose. When ratings are used exclusively for training purposes, perceived consequences are likely to be minimal; consequently, high self-monitors are unlikely to be motivated to inflate or otherwise distort ratings. From White and Gerstein ' s (1987) study it appears that high self-monitors are unlikely to be lenient when consequences are less severe. In contrast to high self-monitors, beh viors of low self-monitors reflect their feelings, attitudes, and judgments without regard to the interpersonal or situational consequences of those behaviors. Therefore, ratings provided by low self-monitors are unlikely to be influenced by consequences that vary with appraisal purposes. We expect, the self-monitoring disposition of raters to moderate the relationship between appraisal purpose (and hence consequences) and leniency of ratings and decisions. Specifically, ratings and decisions of low self-monitors are not expected to vary across appraisal purposes. While ratings and decisions of high self-monitors are expected to be significantly lenient in " administrative purpose with severe consequences " condition, they are not expected to significantly differ from those of low self-monitors in the other two conditions.

Hypotheses

H1: Ratings provided in " administrative purpose with severe consequences " condition will be more lenient than those provided in " training purpose " (and " administrative purpose with less severe consequences ") condition.

H2: Decisions in " administrative purpose with severe consequences " condition will be more lenient than decisions in " training purpose " (and " administrative purpose with less severe consequences ") condition.

H3: Purpose and self-monitoring will interact to affect leniency. Ratings provided by high self-monitors vil] be most lenient in the " administrative purpose with severe consequences " condition.

METHOD

Subjects

A pilot study was conducted with 60 subjects (20 subjects in each appraisal purpose condition) to validate the appraisal purpose manipulations. Results of the pilot study and a power analysis (to detect an effect of size = .4, with a power of .8) indicated a required sample size of at least 40 subjects per experimental condition. The initial sample consisted of 265 undergraduate students enrolled in management classes at a large state university. Due to attrition, and incomplete responses, the final sample was reduced to 240 subjects with 40 subjects in each experimental condition. There were 86 juniors, and 154 seniors. Of these 133 were male and 107 were female.

Experimental Design and Procedure

The study was a 3 (administrative purpose with severe consequences, administrative purpose with less severe consequences, training purpose) X 2 (low self-monitor, high self-monitor) factorial design. This study was conducted in two phases separated by one month. In the first phase of the study, subjects completed the self-monitoring scale constructed by Snyder and his colleagues

(Snyder & Gangestad, 1986). Snyder and his colleagues ' new 18-item measure has an internal consistency of .70. Although a detailed review of the literature is beyond the present scope of this article, note that there has been substantial research demonstrating the construct validity of self-monitoring and the self-monitoring scale, in particular (see for example Snyder, 1987, 1979; Gangestad & Snyder, 1991, 1985). Consistent with past research on self-monitoring (see Gangestad & Snyder, 1985), subjects with self-monitoring scores equal to or greater than 11 were categorized as high self-monitors and those with scores equal to or less than 10 were categorized as low self-monitors. Snyder ' s self-monitoring scale has a classification accuracy (i.e., accuracy of correctly identifying an individual as a low or a high self-monitor) of 87 % (Gangestad & Snyder, 1985). Furthermore, consistent with prior research, the distribution of self-monitoring scores in this study was bimodal with peaks at scores of 8 (low self-monitors) and 11 (high self-monitors). 43 % of the subjects who completed the self-monitoring scale were identified as high self-monitors and.the remaining as low self-monitors. The percentage of high and low self-monitors in the present sample is consistent with the 40/60 split of high and low self-monitors reported in past research (see Gangestad & Snyder, 1985, p. 334; see also Snyder, 1987).

In the second phase of the study, subjects were blocked on the self-monitoring variable and low and high self-monitors were randomly (and independently) assigned to each of the three conditions. Each subject received an information packet containing a letter, scenario, appraisal purpose (either ' administrative purpose with severe consequences " or " administrative purpose with less severe consequences " or " training purpose "),. and performance stimuli. The scenario contained (1) a brief description of a mail-order company specializing in a wide range of outdoor products, and (2) a job description of sales representatives. Performance information was presented to subjects in the form of critical incidents. For instance, one critical incident read " made a recommendation about adding Spencer fishing poles because of numerous customer suggestions " and another " lost temper when dealing with an upset customer. " Twenty-five such incidents captured performance of each of the two subordinates, Pat and Chris. Additionally, the order in which critical incidents capturing performance of Pat and Chris were presented was counterbalanced within each cell. Although performance information of two subordinates, Pat and Chris was provided, all raters were instructed to rate performance of Pat only. We provided performance information on Chris so that raters would have a " standard " for evaluating Pat. Also, performance information was rigged so that Chris was clearly the better performer of the two. Pat ' s performance relative to that of Chris was portrayed as poor to avoid a ceiling effect and allow room for raters to inflate ratings.

In the letter, subjects were instructed to first familiarize themselves with the scenario, performance appraisal form, critical incidents and then evaluate performance of Pat on a Behaviorally Anchored Rating Scale (BARS) designed for this study. The BARS was developed from the job description of ratees (sales representative) to measure performance on the five dimensions of interpersonal and communication skills, dependability, quality of work, knowledge of company products and sales procedures, and initiative. Additionally, depending on the experimental condition, subjects made a decision involving either merit raise or training. After administering manipulation checks for purpose and consequences, subjects were debriefed and thanked for their participation.

Manipulation

Administrative purpose with severe consequences: Please note that this company uses performance appraisal ratings for merit raise (that is, pay increases based on performance) purpose only. This year the company made unusually large profits and consequently funds in the pay raise budget have been tripled. So this year there will be plenty of funds/money for pay increases. Funds in the pay-raise budget are expected to remain at the current level for at least 2 or 3 years.

Administrative purpose with less severe consequences: Please note that this company uses performance appraisal ratings for merit raise (that is, pay increases based on performance) purpose only. This year the company incurred heavy losses and consequently there are no funds in the .pay raise budget. So this year there will be no funds/money for pay increases. Fuids in the pay raise budget are not expected to increase dramatically for at least 2 or 3 years.

Training purpose: Please note that this company uses performance appraisal ratings for training purpose (that is, identifying employees ' training needs) only. Training is provided to improve job knowledge, skills or abilities. The duration of the training typically varies from 1 to 3 days. When employees attend training programs, the company provides regular wages/salary and a temporary worker is assigned to replace the employee during the employee ' s absence.

Dependent Variables

True scores. The most widely used procedure for computing true score estimates was developed by Borman (1977). Briefly, this procedure involves the use of multiple experts who evaluate performance under optimal conditions. The mean rating of expert judges provides a " true-score " measure of a specific ratee ' s performance in that this mean rating approximates the expected value of the rating obtained from an expert who is observing behavior under optimal conditions. If these true score measures are collected, it becomes possible to assess a subject ' s accuracy in rating a ratee ' s performance on several dimensions by comparing the ratings with this true score. Using expert ratings as true scores is widely accepted in the appraisal literature (see McIntyre, et al., 1984; Murphy & Cleveland, 1991).

In this study, two management faculty members with a combined experience of over 30 years in teaching, research, and consulting related to performance appraisal served as expert raters. These expert raters were thoroughly familiarized with ratees ' job description, exhibited job behaviors, the nature and contents of performance dimensions, and the performance appraisal instrument. A written copy of this information was also provided and the experts were encouraged to refer to this material while evaluating performance. Agreement between raters was very high (reliability, _r = .86). Consistent with prior studies, mean expert ratings generated through this procedure were used as true score measures of performance.

Performance Ratings. Leniency (severity) is defined as a rater ' s tendency to assign ratings that are higher (leniency) or lower (severity) than the true scores. Ratings provided by subjects were manipulated using the formula presented in McIntyre, et al., (1984) to compute leniency. Higher values indicate higher levels of leniency. Our measure of leniency is the same as Cronbach ' s (1955) measure of elevation accuracy.

Personnel Decisions. Following Bernardin and Cooke (1992) we used the same scale for both administrative (i.e., merit raise) and training decisions. Bernardin and Cooke used the following scale points: 1 = strongly oppose personnel decision, 4= neutral regarding personnel decision, and 7 = strongly support personnel decision. In follow-up questions, some subjects in our pilot study indicated that the words " personnel decision " lead them to infer that the personnel decision had already been made and that they were to either oppose or support this decision. To avoid such confusion we used the following anchors: 1 = strongly oppose, 3 = somewhat oppose, 5 = neutral, 7 = somewhat support, 9 = strongly support. This scale was prefaced by two statements. Depending on the appraisal purpose, the wordings of these statements were slightly altered. The first read " please make a training (or merit raise) decision. " The second read " decision is whether to send subordinate for training (or give a merit raise). "

Theoretical Measures. To support the proposed theoretical rationale underlying the study, the following operations were performed. First, unlike previous studies, we did not permit consequences to vary randomly with manipulations of appraisal purpose. Instead, we manipulated appraisal purpose such that the severity of consequences would decrease from one treatment condition to another in the order of administrative purpose with severe consequences, administrative purpose with]ess severe consequences, and training purpose. Second, un]ike previous studies, we measured subjects ' perceptions of consequences of ratings and decisions to ascertain if, in fact, subjects ' perceptions of consequences varied as predicted. Two items, " The performance ratings that you just provided will significantly affect your subordinate ", and " The performance ratings you provided will not have any consequences for your subordinate, " both rated on a 9-point scale (1- strongly disagree, 3 - disagree, 5 - neither disagree nor agree, 7 - agree, 9 - strongly agree) were used to measure perception of consequences. Additionally, we also measured

the extent to which subjects considered those consequences while evaluating performance and making administrative (merit raise) and training decisions. Two items, " To what extent did you consider consequences of performance ratings for the ratee while evaluating performance, " and " While recommending merit-raise (or training, depending on purpose condition), to what extent did you consider how your merit-raise (Or training) decision would affect ratee, " both rated on aS - point scale (1 - did not consider at all, 3 - somewhat considered it, 5 - considered it a great deal) were used. While high and low self-monitors (within each condition) were not expected to differ in their perceptions of consequences, we expected the former in comparison to the latter to appraise and make personnel decisions in anticipation of consequences of those decisions.

RESULTS

Manipulation Checks:

In this study, consequences were manipulated to vary with appraisal purposes. As expected, perception of consequences was higher (mean 6.8) in the " administrative purpose with severe consequences " condition than in the " administrative purpose with less severe consequences " condition (mean 6.3), which in turn, was higher than perceived consequences in the " training purpose " condition (mean 4.6). T tests, conducted to ascertain the efficacy of manipulation of consequences indicated that means in the three purpose conditions were significantly different ($p < .05$) from each other. Second, results of a T test indicated that the appraisal purpose manipulation was successful, as subjects in the administrative and training purpose conditions correctly identified ($p < .001$) the purpose for which they evaluated performance. Third, as expected, no order effects were noted (see section on experimental procedure). Additionally, the scores of low self-monitors and those of high self-monitors across the three experimental conditions were equivalent.

Means and standard deviations of dependent variables leniency and decision are presented in Table 1. Since calculation of leniency involves subtracting rater ' s ratings from true scores (see formula, McIntyre, et a]., 1984, p. 151), a negative number indicates leniency. In all conditions, raters assigned ratings higher than true scores, indicating lenient ratings. We decided to drop the negative sign when reporting those values in the manuscript. In Table 2 a summary of the analysis of variance for the dependent variable leniency is reported.

Table 1
Means and Standard Deviations of Dependent Variables

Dependent Variables	Self-Monitoring	Appraisal Purpose					
		APWSC		APWLSC		Training	
		M	SD	M	SD	M	SD
Leniency	LSM	0.57	0.50	0.57	0.56	0.41	0.43
	HSM	1.32	0.55	0.27	0.50	0.41	0.50
Decision	LSM	3.45	1.97	2.75	1.63	8.35	0.95
	HSM	5.85	1.41	2.18	1.17	8.08	1.31

Note: LSM — Low self-monitors, HSM High self-monitors,
APWSC — Administrative purpose with severe consequences,
APWLSC — Administrative purpose with less severe consequences.

Table 2
Results of Analysis of Variance for Dependent Variable Leniency

Source	DF	SS	MS	F-Value	Pr>F
Self-Monitoring	1	1.21	1.21	4.70	.03

Purpose	2	15.15	7.57	29.54	.0001
Self-Monitoring X Purpose	2	12.01	6.0	23.44	.0001
R^2	0.32				

Note: Degrees of freedom for model -5, and error -233. Mean Square for Leniency (model-5.67, error -0.26). F and R^2 values are based on partial sums of squares (type III SS) that are invariant to the ordering of effects in the model.

Hypothesis Testing:

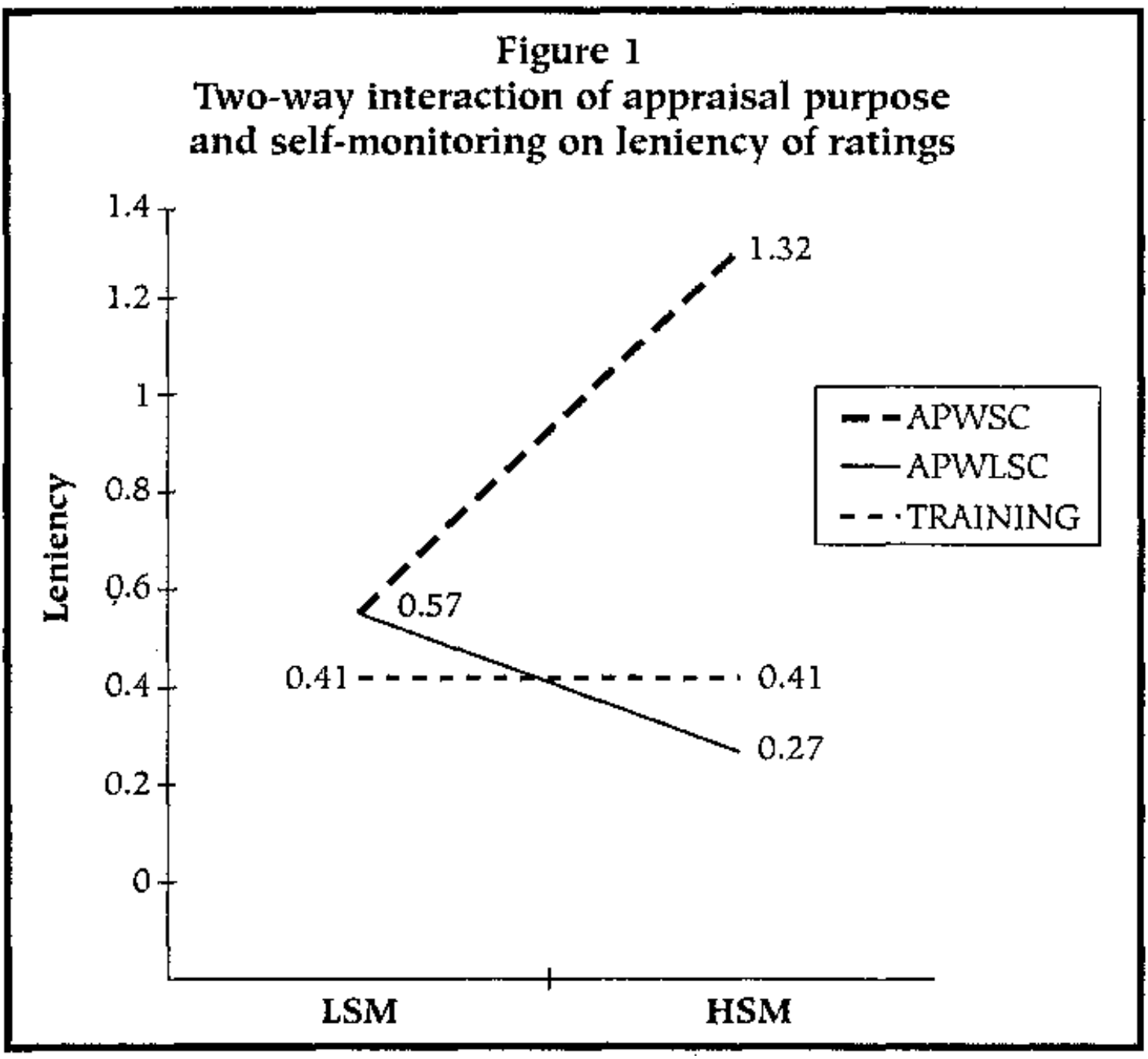
As predicted in hypothesis 1, appraisal purpose significantly influenced leniency ($F_{2,233}$ = 29.54, p < .0001, see Table 2) of appraisal ratings. More importantly, ratihgs (mean 0.95) obtained for " administrative purpose with severe consequences " were significantly more lenient (t_{158} = 6.825, p < .001) than ratings (mean 0.41, see Table 1) obtained for " training purposes, " whereas ratings (mean 0.42) obtained for " administrative purpose with less severe consequences " were not significantly different (p > .5) from those (mean 0.41) obtained for " training purposes. " As expected, effect size (see Cohen, 1988) for the former comparison (d = 1.1) was much larger than that (d = .02) of the latter comparison. Additionally, ratings (mean 0.95) obtained for " administrative purpose with severe consequences " were significantly more lenient (t_{158} = 6.324, p < .00 1) than those (mean 0.42) obtained for " administrative purpose with less severe consequences. "

As predicted in hypothesis 2, the effect of appraisal purpose was also tested with decision as the dependent variable. As mentioned before, a 9 point scale with scale anchors of 1 = strongly oppose, 3 = somewhat oppose, 5 = neutral, 7 = somewhat support, and 9 = strongly support, was used for merit raise and training decisions. Since the true performance level of the ratee was low (see section on experimental procedure), subjects in the merit raise conditions generally opposed giving the ratee a merit increase (see means, Table 1) whereas those in the training condition generally supported remedial training for the ratee (see Table 1). A subject ' s administrative or training decision is actually a data point in the distribution of administrative or training decisions. Hence, administrative decisions and training decisions are data points in two different distributions. Since these two distributions and hence the data points contained in them, are not directly comparable (see Zedeck & Cascio, 1982), unlike previous studies that used standard deviations of subjects ' decisions as data points to test for purpose effects, we used standardized values of decisions. We obtained standardized values (z = sigma (x_1 - X)/s_1) of " administrative decisions with severe consequences, " " administrative decisions with less severe consequences, " and " training decisions. " As predicted, the mean standardized estimate (z_1 = .72) of " administrative decisions with severe consequences " was significantly larger (d = z_1 - z_3 = .61) than the mean standardized estimate (z_3 = .12) of " training decision , " whereas, the mean standardized estimate (z_2 = .21) of " administrative decision with less severe consequences " was not significantly different (d = z_2 - z_3 = .09) from that (z_3 = .12) of the " training decision. " Additionally, the mean standardized estimate of " administrative decisions with severe consequences " was significantly larger (d = z_1 - z_2 = .51) than that of " administrative decisions wherein consequences were less severe. " Since a d of .2 is considered a small effect size, and a d of .5 is considered a medium effect size (see Cohen, 1988, pp. 25-26), we conclude support for hypothesis two.

Hypothesis 3 predicted that purpose and self-monitoring will interact to affect leniency. Additionally, ratings provided by high self-monitors were hypothesized to be most lenient in the " administrative purpose with severe consequences " condition.

As hypothesized, purpose and self-monitoring interacted to influence leniency ($F_{2,233}$ = 23.44, p < .0001; see Table 2). Figure 1 depicts the observed interaction. Additionally, the interaction hypothesis predicted specific differences in the ratings of high and low self-monitors across experimental conditions. To test these differences, the LSMEANS procedure (Searle, Speed, & Millikin, 1980) was used to make preplanned comparisons between means.

Figure 1
Two-way interaction of appraisal purpose and self-monitoring on leniency of ratings



Note: LSM — Low self-monitors, HSM — High self-monitors,
APWSC — Administrative purpose with severe consequences, and
APWLSC — Administrative purpose with less severe consequences.

As expected, ratings of low self-monitors did not differ ($p > .1$) across appraisal purposes (see Table 1). Ratings of high self-monitors were significantly more lenient ($p < .0001$) in the "administrative purpose with severe consequences" (mean 1.32) condition than in "administrative purpose with less severe consequences" (mean 0.27) and "training purpose" (mean 0.41) conditions. Additionally, ratings of high self-monitors were not significantly different ($p > .1$) in the latter two conditions; and as expected, high self-monitors were significantly more lenient ($p < .0001$) than low self-monitors in the "administrative purpose with severe consequences" condition.

DISCUSSION

The major impetus for this study was the inconsistency of prior research investigating the effects of appraisal purpose on leniency of ratings. For example, while Bernardin and Orban (1990) found a significant relationship between appraisal purpose and leniency, Meier and Feldhusen (1979.) did

not. Studies examining the influence of appraisal purpose on decision characteristics have also reported inconsistent results. For instance, while studies by Zedeck and Cascio (1982) and Bernardin and Cooke (1992) found a significant relationship between appraisal purpose and decision inflation, Bernardin, Abbott and Cooper (1985) failed to find such a relationship. Addressing this inconsistency is important not only from a researcher ' s internal validity concerns but also for external validity concerns that are important for organizational application.

In this study, we offered and tested two explanations for the inconsistent results. First, we argued that purpose and consequences are naturally confounded such that purpose effects are likely when consequences of ratings collected for one purpose are more (or less) severe than those collected for other purposes. Consequently, purpose effects reported in some studies (in contrast to those that failed to find purpose effects) may be due to consequences that vary with appraisal purposes, and not necessarily due to appraisal purpose per Se. To test this possibility, ratings and decisions collected in two administrative purpose conditions (identical in all aspects except for the severity of consequences) were compared with those collected for training purposes. Ratings (hypothesis 1) and decisions (hypothesis 2) in " administrative purpose with severe consequences " condition were significantly more lenient than those in " administrative purpose with less severe consequences " and " training purpose " conditions; whereas, there were no significant differences between ratings and decisions in the latter two conditions. Support for hypotheses 1 and 2 suggests that consequences that vary with appraisal purpPse have more influence on leniency than appraisal purpose per se.

Second, we suggested that the inconsistent results may be due to the fact that previous studies ignored individual differences among raters, and argued that raters ' self-monitoring disposition would moderate the relationship between appraisal purpose and leniency. The interactionhypothesis based on the premise that high self-monitors (in contrast to low self-monitors) would provide ratings in anticipation ofconsequences contingent upon those ratings was fully supported. Preplanned comparisons further confirmed the hypothesized pattern of the interaction. Specifically, while ratings of low self-monitors did notvary across treatment conditions, high self-monitors were most lenient when consequences were high as was the case when those ratings were collected for " administrative purpose with severe consequences. " Our results are strengthened as low and high self-monitors did not differ in their perceptions of consequences (t_{238} = 1.1, p > .05), but as expected, in contrast to low self-monitors (mean 2.85), high self-monitors (mean 3.21) took consequences into consideration while evaluating performance (t_{238} = 2.88, p < .001). The entire pattern of results suggest that " purpose effects " are due to consequences that underlie or vary with appraisal purposes for which ratings are obtained, and confirm the tendency of high self-monitors to evaluate performance in anticipation of those consequences.

The inability to discriminate inaccurate (lenient) from accurate raters and reward the latter has been noted as the most significant barrier to achieving accuracy in performance appraisal ratings (e.g., Murphy & Cleveland, 1991). Our study is one of the few to identify raters and conditions under which those raters are likely to be lenient. Consistent with the conditional approach to dispositional constructs (Wright & Mischel, 1987), when consequences were high, high self-monitors rated more leniently than low self-monitors, but ratings of low and high self-monitors did not differ when consequences were less severe. Our results corroborate Kane, et al. ' s (1995) proposition that given equal situational parameters (consequences of appraisal purpose, in this study), difference in the tendency to rate leniently or severely might reflect personality differences (e.g., self-monitoring) among raters.

Potential Limitations

The lack of consensus for purpose effects in the extant literature was the major impetus for this study. Given that this study was essentially concerned with theory-testing, concerns about addressing threats to internal validity took precedence over generalizability of results (Cook & Campbell, 1979); and consequently, this study was conducted in a laboratory setting. In this study,

we were able to manipulate perceptions of some raters that consequences of administrative rat-j ings are severe while simultaneously leading other raters to believe the opposite. In contrast to the]aboratory, ongoing organizations are unlikely to permit the use qf such designs. Another major benefit of this laboratory design is that previous performance and supervisor-subordinate interactions do not contaminate dependent measures. A potential limitation of this study is the use of undergraduate students as subjects. However, college students do have experience in performance appraisals, i.e., evaluating faculty; and on many college campuses, those evaluations do serve as input for important administrative decisions including faculty tenure decisions. Nonetheless, using organizational raters as subjects would have enhanced the ecplogical validity of the study.

As severity of consequences increase, raters are likely to feel accountable (Tetlock, 1985) for the ratings they assign to ratees. Results of our study suggests that high self-monitors are more likely to succumb to accountability pressures and inflate ratings, than low self-monitors. Inspite of the fact that no real consequences were made available, subjects ' perceptions of consequences had the predicted effect. Clearly, considering the interdependent nature of the rater-ratee relationship, consequences of ratings are likely to be very high (and real) in the field (see Ilgen & Favero, 1985). If so, one might contend that the severe consequences of ratings (in organizations) will increase accountability pressures (Tetlock, 1985), overwhelm rater individual differences and elicit lenient ratings from both low and high self-monitors. We believe, that the severity of consequences of ratings and accompanying accountability pressures will actually activate (and perhaps accentuate) individual differences so that ratings of high self-monitors will be even more lenient than those of low self-monitors. For instance, in a field experiment involving customer service employees, decision makers who were high self-monitors engaged in more information manipulation to justify their decisions than low self- monitors. More importantly, these effects were more pronounced under conditions of high accountability (Fandt & Ferris, 1990), and are consistent with much of the earlier research on self-monitoring (see Caldwell & O ' Reilly, 1982a; Snyder, 1979). Therefore, effect sizes reported in this study may actually be an underestimate of what may be observed with raters in ongoing organizations and real consequences. Indeed, reviews of performance appraisal literature (e.g., Bernardin & Villanova, 1986; Murphy & Cleveland, 1991) have documented greater leniency effects in ongoing organizations in comparison to laboratory investigations.

Implications for Research and Practice

This study investigated one avenue through which appraisal purpose influences ratings and related personnel decisions. Support for hypotheses 1 and 2 suggests that ratings and decisions are more likely to be distorted as the severity of consequences increase.

Nonetheless, out results do not fully explain conflicting results reported in prior appraisal purpose research. For instance, while support for hypotheses 1 and 2 by these data can explain ernardin and Cooke ' s (1992) results of greater leniency for retention versus merit raise and employee development decisions, our results are inconsistent with Zedeck and Cascio ' s (1982) study that found greater leniency for merit raise decisions in comparison to retention and employee development decisions. Since prior research did not measure consequences, we can only speculate that retention decisions are likely to have more severe consequences than merit raise or employee development decisions. Should study results be replicated in field settings, future research should examine the extent to which consequences vary, with manipulations of different appraisal purposes and also investigate rater individual differences with potential to moderate the effects of appraisal purpose on rating characteristics.

Other avenues through which appraisal purpose may influence ratings include basic cognitive processes. DeNisi and his colleagues (e.g., DeNisi & Williams, 1988; Williams, DeNisi, Blencoe & Cafferty 1985) have conducted a number of studies that indicate that appraisal purpose affects information acquisition as well as the type of information acquired. There is also some evidence that raters weigh, combine and integrate individual dimensions of performance differently, depending on the purpose of appraisal. More research using verbal protocol analysis (Martin &

Klimoski, 1990) and policy-capturing.(Zedeck & Cascio, 1982) is needed to uncover the cognitive processes underlying appraisal purpose effects.

Given the desire to reduce leniency bias in performance appraisals (see Kane, et al., 1995), the primary contribution of this study is the support for the moderating effect of rater self-monitoring on the relationship between consequences that vary with appraisal purpose and leniency. Very few studies have examined individual differences with potential to moderate purpose effects. In a laboratory study involving undergraduate students, Dobbins, Cardy, and Truxillo (1988) found that when appraisals were made for administrative purposes, raters with traditional stereotypes of women evaluated female ratees less accurately than those with nontraditional stereotypes of women. In another study,. although raters ' trust in appraisal process (TAPS) did not interact with appraisal purpose as hypothesized, it was negatively related to leniency (Bernardin & Orban, 1990). While individual differences in " rater discomfort, " as measured by the performance appraisal discomfort scale, has been associated with leniency (Villanova, Bernardin, Dahmus, & Sims, 1993), " rater discomfort " has not been investigated in the context of appraisal purpose. In contrast to ratings obtained for research or employee development purposes, ratings obtained for administrative purposes (e.g., retention) are likely to highten rater discomfort and the tendency to be lenient. Future research should focus on identifying raters as well as reasons why those raters are likely to bias ratings for some purposes versus others.

Support for hypothesis 3 suggests that raters ' motivation to evaluate employee performance and make accurate decisions may be more important than generally believed. Leniency may actually be an instance of adaptive behavior. Given the ongoing nature of the rater-ratee relationship, raters who doubt their ability to cope with consequences of their harsh but accurate ratings (e.g., ratee ' s fury, unpleasant work atmosphere, etc.) may be motivated to rate leniently and thereby avoid unpleasant confrontations. For instance, there is ample evidence that people fear and avoid threatening situations they believe exceed their coping abilities, whereas they behave assuredly when they judge themselves capable of managing situations that otherwise intimidate them (Bandura, Adams, & Beyer, 1977; see also Bandura, 1977, 1982). Therefore enhancing coping-efficacy of raters may be just as important as increasing rater ' s ability to accurately evaluate performance. Alternatively, by encouraging raters to provide feedback at regular intervals throughout the appraisal period, raters ' apprehensiveness to rate accurately could be reduced. To reward accurate raters and educate and/or train inaccurate raters, those raters and the conditions that activate rater differences must first be identified. Our study results suggest that as the severity of consequences increase, performance appraisals and decisions of high self-monitors, in contrast to those of low self-monitors, are likely to be lenient.

Theories and results of empirical research may be bounded by culture (Hofstede, 1993). Support for a phenomenon in one culture, may be stronger or weaker in another culture depending on the extent to which cultural differenc.es strengthen or weaken the psychological processes that govern the behavior of interest. Our results suggest that when consequences of ratings are severe, high self-monitors will provide lenient ratings in anticipation of those consequences. While self-monitoring in individualistic cultures involves behaving as prototypic persons would behave in the situation at hand, Snyder (1979) suggested that high self-monitors in collectivistic cultures must take into consideration the specific individuals present in the situation and their status relationship with them in deciding how to behave in a particular situation (Gudykunst, Yang, & Nishida, 1987).

Because of the public nature of performance evaluations, harsh but accurate ratings, are likely to result in " loss of face. " Saving face is a very fundamental concern in cultures such as the Asian cultures that are higher in terms of collectivism and uncertainty avoidance than the North American culture (Hofstede, 1980). That Taiwanese were less likely to believe that " supervisors should be honest with criticism " than North Americans (McEvoy & Cascio, 1990) highlights this concern. Consequently,.McEvoy and Cascio noted that " American performance appraisal practices should be modified to provide less direct and open confrontation between supervisors and subordinates. Negative feedback from the boss is likely to cause serious problems and must be provided in a very tactful manner (1990, p. 217). " Oe avenue to do so would be to provide

negative feedback in private, but commend the subordinate in public or simply turn in more lenient performance evaluations. To the extent that harsh but accurate evaluations lead to loss of face and other negative consequences, the results of this study are likely to be replicated in Asian cultures as well.

REFERENCES

Ajzen, I., Timko, C., & White, J. B. (1982). Self-monitoring and the attitude-behavior relation. *Journal of Personality and Social Psychology*, 42, 426-435.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on theeffect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, 9, 67-84.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37, 127-147.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84 (2), 191-2 15.

Bandura,.A., Adams, N. E., & Beyer, J. (1977). Cognitive processes mediating behavioral change. *Journal of Personality and Social Psychology*, 35, 125-139.

Berkshire, I. R., & Highland, R. W. (1953). Forced - choice performance rating - A methodological study. *Personnel Psychology*, 6, 355-3 78.

Bernardin, H. I., Abbott, J., & Cooper, D. (1985). The effects of appraisal purpose and rater training on rating characteristics. Paper presented at the *Annual Academy of Management Meetings*, San Diego.

Bernardin, H. J., & Cooke, D. K. (1992). Effects of appraisal purpose on discriminability and accuracy of ratings. *Psychological Reports*, 70, 1211-1215.

Bernardin, J. H., & Orban, J. A. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology*, 5(2), 197-211.

Bernardin, H. I., & Villanova, P. (1986). Performance appraisal. In E.A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 43-62). Lexington, MA: Lexington.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20, 238-252.

Borrensen, H. A. (1967). The effects of instruction and item content on three types of ratings. *Educational and Psychological Measurement*, 27, 855-862.

Caidwell, D., & 0 ' Reilly, C. (1982a). Responses to failure: The effects of choice and responsibility on impression management. *Academy of Management Journal*, 25, 121-136.

Caldwell, D., & 0 ' Reilly, C. (1982b). Boundary spanning and individual performance: The impact of self-monitoring. *Journal of Applied Psychology*, 67, 124-127.

Centra J. E. (1976). The influence of different directions on student ratings of instruction. *Journal of Educational Measurement*, 13, 277-282.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of perforrnance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, 74, 130-135.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Eribaum Associates.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Chicago: RandMcNally.

Cronbach, L. (1955). Processes affecting scores on " understanding of others " and assumed similarity, *Psychological Bulletin*, 59, 177-193.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review*, 3, 635-646.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance*, 33, 360-396.

DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In K.M. Rowland and G.R. Ferris (Eds.), *Research in Personnel and Human Resource Management*, 6, 109-155. Greenwich, CT: JAI Press.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology*, 73, 55 1-558.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1986). Effects of ratee sex and purpose of appraisal on the accuracy of performance evaluations. *Basic and Applied Social Psychology*, 7, 225-24 1.

Driscoll, L. A., & Goodwin, W L. (1979). The effects of varying information about the use and disposition of results on university students ' evaluations of faculty and courses. *American Educational Research Journal*, 16, 25-3 7.

Fandt, P. M., & Ferris, G. R. (1990). The management of information and impressions: When employees behave opportunistically. *Organizational Behavior and Human Decision Processes*, 45, 140-158.

Gangestad, S. W, & Snyder, M. (1991). Taxonomic analysis redux: Some statistical considerations for testing a latent class model. *Journal of Personality and Social Psychology*, 61, 141-146.

Gangestad, S. W, & Snyder, M. (1985). " To carve nature at its joints ": On the existence of discrete classes in personality. *Psychological Review*, 92, 317-349.

Gmelch, WH., & Glasman, N. S. (1977). The effects of purpose on student evaluation of college instructors. *Educational Research Quarterly*, 2, 45-55.

Gudykunst, W B., Yang, S. M., & Nishida, T. (1987). Cultural differences in self- consciousness and self-monitoring. *Communication Research*, 14 (1), 7-34.

Harris, M. M., Smith, D. E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research versus administrative based ratings. *Personnel Psychology*, 48, 151-160.

Hofstede, G. (1993). Cultural constraints in management theories. *The Academy of Management Executive*, 7 (1), 81-94.

Hofstede, G. (1980). *Culture ' s consequences: International differences in work-related values*. Beverly Hills, CA: Sage Publications.

Hollander, E. P. (1965). Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology*, 49, 434-438.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology*, 41, 85-90.

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

Kane, J. S., Bernardin, J. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, 34(4), 1036-1051.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment*. New York, New York: University Press.

Lippa, R. (1978). The effects of expressive control on expressive consistency and on the relation between expressive behavior and personality. *Journal of Personality*, 46, 438-46 1.

Martin, S. L., & Kllmoski, R. 1. (1990). Use of verbal protocols to trace cognitions associated with self and supervisor evaluations of performance. *Organizational Behavior and Human Decision Processes*, 46, 135-154.

McEvoy, G. M,, & Cascio, W F. (1990). The United States and Taiwan: Two different cultures look

at performance appraisal. In B. B. Shaw and J. E. Beck (Guest Eds.) and G. R. Ferris and K. M. Rowland (Eds.), *Research in Personnel and Human Resources Management*, (Supplement 2, pp. 201-220), Greenwich, CT: JAI Press.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, 69, 149-156.

Meier, R. A., & Feidhusen, I. F. (1979). Another look at Dr. Fox: Effect of stated purpose of evaluation, lecturer expressiveness, and density of lecture content on student ratings. *Journal of Educational Psychology*, 71, 339-345.

Meyer, H. H., Kay, E., & French, 1. (1965). Split roles in performance appraisal. *Harvard Business Review*, 43, 123-129.

Murphy, K. R., Balzer, W K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. *Journal of Educational Psychology*, 76, 45-5 4.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective.* Allyn and Bacon, Boston.

Reilly, C. E., & Balzer, V. K. (1988). Effect of purpose on observation and evaluation of teaching performance. *Unpublished manuscript*, Bowling Green State University.

Searle, S. R., Speed, F. M., & Millikin, G. A. (1980). Population marginal means in the linear model: An alternative to the least squares mean. *The American Statistician*, 34, 216-221.

Shaffer, D. R., Ogden, J. K., & Wu, C. (1987). Effects of self-monitoring and prospect of future interaction on self-disclosure reciprocity during the acquaintance process. *Journal of Personality*, 55, 75-96.

Sharon, A. (1970). Eliminating bias from student ratings of college instructors. *Journal of Applied Psychology*, 54, 278-281.

Sharon, A., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology*, 22, 251-263.

Snyder, M. (1987). *Public appearances private realities: The psychology of self- monitoring*. New York: WH. Freeman & Co.

Snyder, M. (1979). Self-monitoring process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (vol. 12). New York: Academic Press.

Snyder, M., & Cantor, N. (1980). Thinking about ourselves and others: Self-monitoring and social knowledge. *Journal of Personality and Social Psychology*, 39, 222-234.

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology*, 51(1), 125-139.

Snyder, M., & Monson, T. C. (1975). Persons, situations, and the control of social behavior. *Journal of Personality and Social Psychology*, 32, 63 7-644.

Taylor, E. K., & Wherry, R .1. (1951). A study of leniency in two rating systems. *Personnel Psychology*, 4, 39-47.

Tetlock, p. (1985). Accountability: The neglected social context of judgment and choice. In B. M., Staw and L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. pp. 297 - 332). Greenwich, CT: JAI Press.

Villanova, P., Bernardin, J., Dahmus, S. A., & Sims, R. L. (1993). Rater leniency and performance appraisal discomfort. *Educational and Psychological Measurement*, 53, 789-799.

White, J. W., & Gerstein, L. H. (1987). Helping: The influence of anticipated social sanctions and self-monitoring. *Journal of Personality*, 55, 41-54.

Williams, K. I., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and*

Human Decision Processes, 35, 3 14-339.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology*, 53, 1159-1177.