

THE UNIVERSITY OF CALGARY

A Comprehensive Examination of the Quality of
Self-Ratings of Performance

by

Lisa M. Keeping

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER
OF SCIENCE

DEPARTMENT OF PSYCHOLOGY

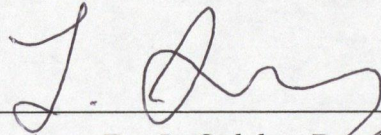
CALGARY, ALBERTA

APRIL, 1995

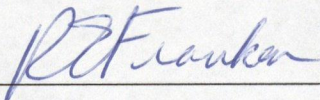
© Lisa M. Keeping 1995

THE UNIVERSITY OF CALGARY
FACULTY OF GRADUATE STUDIES

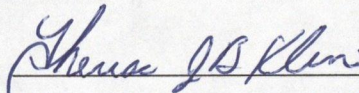
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "A Comprehensive Examination of the Quality of Self-Ratings of Performance" submitted by Lisa Keeping in partial fulfillment of the requirements for the degree of Master of Science.



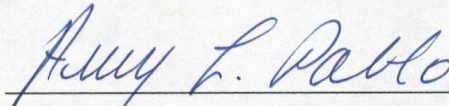
Supervisor, Dr. L. Sulsky, Psychology



Dr. R.E. Franken, Psychology



Dr. T.J.B. Kline, Psychology



Dr. A. Pablo, Management

May 1, 1995
Date

Abstract

The effects of appraisal purpose, validation expectation, and social comparison information on the validity and leniency of self-ratings were investigated on four performance dimensions. A 2 (reward versus research) by 2 (validation expectation versus no expectation) by 2 (social comparison versus no social comparison) experimental design was employed. Rating validity was maximized for all performance dimensions when a reward was expected, social comparison information was provided, and rating validation was expected. For two of the dimensions, ratings were most lenient when the appraisal was conducted for reward purposes, social comparison information was provided, and rating validation was not expected. In contrast, ratings were relatively severe, again when the appraisal was conducted for reward purposes and social comparison information was provided, but when rating validation was expected. Methodological and theoretical issues relating to research on self-ratings of work performance are discussed and directions for future research are proposed.

Acknowledgments

I would like to express my sincere appreciation and gratitude to Dr. Lorne Sulsky, who served as my supervisor for this thesis. His guidance, expertise, and insight helped to create both a paper and an experience far exceeding my expectations. Without his dedication, and most of all, his encouragement and patience, I still might be lost in a mountain of data, and he still might be waiting for that infamous first draft.

I would also like to thank Drs. Theresa Kline and Bob Franken for their encouragement and assistance throughout this entire process.

Finally, I would like to thank Doug Brown, Becky Hooey, and the rest of my friends and colleagues for enduring my moans, groans, and panic and for assuring me that someday I would actually finish this thesis (although not necessarily in a timely manner).

TABLE OF CONTENTS

Approval Page	ii
Abstract.....	iii
Acknowledgments.....	iv
Table of Contents	v
List of Tables.....	vii
List of Figures.....	viii
 Introduction.....	 1
Self-Ratings of Performance.....	3
Advantages Associated with Self-Ratings of Performance.....	4
Disadvantages Associated with Self-Ratings of Performance.....	6
Comparison of Self- and Supervisor Ratings.....	6
Leniency Error and Range Restriction	7
Alternative Levels of Performance.....	8
Validity of Self- and Supervisor Ratings.....	10
Conceptualizing Rating Quality	11
Rating Validity	11
Rating Leniency.....	15
The Present Study.....	21
 Method.....	 24
Subjects.....	24
Procedure	24
Appraisal Purpose Manipulation.....	25
Validation Manipulation.....	25
Social Comparison Manipulation.....	26
Stimulus Materials.....	26
Response Materials.....	27
Self-Evaluations.....	27
Objective Performance Measures.....	27
Dependent Measures.....	28
Rating Validity	28
Rating Leniency.....	28

TABLE OF CONTENTS (Continued)

Results.....	29
Manipulation Check for Social Comparison.....	31
Self-Rating Validity.....	31
Self-Rating Leniency.....	34
Residualized Difference Score Approach.....	34
Leniency for Quantity of Proofreading.....	34
Leniency for Quality of Proofreading.....	37
Leniency for Quantity of Comments.....	37
Leniency for Quality of Comments.....	37
Absolute Approach.....	38
Leniency for Quantity of Proofreading.....	38
Leniency for Quality of Proofreading.....	41
Leniency for Quantity of Comments.....	41
Leniency for Quality of Comments.....	41
Discussion.....	42
Self-Rating Validity.....	44
Self-Rating Leniency.....	45
Implications and Future Research.....	48
Quality Versus Quantity of Performance.....	48
Conceptualizing the Independent Variables.....	50
Relative Versus Absolute Processing.....	52
Training for Self-Raters.....	54
Limitations and Conclusions.....	56
References.....	59

LIST OF TABLES

<u>Table</u>	<u>Title</u>	<u>Page</u>
1	Operationalizations of Self-Rating Validity in the Self-Appraisal Literature.....	12
2	Operationalizations of Self-Rating Leniency in the Self-Appraisal Literature.....	16
3	Means and Standard Deviations of Self-Ratings for the Four Performance Dimensions	30
4	Means and Standard Deviations for Commensurate Objective Scores on Four Performance Dimensions	32
5	Average Correlations Between Self-Ratings and Objective Scores on Four Performance Dimensions for Each Experimental Condition.....	33
6	Means and Standard Deviations for Leniency Using the Residualized Difference Score Approach for the Four Performance Dimensions	35
7	Means and Standard Deviations for Leniency Using the Absolute Approach for the Four Performance Dimensions.....	39

LIST OF FIGURES

<u>Figure</u>	<u>Title</u>	<u>Page</u>
1	Validation Expectation by Social Comparison in the Presence of a Reward for Quantity of Proofreading - Residualized Approach.....	36
2	Validation Expectation by Social Comparison in the Presence of a Reward for Quantity of Proofreading - Absolute Approach.....	40
3	Validation Expectation by Social Comparison in the Presence of a Reward for Quality of Comments - Absolute Approach.....	43

A Comprehensive Examination of the Quality of Self-Ratings of Performance

Performance appraisal is frequently performed in organizations for a variety of purposes, including administrative decisions (e.g., raise, promotion), feedback and development, and personnel research. Thus, performance appraisal is among the most important human resource systems in organizations insofar as it represents critical decisions integral to a variety of human resource actions and outcomes (Judge & Ferris, 1993). Because of its prevalence and importance in organizations, performance appraisal is also one of the most widely researched areas in Industrial-Organizational psychology (Pearce & Porter, 1986).

Most appraisal systems rely heavily upon subjective ratings of an employee's performance, usually completed by the employee's immediate supervisor (Bernardin & Beatty, 1984; Murphy & Cleveland, 1991). Research in the area of supervisor ratings has indicated that there are a number of psychometric problems associated with these ratings, such as halo and rating leniency/severity (Landy & Farr, 1980, 1983).

One common perspective in the extant appraisal research literature is that psychometric rating errors arise from the fact that performance appraisal represents a difficult cognitive task. Consequently, rater limitations in the cognitive processing of ratee performance information lead to reductions in rating quality (e.g., DeNisi, Cafferty, & Meglino, 1984).

A second perspective is that performance rating is a "motivated" behaviour and thus, rating quality becomes compromised insofar as raters intentionally distort their evaluations due to a host of possible motivational factors, such as politics and affect toward ratees (e.g., Banks & Murphy, 1985; Longenecker, Sims, & Gioia, 1987).

A third and final perspective on the psychometric problems associated with performance ratings is that raters have sufficient ability and motivation to produce quality ratings; however, logistical constraints, such as inadequate opportunity to observe ratee performance adversely affect ratings (e.g., Cascio, 1987; Feldman, 1981).

Given the variety of cognitive ability, motivational, and situational factors, all of which may lead to reductions in rating quality, a great deal of research attention has been focused upon the psychometric quality of supervisor ratings, and has examined possible ways of enhancing their validity. Some of the strategies suggested for protecting against the biases and inaccuracies associated with supervisor ratings have included rater training programs and the construction of alternative rating instruments and methods (Cascio, 1987; Landy & Farr, 1980; Murphy & Balzer, 1989; Riggio & Cole, 1992).

Given the potential problems associated with supervisor ratings, some researchers have investigated alternative or additional rating sources. One source which has received some research attention is the employee, engaging in self-appraisals of performance (Bernardin & Klatt, 1985; Harris & Schaubroeck, 1988; Steel & Ovalle, 1984). As is the case with any appraisal rating source, self-ratings may also be associated with potential psychometric problems, such as rating leniency, which has led to the examination of the effects of certain key variables (e.g., appraisal purpose) on self-ratings. The purpose of this study was to examine the quality of self-ratings of performance by consolidating three variables assumed to influence rating quality: the purpose of the appraisal, the expectation that ratings will be validated by an external source, and access to social comparative information.

Although these variables have been investigated individually in previous research (e.g., Farh & Dobbins, 1989a; Farh & Werbel, 1986), to date, no research has investigated their combined effects.

Before presenting the details of the present study, however, a general discussion of self-ratings will be presented, followed by an examination of research associated with the variables pertinent to this thesis. Then, the concept of "rating quality" is considered and the study's central hypothesis is presented.

Self-Ratings of Performance

As mentioned previously, self-ratings of performance have been considered as an alternative or as an addition to the more traditional supervisor ratings. The incorporation of self-ratings into formal performance appraisal systems has become increasingly popular as more organizations have become concerned with such issues as employee participation, satisfaction, and productivity. In addition, a heightened awareness of the potential implications regarding fair employment practices has resulted in an increase in the attention focused on self-ratings and performance appraisal in general (Pearce & Porter, 1986). Thus, there has been a recent surge in research investigating self-ratings of performance (e.g., Arnold & Davey, 1992; Campbell & Lee, 1988; Eder & Fedor, 1989; Farh & Dobbins, 1989a, 1989b; Farh & Werbel, 1986; Farh, Dobbins, & Cheng, 1991; Farh, Werbel, & Bedeian, 1988; Fox & Dinur, 1988; Fox, Caspy, & Reisler, 1994; Furnham & Stringfield, 1994; Harris & Schaubroeck, 1988; Lane & Herriot, 1990; Levy, 1993; Riggio & Cole, 1992; Roberson, Torkel, Korsgaard, Klein, Diddams, & Cayer, 1993; Williams & Levy, 1992).

Researchers examining self-ratings of performance have primarily discussed two different methods of incorporating self-ratings into a formal appraisal system. In one method, the subordinate completes the appraisal form alone (Bassett & Meyer, 1968), which is very rarely used in practice. The second, more common approach, involves independent performance assessments by both supervisor and subordinate (Bernardin & Beatty, 1984; Teel, 1978). Thus, although theoretically regarded as an alternative to supervisor ratings, in practice, self-ratings are generally used in conjunction with ratings from another source, usually supervisor ratings (Campbell & Lee, 1988; Landy & Farr, 1980). Moreover, self-ratings most commonly function in the appraisal system as a focus for the appraisal interview, which occurs subsequent to the actual rating of performance. That is, they facilitate the appraisal interview by serving as a supplemental source of information around which the interview can proceed (Roberson et al., 1993).

Advantages Associated with Self-Ratings of Performance

Researchers have long recognized that self-appraisals can serve several distinct functions in an organization and have suggested numerous advantages associated with their use. For example, proponents for including self-ratings in appraisal systems have contended that no one is more aware of a ratee's performance on the job than the ratee. Thus, employees are better equipped to rate their own performance because they are more knowledgeable about their performance compared to their supervisor (Levine, 1980; Riggio & Cole, 1992).

Second, it has been suggested that self-appraisals may increase ratee participation in the appraisal interview (Farh et al., 1988; Latham & Wexley,

1981), which may make them more committed to performance goals and more accepting of criticism (Riggio & Cole, 1992). Third, it has been proposed that self-appraisals may also reduce subordinate ambiguity regarding performance standards and managerial expectations (Bassett & Meyer, 1968; Farh et al., 1988). Fourth, incorporating self-appraisals, or at least allowing a subordinate to express his/her feelings as part of the appraisal, has been found to promote greater supervisor and subordinate perceptions of fairness, accuracy, acceptance, and satisfaction with the appraisal process, as well as increasing subordinates' motivation to improve performance (Bassett & Meyer, 1968; Dipboye & de Pontbriand, 1981; Greller, 1975; Farh et al., 1988; Landy, Barnes, & Murphy, 1978). Finally, self-appraisals have been found to increase communication between supervisors and subordinates as well as increase employees' sense of control, which are important aspects of perceived procedural justice and fairness (Farh et al., 1988; Folger & Greenberg, 1985).

Although the above noted benefits of self-appraisals have been regarded as intuitively plausible, it should be noted that there have actually been very few experimental investigations of these effects (Campbell & Lee, 1988). Therefore, to date, many of the potential benefits of self-appraisals can only be regarded as speculation (Roberson, et al., 1993). Moreover, in some studies often cited as support for self-appraisals (e.g., Landy, Barnes, & Murphy, 1978) the documented advantages were related to allowing subordinates to express their feelings and/or opinions during the appraisal process, which is not equivalent to actually incorporating self-appraisals into the performance appraisal system.

A study by Roberson et al. (1993) attempted to experimentally confirm

some of the potential benefits of self-appraisals. They found that self-appraisals had no effect on ratee perceptions of their contributions to the appraisal interview or satisfaction with the appraisal. In fact, ratees who performed self-appraisals perceived less influence over the appraisal discussion than did ratees who did not have the opportunity to rate their own performance.

Disadvantages Associated with Self-Ratings of Performance

Despite the potential advantages associated with self-ratings, the primary reason some researchers have argued against their use has been on the grounds that they tend to be more lenient than either supervisor or peer ratings (e.g., Farh & Werbel, 1986; Holzbach, 1978). This apparent disadvantage of self-ratings will be discussed more fully in the following section.

There are other potential disadvantages associated with self-ratings. For example, some employees may not have well developed performance standards, especially newer employees who are relatively unfamiliar with the job. Also, many of the cognitive limitations associated with supervisor ratings are applicable to self-ratings, such as problems with storage, retrieval, and integration of information as well as informational (e.g., role ambiguity) and affective (e.g., defense mechanisms) constraints (Campbell & Lee, 1988).

Comparison of Self- and Supervisor Ratings

The majority of the self-rating research has compared self-ratings to ratings obtained from other sources, such as peers and supervisors. Moreover, many of these studies have examined the correlations between

self- and supervisor ratings (Harris & Schaubroeck, 1988; Mabe & West, 1982). Overall, correlations between self-ratings and supervisor ratings have been found to be moderately low (Harris & Schaubroeck, 1988; Landy & Farr, 1980; Mabe & West, 1982). For example, a recent meta-analysis by Harris and Schaubroeck (1988) found a mean self-supervisor correlation of .35. The somewhat low correlations found between self- and supervisor ratings have led some researchers to hypothesize why these ratings do not converge to a greater extent. Three distinct hypotheses have been invoked to explain the general lack of convergence between these rating sources. A discussion of these hypotheses follows next. Included in the discussion is a consideration of the research relevant to the three variables examined in this study, which are presumed to affect self-ratings: appraisal purpose, validation expectation, and social comparison information.

Leniency Error and Range Restriction

One possibility for the lack of convergence between self- and supervisor ratings is a restriction of range in self-appraisals due to leniency error (Holzbach, 1978). It seems intuitive that ratees might spuriously inflate their self-ratings, especially if increasing the rating level benefits them in some way. A generally consistent inflation in ratings would result in restriction of range, thus making it difficult for self-ratings to correlate highly with supervisor (or any other) ratings.

In a study examining the issue of self-rating leniency, Farh and Werbel (1986) found that subjects rating their class participation for extra course credit produced more lenient ratings than those rating their participation for research purposes. Thus, self-ratings for reward purposes were significantly

higher than those for research purposes. Their results also indicated that when subjects expected their ratings to be validated by an external source, the ratings produced were less lenient than those for subjects without this validation expectation. In addition, Mabe and West (1982) found that even when validation is implied, subjects' self-ratings are less lenient than when there is no expectation of validation at all. Thus, it appears that the expectation of validation and appraisal purpose might be important factors to consider when attempting to reduce self-rating leniency and thereby improve the convergence between self-ratings and other performance measures.

Alternative Levels of Performance

A second possibility suggested as an explanation for the low correlations typically found between supervisor and self-ratings is that supervisors and subordinates possess different levels of performance standards, or unique perspectives, and therefore attend to different aspects of performance (Borman, 1974; Schmitt, Noe, & Gottschalk, 1986; Thornton, 1980). In sum, they lack a common frame of reference regarding what constitutes good and poor performance (Farh & Dobbins, 1989a).

With this idea in mind, Farh and Dobbins (1989a) proposed that one source of performance standards is social comparison information. More specifically, based on Festinger's (1954) social comparison theory, they suggested that information regarding the performance of others may be a better standard against which to evaluate performance than some absolute set of criteria. Festinger proposed that individuals are motivated to form an accurate perception of their own performance because such knowledge helps to predict success and avoid failure in the future. Thus, individuals seek out

information to help with this self-assessment. Ideally, they seek to assess their performance or abilities based on some sort of objective standard, such as the time required to complete a particular task. In the absence of this information, however, they tend to compare themselves with others and use social comparative information as a means of assessing self-performance. A substantial body of research lends support to these basic tenets of social comparison theory (see Sulls & Miller, 1977, for a review).

In this vein, Farh and Dobbins (1989a) argued that in most organizations, supervisors have access to comparative subordinate performance information, while subordinates themselves typically do not. In addition, clearly defined, objective standards of performance, which social comparison theory predicts we consult first, are very rare in most organizations, thus making social comparison information a valuable assessment source. Farh and Dobbins (1989a) hypothesized that social comparative information provided to subordinates would result in higher correlations between self-ratings and supervisor ratings as well as between self-ratings and objective performance measures.

In a laboratory study investigating these hypotheses, subjects were run in groups and performed a proofreading task. After completing the task, they were required to rate themselves on five performance dimensions regarding both the quality and quantity of their proofreading. A second group of subjects were required to review the proofreading work of the first group and rate this work on the same performance dimensions as those used for the self-ratings. The work of the proofreaders was also evaluated according to objective criteria (e.g., the number of lines of text proofread by a subject). Thus, the study included self-ratings, supervisor ratings, and objective scores.

The authors manipulated social comparison information such that proofreading subjects in the social comparison condition were given the opportunity to review the work of others in their group before rating their own work. Those subjects in the no social comparison group rated their work immediately after task completion. Results indicated that, as predicted, the provision of social comparison information resulted in higher correlations between self- and supervisor ratings, as well as between self-ratings and objective scores.

Validity of Self- and Supervisor Ratings

A third explanation for the lack of convergence between self- and supervisor ratings is the possibility that supervisor ratings are not a proper or valid standard of comparison. As Murphy and Cleveland (1991) noted, it is possible that self-ratings are accurate and supervisor ratings are unduly harsh. Thus, any low correlations between self-ratings and supervisor ratings are difficult to interpret because supervisor ratings may not be the optimal criteria against which to evaluate self-ratings (Murphy & Cleveland, 1991). Indeed, as mentioned previously, the extant models of the rating process (e.g., DeNisi et al., 1984), along with various motivational and situational variables, predict a variety of ways in which the quality of supervisor ratings may be compromised. On the other hand, even if leniency error is not a problem with self-ratings, other psychometric problems might plague self-ratings, rendering them invalid for the purposes intended. In sum, the low correlations obtained between supervisor and self-ratings may be due to psychometric problems associated with the supervisor ratings, the self-ratings, or both.

Conceptualizing Rating Quality

Because this study was designed to examine conditions hypothesized to improve the quality of self-ratings of performance, it is useful at this point to consider what is meant by the term "rating quality".

Rating Validity

One common approach for determining rating quality in the performance appraisal literature examining alternative rating sources (e.g., self- versus supervisor ratings) is to examine rating validity. Table 1 presents previous studies that examined self-rating validity in a variety of contexts and indicates the different ways validity has been operationalized. As Table 1 illustrates, validity has most often been assessed by employing a convergent validity strategy, correlating self-ratings with ratings from other sources, such as supervisors or peers (e.g., Farh & Dobbins, 1989a; Fox, Caspy, & Reisler, 1994). The predictive validity and concurrent validity approaches essentially represent the correlation between self-ratings and objective scores (e.g., Fox & Dinur, 1988; Lane & Herriot, 1990; Steel & Ovalle, 1984). A validity index in this context provides information regarding the strength of the relationship between a set of self-ratings and a corresponding set of objective scores.

Self-rating validity has also been examined with a discriminant validity strategy following the approach set out by Campbell and Fiske (1959). While convergent validity assesses convergence between variables that should theoretically correlate, discriminant validity provides an assessment of whether a variable does not correlate significantly with variables from which it should theoretically differ (Anastasi, 1988).

Table 1

Operationalizations of Self-Rating Validity in the Self-Appraisal Literature

Study	Rating Sources	Operational Definition
Farh & Dobbins (1989a)	Self, Supervisor, Objective	Convergent - by dimension and averaged across dimensions
Farh, Werbel, & Bedeian (1988)	Self, Supervisor, Objective	Convergent - by dimension Predictive - regressing ratings on objective scores by dimension
Fox & Dinur (1988)	Self, Peer, Supervisor, Objective	Predictive - by dimension Predictive - regressing objective scores on ratings across dimensions Convergent - by dimension Discriminant
Fox, Caspy, & Reisler (1994)	Self, Peer Supervisor	Convergent - averaged across dimensions

Table 1 Continued

Study	Rating Sources	Operational Definition
Furnham & Stringfield (1994)	Self, Supervisor, Subordinate	Convergent - by dimension Discriminant
Heneman (1974)	Self, Supervisor	Convergent - by dimension Discriminant
Holzbach (1978)	Self, Peer, Supervisor	Convergent - by dimension Discriminant
Lane & Herriot (1990)	Self, Supervisor, Objective	Predictive - by dimension Predictive - regressing ratings on objective scores by dimension
Levine, Flory, & Ash (1977)	Self, Supervisor, Objective	Convergent - by dimension Discriminant
Riggio & Cole (1992)	Self, Subordinate	Convergent - by dimension

Table 1 Continued

Study	Rating Sources	Operational Definition
Steel & Ovalle (1984)	Self, Supervisor, Objective	Convergent - by dimension Discriminant Concurrent - Spearman Rank
Thornton (1968)	Self, Supervisor, Objective	Convergent - by dimension and averaged across dimensions
Williams & Levy (1992)	Self, Supervisor	Convergent - one overall rating

Rating Leniency

A second approach for determining rating quality in the source of appraisal literature is to examine rating leniency/severity. As mentioned previously, leniency has been suggested as a possible explanation for the lack of convergence between self- and supervisor ratings. In this way, it can be considered an "indirect" method for investigating rating validity. There appears to be a general consensus in the literature on the conceptual definition of leniency/severity: leniency/severity refers to the tendency of a rater to be overly lenient or harsh (Sulsky & Balzer, 1988). However, a few alternative operational definitions for leniency/severity have been proposed. Table 2 presents a list of studies which have examined self-rating leniency/severity, along with the operational definitions.

Traditionally, the overall appraisal literature has assessed leniency/severity by examining the absolute level of the ratings or by considering the distance of assigned scores from the scale's midpoint (Saal, Downey, & Lahey, 1980). These traditional indices of leniency/severity may be poor surrogates of rating validity as they contain no standard of comparison against which to examine the true leniency/severity of a rater's ratings (Murphy & Balzer, 1989). For example, a rater may rate himself as a 7 on a 7-point scale, which would be considered lenient by traditional standards (because it is as far away as possible from the midpoint). Without any information regarding this self-rater's actual performance, however, it is difficult to determine whether the rating of "7" is actually lenient or deserved.

Rating leniency/severity has frequently been assessed in the self-rating literature by examining the difference or distance between self-ratings and

Table 2

Operationalizations of Self-Rating Leniency in the Self-Appraisal Literature

Study	Rating Sources	Operational Definition
Arnold & Davey (1992)	Self, Supervisor	Mean Differences by dimension
Farh & Dobbins (1989a)	Self, Supervisor	Squared Difference Scores by dimension
Farh & Dobbins (1989b)	Self, Objective	Residualized Difference Scores
Farh & Werbel (1986)	Self, Observed	Percentile Difference Scores (one dimension) Mean Differences (one (dimension)
Farh, Werbel, & Bedeian (1988)	Self, Supervisor	Mean Differences by dimension
Fox & Dinur (1988)	Self, Supervisor	Mean Differences
Fox, Caspy, & Reisler (1994)	Self	Mean Differences by dimension
Furnham & Stringfield (1994)	Self, Supervisor, Subordinate	Mean Differences by dimension Difference Scores by dimension

Table 2 Continued

Study	Rating Sources	Operational Definition
Heneman (1974)	Self, Supervisor	Mean Differences by dimension
Holzbach (1978)	Self, Peer, Supervisor	Mean Differences by dimension
Levine, Flory, & Ash (1977)	Self, Supervisor, Objective	Mean Differences by dimension
Shore & Thornton (1986)	Self, Supervisor	Mean Differences by dimension
Steel & Ovalle (1984)	Self, Supervisor	Mean Differences by dimension
Thornton (1968)	Self, Supervisor	Mean Differences averaged across dimensions Difference Scores by dimension

Table 2 Continued

Study	Rating Sources	Operational Definition
Williams & Levy (1992)	Self, Supervisor	Mean Difference Score - overall rating Mean Difference - overall rating Subjective Comparisons
Zammuto, London, & Rowland (1982)	Self, Peer, Supervisor	Mean Differences by dimension

ratings from another source. This has primarily been accomplished in one of two ways: either by testing for mean differences between the ratings, or by computing difference scores. As illustrated in Table 2, the mean difference approach has been the most popular method of examining self-rating leniency (e.g., Arnold & Davey, 1992; Farh, Werbel, & Bedeian, 1988; Fox & Dinur, 1988; Fox, Caspy, & Reisler, 1994; Heneman, 1974; Levine, Flory, & Ash, 1977; Shore & Thornton, 1986; Steel & Ovalle, 1984; Zammuto, London, & Rowland, 1982). In this approach, self-ratings are compared to ratings from another source (usually supervisor) and the means are tested for possible statistically significant differences using either parametric (e.g., analysis of variance) or non-parametric (e.g., sign tests) analytical techniques.

Although the mean difference approach attempts to assess the difference between self-ratings and ratings from another source, it is really equivalent to the midpoint approach (see above) when supervisor (or peer) ratings are employed as the standard of comparison. This is because in the absence of a more "objective" standard there is actually no reliable indication of whether or not an individual deserves the self-ratings.

The difference score approach has also been used in the self-appraisal literature as a method for examining the distance between self-ratings and ratings from another source. In this approach, difference scores are computed by subtracting the self-ratings from ratings from another source (e.g., Farh & Dobbins, 1989a; Farh & Werbel, 1986; Furnham & Stringfield, 1994; Williams & Levy, 1992; Thornton, 1968). These difference scores then serve as dependent variables. It turns out that this approach is actually operationally equivalent to the mean difference approach and yields identical results. Thus, with supervisor (or peer) ratings as the comparative standard, there are

no advantages associated with difference scores as the unit of analysis.

In the overall appraisal literature, a more contemporary method of operationally defining leniency/severity is to utilize "true" or comparison scores as the standard of comparison against which to examine self-ratings. These comparison scores are generated from "expert" raters using one of a number of alternative methodologies (see Sulsky & Balzer, 1988 for a more detailed discussion). An analogous method used in the self-appraisal literature is to compute residualized difference scores between self-ratings and "objective" measures of performance (e.g., Farh & Dobbins, 1989b). These residualized difference scores are computed by regressing self-ratings on "objective" performance scores. The variance in the self-ratings unaccounted for by the objective scores represents the residualized difference, which serves as a dependent measure. This operationalization of leniency is sensitive to the distance between the two sets of scores and thus can be conceptualized as an index of rating accuracy (Sulsky & Balzer, 1988). Because validity is a necessary, but not sufficient condition for accuracy, evidence of rating accuracy ostensibly provides excellent evidence for validity (Sulsky & Balzer, 1988). However, because the difference obtained using the residualized approach is based upon predicted scores obtained from group data, this difference measure does not provide a clear indication of the actual difference between self- and objective scores for a given self-rater. An absolute difference measure would accomplish this, and is commonly used in the appraisal literature (Sulsky & Balzer, 1988). Interestingly, this absolute measure has not been adopted in studies investigating self-ratings of performance.

It should be pointed out that the residualized approach attenuates a

problem associated with difference scores, because random error associated with individual raters is not aggregated (cf. Edwards, 1993). However, this approach also removes variance in self-ratings that can be predicted by true performance scores (Farh & Dobbins, 1989b). In sum, the residualized approach would appear to be methodologically superior (see Cronbach & Furby, 1970 for a more detailed discussion) although the absolute difference approach might be preferable on conceptual grounds.

It is interesting to note that validity has recently been conceptualized in the Industrial-Organizational psychology literature as the validity of inferences from a set of scores or ratings (Society for Industrial and Organizational Psychology, Inc., 1987). In this regard, it is possible at the level of operations to achieve the somewhat paradoxical situation whereby a given rater is less lenient/severe (i.e., more accurate) than a second rater, yet certain specific inferences from the first rater's (more accurate) ratings are less valid.

It is apparent then, that both validity and leniency (i.e., accuracy) provide potentially valuable information regarding self-appraisals of performance. Because of this, and because we might expect the leniency (i.e., accuracy) and validity indices to be somewhat independent, both conceptualizations of rating quality were included in this study (see below).

The Present Study

Overall then, self-ratings have generally been advocated by researchers for the wide variety of potential benefits they present, such as increased employee participation and motivation, as well as increased perceived satisfaction, fairness, accuracy, and acceptance of the appraisal by supervisors and subordinates, to name a few (Bassett & Meyer, 1968; Farh et al., 1988;

Latham & Wexley, 1991; Landy, Barnes, & Murphy, 1978; Meyer, 1991; Riggio & Cole, 1992). In the vast majority of studies investigating self-ratings, however, they have not been found to correlate strongly with supervisor ratings, although there have been exceptions (e.g., Heneman, 1974). One immediate interpretation of this result might be that the quality of self-ratings of performance is generally poor. As noted above, one explanation for this lack of convergence is the possibility that the quality of supervisor ratings as the standard for comparison is questionable in some instances. Without supervisor ratings, what should serve as the proper standard for comparison? In short, how can we better examine the quality of self-ratings of performance? Moreover, what factors are likely to influence rating quality in the first place?

The purpose of this study was to attempt to provide preliminary answers to these questions. To this end, "objective" measures of performance have been employed as the standards of comparison for self-ratings of performance. Thus, unlike the majority of previous self-rating studies, which have correlated self-ratings with ratings from other subjective sources, the present study employed a more objective standard for comparison (cf. Farh & Dobbins, 1989a). This approach is consistent with current appraisal research which assesses rating accuracy by employing a standard of comparison presumed to be valid (i.e., true or comparison scores) (see Sulsky & Balzer, 1988 for a more detailed discussion).

In an attempt to further understand the quality of self-ratings, I examined the effects of appraisal purpose, validation expectation, and social comparison information, as well as their interactions, on rating leniency and validity. This approach has the potential to contribute to the literature on

self-ratings for both theoretical and methodological reasons. First, as mentioned previously, the variables of appraisal purpose, validation expectation, and social comparative information have been shown to significantly affect self-ratings of performance, individually. This study was designed to contribute to the growing body of knowledge concerning factors that potentially affect the quality of self-ratings by consolidating these three variables in one framework. Thus, the possible individual and interactive effects of these variables could be explored. Second, the present study computed an index of validity incorporating objective scores as the standard for comparison. In addition, leniency was computed in two ways: (a) using the residualized difference score approach, utilized by Farh and Dobbins (1989b), and (b) by computing an absolute leniency/severity index for each rater. As noted earlier, this latter approach has not been adopted in previous self-rating research, although it provides a potentially useful index of discrepancy from objective scores for each self-rater.

Although the present study was designed to be primarily exploratory in nature, previous self-appraisal research examining appraisal purpose, validation expectation, and social comparison information provides some indication of what might be expected when all of the variables are considered in combination. Thus, a three-way interaction was hypothesized such that rating quality would be maximized when the ratings are conducted for research purposes, when validation is expected, and when social comparison information is provided. This hypothesis was formulated in light of previous research indicating that rating quality is improved when social comparison information is provided, when subjects expect their ratings to be validated, and when the appraisal is conducted for research purposes, rather

than for a reward (e.g., Farh & Dobbins, 1989a; Farh & Werbel, 1986).

Method

Subjects

Subjects consisted of 147 students enrolled in undergraduate psychology courses at The University of Calgary who volunteered for participation. The data for 24 of these subjects was excluded due to either the failure of the reward manipulation (see below), or due to incomplete data. Thus, the data collected for 123 subjects (88 females, 35 males) were utilized for the present study.

Procedure

Subjects were randomly assigned to one of eight experimental cells (see manipulations below) in a 2 (reward/research) \times 2 (validation expectation/no validation expectation) \times 2 (social comparison/no social comparison) factorial design, and were run in groups of three, four, or five. Subjects were told that the study was investigating the cognitive components of reading and reading error detection. The task involved proofreading a series of articles for a period of 20 minutes. Specifically, subjects were instructed to correct typographical and grammatical errors, and to evaluate each article on a 7-point scale, supporting this evaluation with comments. Subjects were informed that they were being asked to rate each article because their data would be used by the experimenter to determine the quality of the articles without having to actually read them. The article evaluation scale ranged from 1, representing very low quality to 7, representing very high quality. At this point subjects were asked if they were still willing to participate in the

study. Those still interested in participating signed informed consent forms and proceeded to complete the task. When the proofreading session was completed, subjects evaluated their work on four performance dimensions. Regardless of the experimental condition to which they had been assigned, all subjects were fully debriefed following completion of the study.

Appraisal Purpose Manipulation

Subjects were randomly assigned to either a reward or research (nonreward) condition. In the reward condition, subjects were told that because the task was somewhat monotonous, it was felt that an incentive (\$100 to top performer) should be offered to motivate subjects to carefully attend to the task. They were asked to rate themselves on the performance dimensions to provide the experimenter with an indication of their performance as well as to determine reward allocation. Subjects were informed that in the event of a tie, reward allocation would be determined by a lottery involving only those subjects who were tied for top performance. Subjects in the research condition were simply asked to evaluate themselves with no mention of a reward.

Subjects in the reward condition were asked, after debriefing, to indicate whether or not they believed a reward really existed. Data obtained from subjects who did not believe that a reward actually existed were excluded from the analyses.

Validation Manipulation

Subjects were randomly assigned to either the validation expectation or no expectation condition, prior to beginning the task. In the validation

expectation condition, subjects were told that, although the experimenter would not have time to check each individual's work, and thus would have to rely on their self-ratings, some of their work would be spot checked by the experimenter to ensure that it was carefully attended to. Subjects in the no validation expectation condition were told that their performance would be determined based solely on their ratings, because the experimenter would not have time to check over the large amount of information proofread.

Social Comparison Manipulation

Subjects were randomly assigned to either the social comparison or no social comparison condition, subsequent to the 20 minute work session. In the social comparison condition, subjects were given four minutes to look at the work of each member in their group (e.g., for a group of four, 12 minutes total), before rating themselves on the performance dimensions. Subjects in the no social comparison condition did not review their peers' work; rather, they performed a filler reading task prior to rating themselves. The filler task involved a series of magazine articles which subjects were told to read as fast as they could until the experimenter told them to stop.

Stimulus Materials

Stimulus materials were adopted from Miceli (1985). They consisted of six articles constructed for the task of proofreading, which contained 440 total lines and 123 possible errors. The articles were generally representative of typical magazine articles and ranged in content from a technical piece on the Metropolitan Museum of Art to a human interest story on a baseball player.

Response Materials

Response materials were adopted from Farh and Dobbins (1989a). Subjective rating scales for four performance dimensions were included for self-evaluation, as well as objective ratings for each of these dimensions. Both the subjective and objective rating dimensions are outlined below.

Self-evaluations. Subjects rated themselves on four performance dimensions: (a) Quantity of proofreading: defined as the number of pages proofread by the subject during the 20 minute session; (b) Quality of proofreading: defined as the average number of mistakes per page, detected by the subject; (c) Quantity of editorial comments: defined as the total number of comments made by the subject as support for his/her evaluation of the article; and (d) Quality of comments: defined as the relevance of the subject's comments and the extent to which these comments would help the experimenter determine the quality of the article. Each dimension was rated on a 7-point likert-type scale ranging from 1, indicating extremely low performance, to 7, indicating extremely high performance, with 4 representing the midpoint. These dimensions were verbally defined for subjects prior to the self-rating task.

Objective performance measures. Four objective performance measures were calculated to provide objective criteria against which to compare self-ratings¹. These included the following: (a) Quantity of

¹ The objective measures were calculated on different scales than the self-evaluations. An attempt was made to make the self-evaluations as parsimonious as possible for subjects as the task involved a subjective evaluation of their performance and was not meant to involve calculations. The definitions given for each dimension were meant to serve as guidelines for evaluation, not to be answered with exact numbers. In contrast, the objective measures were calculated as accurately as possible, to try and obtain a clear indication of a subject's true performance. Thus, it was necessary to calculate at a more micro level. For example, quantity of proofreading was calculated as the number of lines proofread, rather than the number of pages proofread to avoid making any subjective decisions regarding partially read pages. In

proofreading: calculated as the total number of lines proofread by the subject during the 20 minute session; (b) Quality of proofreading: calculated as the average number of errors detected by the subject per 20 lines of proofread material; (c) Quantity of comments: calculated as the total number of words written by the subject in support of his/her evaluation²; and (d) Quality of comments: Comments were separately rated by two psychology graduate students on a 7-point scale with 1 representing few one-word remarks not related to the stories, 4 representing comments about grammar, sentence structure, typographical, and spelling errors, and 7 representing comments about the story content (e.g., lack of organization, incoherent arguments, unsupported inferences). The initial interrater agreement, as indicated by an intraclass correlation was .65. The graduate students then met to discuss any discrepancies and reached consensus on each rating.

Dependent Measures

Rating validity. The validity of self-ratings was operationalized as convergent validity, consistent with previous research (e.g., Farh & Dobbins, 1989a). Specifically, the correlation between self-ratings and objective scores was calculated for each of the eight experimental conditions, separately for each performance dimension.

Rating leniency. Leniency was conceptualized as the distance between self-ratings and objective scores. First, leniency was operationalized as the

addition, the number of potential errors detected varied directly with the number of lines proofread, not the number of pages proofread.

² The total number of words was used to operationally define quantity of comments because subjects were instructed to support their article evaluations with comments and a quantitative measure was required to distinguish between those subjects who simply jotted down one-word comments and those who wrote more extensively in defense of their evaluation.

residualized difference score between self-ratings and objective scores. The residualized difference scores were computed separately for each dimension by regressing the self-ratings on the objective scores. As mentioned previously, the residualized difference score represents the variance in the self-rating unaccounted for by the objective score. Positive values indicate that the self-ratings are higher than what would be predicted from the objective scores; negative values indicate that the self-ratings are less than what would be predicted from the objective scores.

Second, leniency was computed using an absolute measure such that all objective scores were transformed to scales commensurate with those used for the self-ratings (i.e., 7-point scales). For example, to transform the number of lines proofread (objective measure for quantity of proofreading) to a 7-point scale, the maximum number of possible lines proofread (440) was equated with a score of 7, and all other objective scores for quantity of proofreading were correspondingly transformed using this calibration. If an objective score for any of the four performance dimensions transformed to a score of less than 1 on a 7-point scale, it was set to 1 because 1 was the lowest possible rating used in the self-rating task. The objective score was then subtracted from the self-rating for each rater on each dimension.

Results

Table 3 presents the means and standard deviations of the self-ratings for each condition. Although the actual analyses were performed on rating validity and leniency, the raw rating scores and their standard deviations are presented for perusal of the mean rating levels. As Table 3 illustrates, many of the mean ratings across conditions are above the scale midpoint (i.e., 4).

Table 3

Means and Standard Deviations of Self-Ratings for the Four
Performance Dimensions

Condition	Proofreading				Comments				
	Quantity	Quality		Quantity		Quality		M	SD
		M	SD	M	SD	M	SD		
Reward									
Validation									
Compare	4.61	1.50	5.23	1.16	4.00	1.63	4.08	1.55	
No Compare	4.56	1.15	5.75	1.00	4.25	1.24	4.75	1.34	
No Validation									
Compare	5.27	0.96	5.67	1.04	4.83	1.58	5.07	1.39	
No Compare	4.15	0.90	5.69	0.75	4.85	1.07	4.92	1.32	
No Reward									
Validation									
Compare	5.05	1.35	5.26	1.05	4.05	1.13	4.29	0.96	
No Compare	3.87	0.74	4.93	1.03	4.13	1.25	3.87	1.06	
No Validation									
Compare	4.82	1.29	5.00	1.22	4.03	0.94	4.18	1.29	
No Compare	4.07	1.03	5.07	1.10	3.87	1.35	4.50	1.37	

Table 4 presents the means and standard deviations of the transformed objective scores for each condition.

Manipulation Check for Social Comparison

Subjects were asked to estimate the number of pages proofread and the number of errors detected by the average person in their group. These estimations and the objective data for quantity and quality of proofreading for the average group member, respectively, were transformed to z-scores. Discrepancy scores were then computed for average pages proofread and average number of errors detected. Subjects provided with social comparison information were significantly more accurate in their estimations than subjects not provided with social comparison information, for both average number of pages proofread ($t(37) = -1.62, p < .05$) and average number of errors detected ($t(37) = -1.62, p < .05$). Thus, it appears that subjects provided with social comparison information had a more accurate perception of the average performance of their peers than did those who were not provided with this comparative information.

Self-Rating Validity

Table 5 contains the correlations found between the self-ratings and corresponding objective scores on each performance dimension for each experimental condition. It was hypothesized that rating validity would be maximized when there was no reward, there was a validation expectation, and social comparison was provided. However, at odds with this prediction was the finding that for each performance dimension, the highest correlation (i.e., highest validity) was obtained in the condition where the appraisal was

Table 4

Means and Standard Deviations for Commensurate Objective Scores
on Four Performance Dimensions

Condition	Proofreading				Comments			
	Quantity		Quality		Quantity		Quality	
	M	SD	M	SD	M	SD	M	SD
Reward								
Validation								
Compare	3.94	1.84	3.88	0.80	3.05	1.74	3.15	1.67
No Compare	2.73	0.92	3.71	0.80	3.75	1.93	3.25	1.81
No Validation								
Compare	2.97	0.65	3.69	0.85	4.05	2.06	2.47	1.60
No Compare	2.82	1.32	3.83	0.72	4.81	1.30	3.85	1.28
No Reward								
Validation								
Compare	3.21	1.23	4.07	0.87	3.25	1.24	3.31	1.06
No Compare	2.75	0.82	3.75	0.86	3.62	1.41	3.07	1.53
No Validation								
Compare	3.03	0.85	3.36	0.79	4.24	1.55	2.88	1.45
No Compare	3.01	0.98	3.62	1.0	3.72	1.86	2.53	1.30

Table 5

Average Correlations Between Self-Ratings and Objective Scores on Four Performance Dimensions for Each Experimental Condition

Condition	Proofreading		Comments	
	Quantity	Quality	Quantity	Quality
Reward				
Validation				
Compare	.7819**	.4437	.8712**	.6678**
No Compare	.7570**	-.0075	.5237*	.4124
No Validation				
Compare	.4748*	.2091	.6899**	.3718
No Compare	.0482	.0600	.5434*	.1895
No Reward				
Validation				
Compare	.5897**	.3609	.1733	.5335**
No Compare	.6525**	.3672	.1802	.3575
No Validation				
Compare	.6304**	-.0119	.5993**	.3128
No Compare	.4764*	.1601	.5051*	.1794

* $p < .05$. ** $p < .01$. (one-tailed)

used for reward purposes, subjects expected their ratings to be validated, and social comparative information was present (see Table 5).

Self-Rating Leniency

Residualized Difference Score Approach

The means and standard deviations for leniency using the residualized difference score approach for each performance dimension are given in Table 6.

Four 2 (reward versus research) X 2 (validation expectation versus no validation expectation) X 2 (social comparison information versus no social comparison information) univariate analyses of variance (ANOVAs) were performed on the residualized difference scores, one for each of the four performance dimensions.

Leniency for quantity of proofreading. For quantity of proofreading, the three-way interaction between appraisal purpose, validation expectation, and social comparative information was significant ($F(1,115)=7.68$, $p<.01$, $\eta^2=.053$). Simple two-way interactions were then examined between validation expectation and social comparison information within each of the two reward conditions. In the presence of a reward, the two-way interaction between validation expectation and social comparison information was significant ($F(1,53)=11.74$, $p<.001$, $\eta^2=.173$). Figure 1 illustrates this simple two-way interaction. Subjects were most lenient when social comparison information was present and there was no expectation that their ratings would be validated. Subjects were less lenient when social comparison information was not present, regardless of whether or not they expected their ratings to be validated. Subjects were least lenient and

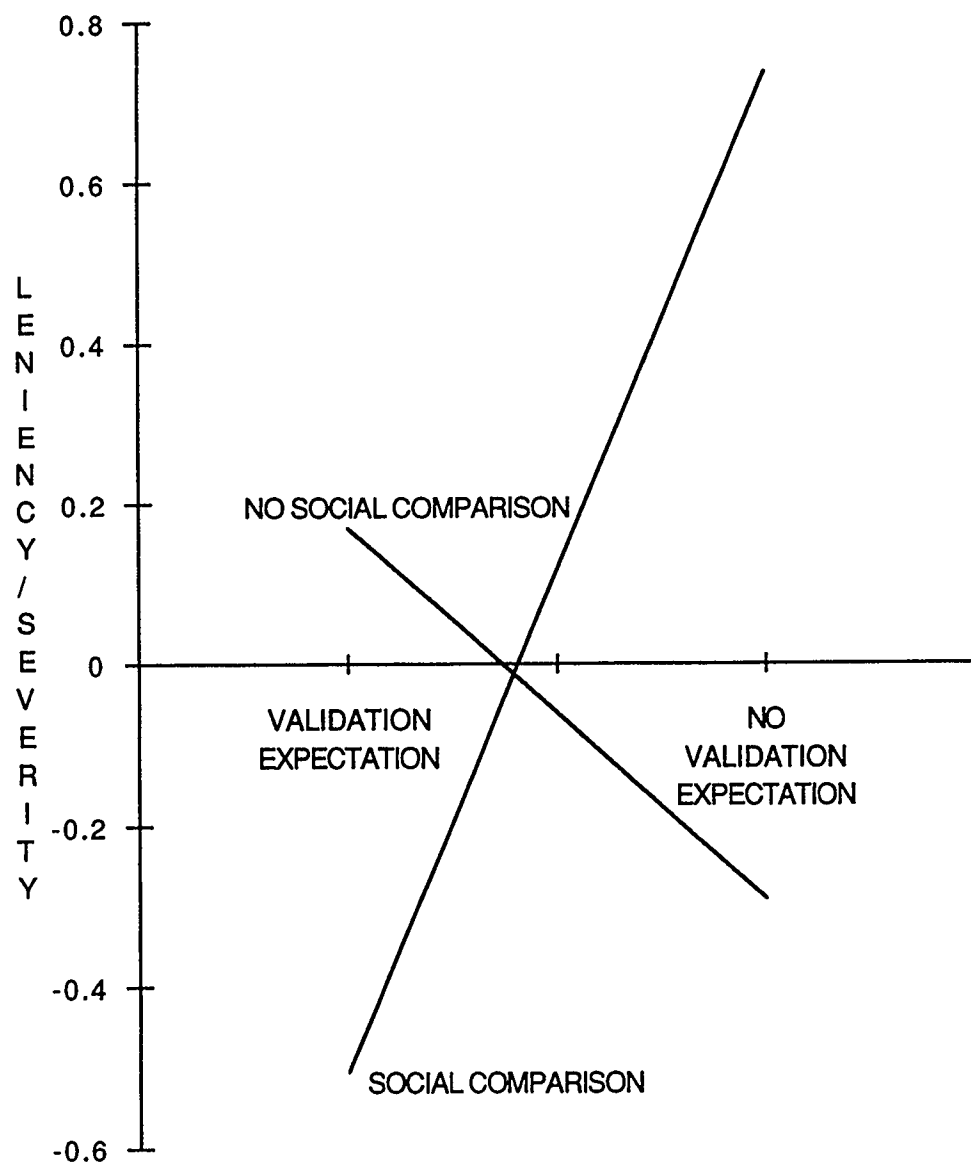
Table 6

Means and Standard Deviations for Leniency Using the Residualized
Difference Score Approach for the Four Performance Dimensions

Condition	Proofreading				Comments			
	Quantity		Quality		Quantity		Quality	
	M	SD	M	SD	M	SD	M	SD
Reward								
Validation								
Compare	-0.51	0.94	-0.08	1.11	0.08	1.07	-0.40	1.27
No Compare	0.17	0.83	0.47	1.01	0.05	1.06	0.24	1.22
No Validation								
Compare	0.74	0.85	0.35	1.02	0.46	1.15	0.80	1.29
No Compare	-0.29	1.12	0.42	0.75	0.19	0.90	0.23	1.30
No Reward								
Validation								
Compare	0.40	1.10	-0.17	0.99	0.05	1.15	-0.24	0.83
No Compare	-0.54	0.56	-0.36	0.98	-0.02	1.27	-0.58	0.99
No Validation								
Compare	0.26	1.06	-0.29	1.24	-0.37	0.75	-0.22	1.22
No Compare	-0.48	0.91	-0.25	1.09	-0.36	1.18	0.21	1.36

Figure 1

Validation Expectation by Social Comparison in the Presence of a Reward for
Quantity of Proofreading - Residualized Approach



somewhat severe in the presence of social comparison information, when they expected their ratings to be validated.

Simple simple effects tests indicated that in the presence of a reward, the difference between the means when social comparison information was present ($M=-.506$) and when it was absent ($M=.169$) was significant, when subjects expected their ratings to be validated ($F(1,27)=4.20$, $p<.05$, $\eta^2=.135$).³ In the absence of validation expectation, there was also a significant difference between the means when social comparison information was present ($M=.738$) and when it was absent ($M=-.291$), ($F(1,26)=7.58$, $p<.01$, $\eta^2=.226$).

In the absence of a reward, the simple effect of social comparison information was significant ($F(1,62)=12.87$, $p<.001$, $\eta^2=.171$) such that ratings were lenient in the presence of social comparison information ($M=.329$) and quite harsh in the absence of social comparison information ($M=-.512$).

Leniency for quality of proofreading. For quality of proofreading, a significant main effect for appraisal purpose was found ($F(1,115)=8.61$, $p<.005$, $\eta^2=.068$), such that subjects were lenient when a reward was present ($M=.289$) and were somewhat severe when it was absent ($M=-.264$). No other effects were significant ($p>.05$).

Leniency for quantity of comments. For quantity of comments, there were no significant main effects or interactions obtained ($p>.05$).

Leniency for quality of comments. The three-way interaction between appraisal purpose, validation expectation, and social comparative information was significant for quality of comments ($F(1,115)=5.32$, $p<.05$, $\eta^2=.041$). Simple two-way interactions were then examined between

³ Some researchers might choose to use a more conservative approach and employ the Bonferroni adjustment. In this case such an approach would render this finding nonsignificant.

validation expectation and social comparison information for each of the two reward conditions. Neither of these two-way interactions was significant, however, nor were any of the simple effects ($p > .05$), rendering the three-way interaction uninterpretable.

Absolute Approach

The means and standard deviations for the four independent leniency measures using the absolute approach for computing leniency are given in Table 7.

Similar to the residualized difference score approach, four 2 (reward versus research) X 2 (validation expectation versus no validation expectation) X 2 (social comparison information versus no social comparison information) ANOVAs were performed, one for each of the four performance dimensions, with absolute leniency serving as the dependent variable.

Leniency for quantity of proofreading. As with the residualized difference score approach, a significant three-way interaction was found for quantity of proofreading ($F(1,115)=7.85$, $p < .005$, $\eta^2=.055$) and simple two-way interactions were then examined for each of the two reward conditions. In the presence of a reward, a significant two-way interaction between validation expectation and social comparison information ($F(1,53)=15.78$, $p < .001$, $\eta^2=.192$) was obtained. Figure 2 illustrates this simple two-way interaction. Ratings were most lenient when subjects did not expect their ratings to be validated and social comparison information was present. Subjects were less lenient when no social comparison information was present, regardless of whether or not they expected their ratings to be validated. Subjects were least

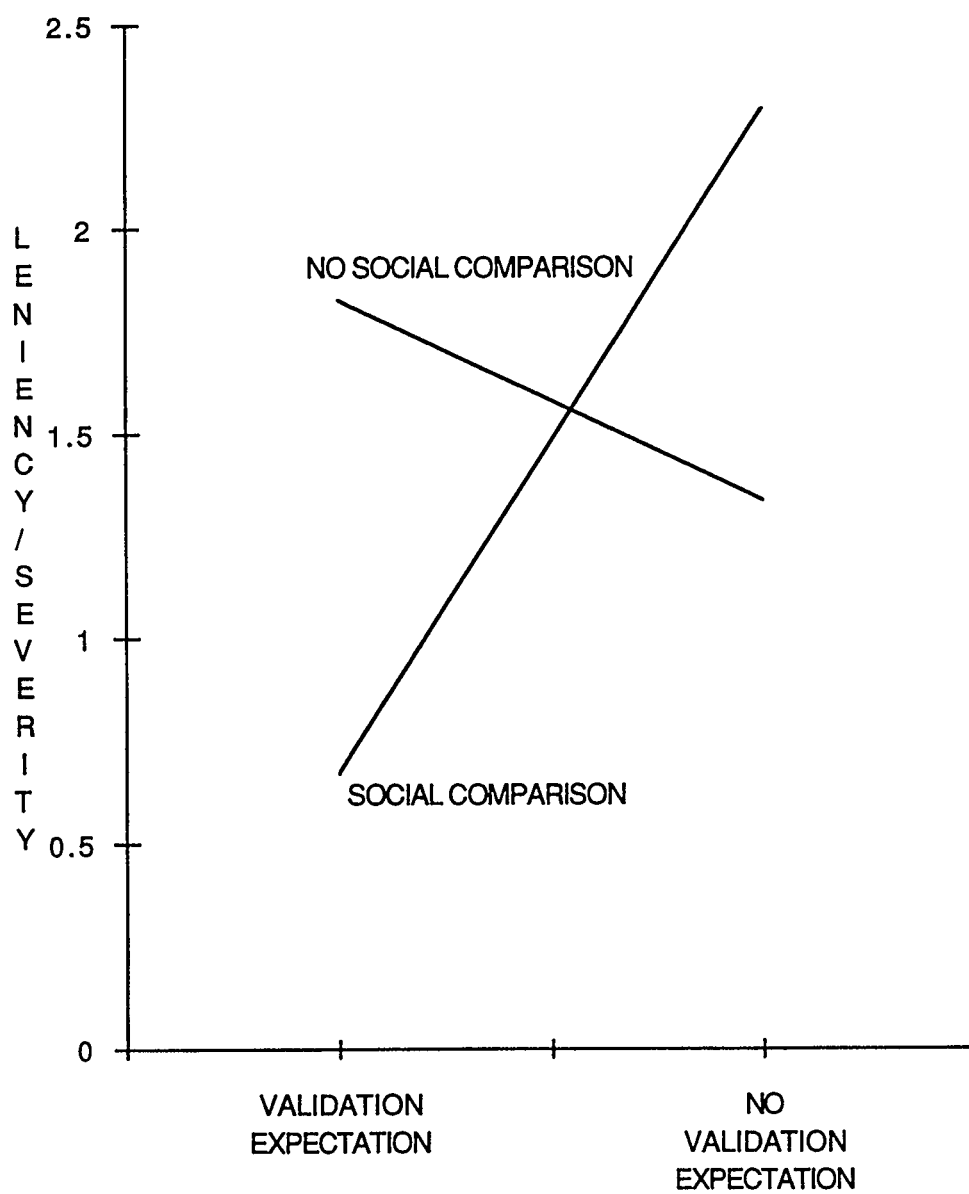
Table 7

Means and Standard Deviations for Leniency Using the Absolute
Approach for the Four Performance Dimensions

Condition	Proofreading				Comments			
	Quantity		Quality		Quantity		Quality	
	M	SD	M	SD	M	SD	M	SD
Reward								
Validation								
Compare	0.67	1.05	1.35	0.99	0.94	0.87	0.92	1.32
No Compare	1.83	0.76	2.03	1.30	0.50	1.66	1.50	1.75
No Validation								
Compare	2.29	0.87	1.97	1.18	0.78	1.52	2.60	1.68
No Compare	1.34	1.56	1.86	0.94	0.04	1.15	1.08	1.66
No Reward								
Validation								
Compare	1.84	1.18	1.19	1.08	0.80	1.53	0.97	0.98
No Compare	1.12	0.66	1.18	1.08	0.51	1.71	0.80	1.52
No Validation								
Compare	1.79	1.00	1.63	1.31	-0.22	1.24	1.29	1.61
No Compare	1.05	1.03	1.44	1.24	0.15	1.64	1.97	1.72

Figure 2

Validation Expectation by Social Comparison in the Presence of a Reward for
Quantity of Proofreading - Absolute Approach



lenient when they expected that their ratings might be validated and social comparison information was present.

Simple simple effects tests revealed that in the presence of a reward, there was a significant difference between the means for the social comparison ($\underline{M}=1.671$) and no social comparison ($\underline{M}=1.83$) conditions when subjects expected their ratings to be validated ($F(1,27)=11.89$, $p<.005$, $\eta^2=.306$). There was also a significant difference found between the means for the social comparison ($\underline{M}=2.29$) and no social comparison ($\underline{M}=1.34$) conditions when there was no expectation of validation ($F(1,26)=4.17$, $p<.05$, $\eta^2=.138$).

In the absence of a reward, there was a significant simple effect found for social comparison information ($F(1,62)=8.77$, $p<.005$, $\eta^2=.124$), such that ratings were more lenient in the presence of social comparison information ($\underline{M}=1.82$) than in its absence ($\underline{M}=1.08$). In sum, for this dimension, the results for absolute leniency corroborated the results when the residualized difference score served as the dependent variable.

Leniency for quality of proofreading. As with the residualized difference score approach, a significant main effect for appraisal purpose was found for quality of proofreading ($F(1,115)=4.46$, $p<.05$, $\eta^2=.036$), such that ratings were more lenient in the presence of a reward ($\underline{M}=1.80$), than when the appraisal was performed for research purposes ($\underline{M}=1.36$).

Leniency for quantity of comments. There were no significant main effects or interactions found for quantity of comments ($p>.05$), replicating the findings when the residualized difference approach was used to calculate leniency.

Leniency for quality of comments. For quality of comments, a significant three-way interaction was found ($F(1,115)=6.96$, $p<.01$, $\eta^2=.053$)

and simple two-way interactions were examined for each of the two reward conditions. In the presence of a reward, a significant two-way interaction between validation expectation and social comparison information ($F(1,53)=15.58, p<.05, \eta^2=.095$) was obtained. Figure 3 illustrates this simple two-way interaction. Ratings were most lenient when subjects did not expect their ratings to be validated and social comparison information was present. Subjects were less lenient when no social comparison information was present, regardless of whether or not they expected their ratings to be validated. Subjects were least lenient when they expected that their ratings might be validated and social comparison information was present.

Simple simple effects tests revealed that in the presence of a reward, there was a significant difference between the means for the social comparison ($M=2.60$) and no social comparison ($M=1.08$) conditions when subjects did not expect their ratings to be validated ($F(1,26)=5.79, p<.05, \eta^2=.182$). When subjects did expect their ratings to be validated, however, there was no significant difference obtained between the means for the social comparison and no social comparison conditions.

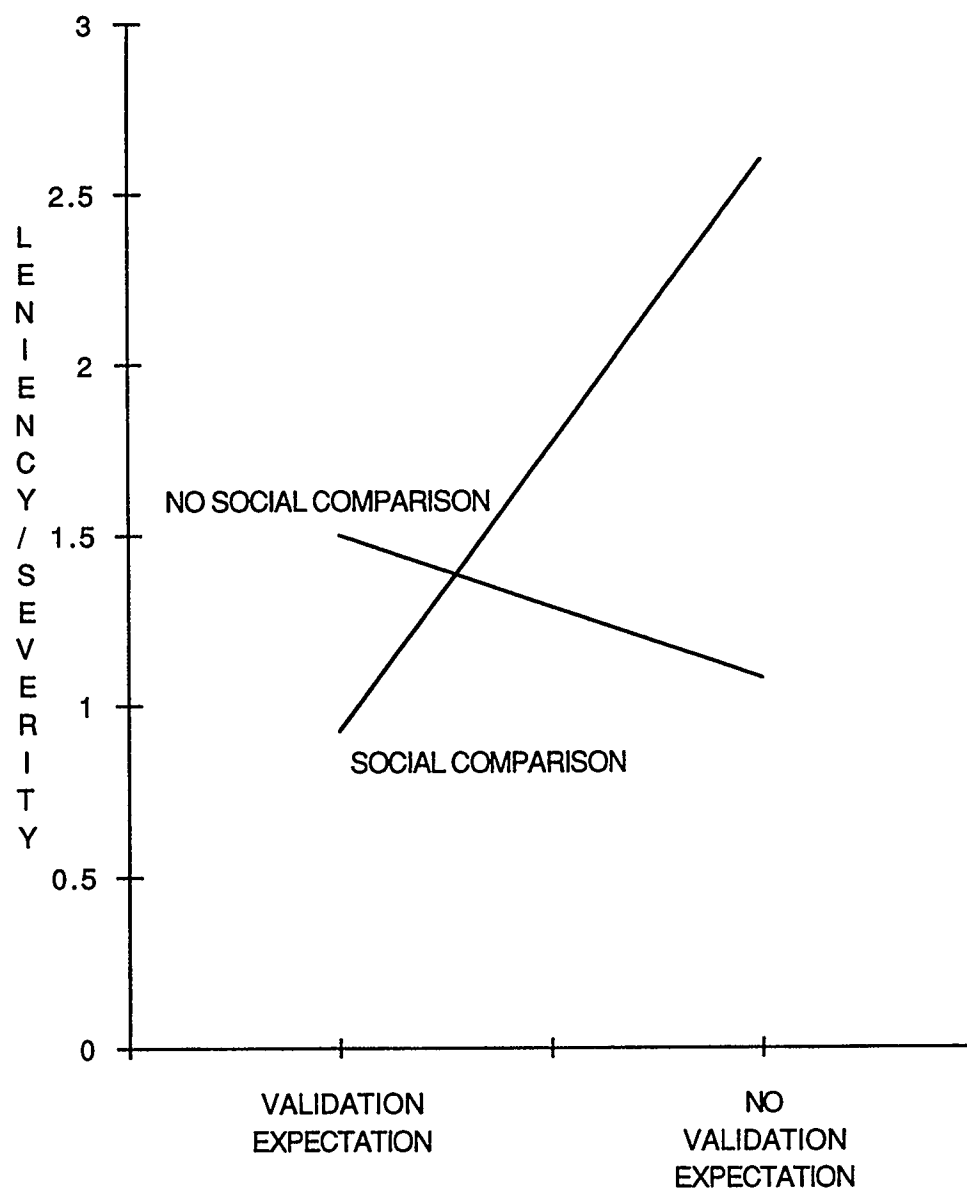
In the absence of a reward, there was a significant simple effect found for validation expectation ($F(1,62)=4.23, p<.05, \eta^2=.062$), such that ratings were more lenient in the absence of a validation expectation ($M=1.63$) and less lenient in its presence ($M=.887$).

Discussion

The primary purpose of the present study was to examine the combined effects of appraisal purpose, validation expectation, and social comparison information on the validity and leniency of self-ratings of

Figure 3

Validation Expectation by Social Comparison in the Presence of a Reward for
Quality of Comments - Absolute Approach



performance. The results pertaining to each of these measures of rating quality will be discussed separately, followed by a discussion of research implications and suggestions for future research.

Self-Rating Validity

Results indicated that self-rating validity was maximized when the appraisal was performed for reward purposes, when rating validation was expected, and when social comparison information was provided. This finding was not consistent with the prediction that the most favourable condition for validity would be when the appraisal is performed for research purposes, when subjects expect validation, and social comparison information is provided. This result also does not correspond with previous research which has indicated that self-ratings performed for reward purposes are less valid than those for research purposes (e.g., Farh & Dobbins, 1989a)

The most valid ratings were obtained in the cell where the appraisal was performed for reward purposes, validation was expected, and social comparison information was provided. It is of interest that this cell also evidenced somewhat severe ratings (see Tables 4 and 6). It appears as though the presence of a reward sensitized subjects to the validation expectation and social comparison variables. That is, when validation was expected, the need to be as accurate as possible in their self-assessments probably became more salient, and social comparison information helped to increase the validity of the ratings (and also decrease leniency), presumably because it provided raters with some kind of standard with which to make their evaluations.

Self-Rating Leniency

It has been well established that subjects are more lenient when an appraisal is performed for reward purposes rather than research purposes and, similarly, when subjects do not expect their ratings to be validated by an external source (rather than when they do have this validation expectation) (e.g., Farh & Werbel, 1986). Farh and Dobbins (1989a) found that providing subjects with social comparative information reduces rating leniency, but they only examined the effect of this variable when the appraisal was conducted for research purposes.

The present study attempted to go a step further and combine the variables of appraisal purpose, validation expectation, and social comparison information. As hypothesized, the three-way interaction between appraisal purpose, validation expectation, and social comparison information was significant, although only for the dimensions of quantity of proofreading and quality of comments. Further, significance was obtained for the latter only when leniency was operationalized using the absolute difference score approach. Arguably, the most interesting finding was the pattern of results obtained for the significant two-way interaction between validation expectation and social comparison information in the presence of a reward for these dependent measures. This pattern indicated that in the presence of a reward, self-ratings were most lenient when social comparison information was provided and rating validation was not expected. In contrast, the least lenient ratings were obtained, also in the presence of social comparison information, but when rating validation was expected. Thus, in the reward condition, both the most and the least lenient ratings were obtained in the presence of social comparison information.

The above findings were unexpected based upon the results of Farh and Dobbins (1989a) who found that ratings were less lenient in the presence of social comparison information than in its absence. Because they did not manipulate appraisal purpose and validation expectation, however, perhaps a fairer comparison would be to examine the current findings when no reward was present. The findings for the quantity of proofreading dimension indicate a significant simple effect for social comparison information such that ratings are more lenient in the presence of social comparison information than in its absence. This finding is also inconsistent with the results of Farh and Dobbins (1989a), but supports Meyer's (1980) contention that "when self-appraisals are obtained on a 'compared to others' basis, the leniency error will be very strong" (p. 295).

Based upon the data, it appears that although subjects possessed a more accurate perception of the performance of their peers when provided with social comparison information (as indicated by the manipulation check), variables such as appraisal purpose and validation expectation seemed to affect how they utilized this comparative information to form their ratings.

One explanation for the obtained significant two-way interaction is that in the presence of a reward and social comparison information, subjects accurately perceived the performance of their peers and therefore knew the level of performance they needed to surpass to obtain the reward. When they did not expect their ratings to be validated, they considered the comparative information and rated themselves leniently in an attempt to gain the reward. Conversely, when there existed an expectation of rating validation, subjects were relatively severe in their self-ratings because they knew anyone validating their ratings would have access to the same comparative

information and thus, they did not want to appear dishonest.

The logic of the above argument would appear to hold only for those subjects who performed poorly in comparison to the top performer in their group. Because in the present study only one person was the top performer for any group, the remaining subjects' performance would have been inferior (which subjects presumably were able to detect according to the manipulation check). Thus, the majority of subjects in the reward condition would have been in a situation where someone else outperformed them and, if this was correctly perceived, could have led to either rating inflation or deflation. The direction of the level bias would have depended on the presence/absence of a validation expectation.

Also at odds with the present results are findings reported by Farh and Werbel (1986). They examined the effects of appraisal purpose and validation expectation and found significant main effects for both variables, but did not obtain a significant interaction between the two. A direct comparison of their results to those of the present study can be made by examining the data when the significant three-way interaction found for quantity of proofreading is split by social comparison information rather than by appraisal purpose. In the situation where subjects were not provided with social comparison information, no significant interactions and one main effect was found. For the quantity of proofreading dimension, a significant main effect for appraisal purpose was obtained such that ratings were more lenient in the presence of a reward than in its absence (when leniency was operationalized using the residualized approach). Although these results do not replicate those of Farh and Werbel (1986), they suggest that social comparison information is an important variable to consider when examining self-ratings.

Implications and Future Research

Quality versus quantity of performance. An examination of the validity coefficients reported in Table 5 conveys that overall, a greater number of significant correlations were obtained for the quantity dimensions than for the quality dimensions. Thus, it appears that overall, ratings were more valid for the quantity dimensions. This introduces a potentially interesting distinction between the rater demands of rating performance quality versus performance quantity, a topic which is rarely examined in the appraisal literature (see Hoffman, Nathan, & Holden, 1991). From an ability perspective, perhaps it was easier for subjects to accurately assess their performance in terms of quantity rather than quality. If one considers the task of proofreading, it seems logical that it would be easier to judge oneself in terms of the number of pages proofread and number of comments written rather than errors detected per page and the "quality" of comments written. Quality is more subjective, while quantity is easier to discern, if not entirely objective.

From a motivational perspective, perhaps subjects were less likely to inflate their ratings on quantity dimensions because quantity is more visibly detectable than quality. In contrast, because quality dimensions are more amorphous and subjective their inflation would be less susceptible to detection. A repeated measures planned comparison using absolute leniency as the dependent measure was performed to examine for possible rating level differences between the quantity and quality dimensions. Averaged across conditions, subjects were indeed more lenient when rating the quality dimensions ($\bar{M}=1.48$), than when rating the quantity dimensions ($\bar{M}=0.982$), ($F(1,22)=18.47$, $p<.001$), lending some credibility to the above argument.

It is interesting to note that none of the correlations obtained between the ratings and the objective scores for the quality of proofreading dimension were significant ($p > .05$) (see Table 5). As mentioned above, part of this may be due to the fact that the ratings were more lenient on this dimension because it was a quality dimension. Perhaps compounding this, however, is the fact that the objective scores on this dimension suffered from a restriction of range, compared to the other dimensions (see Table 3). Thus, it appears that not only were the ratings on this dimension lenient (restricting the range of the ratings), but the range of the objective scores was also restricted, thus making it more difficult to achieve convergence.

Future research should examine the effects of including quality versus quantity dimensions on self-ratings of performance. For instance, the goal-setting literature suggests that there is a trade-off between performance quantity and performance quality such that a person cannot successfully accomplish both for the same goal (Gilliland & Landis, 1992). Is this trade-off reflected in performance ratings? Is there a difference in the performance ratings obtained for quality and quantity when a supervisor conducts the appraisal compared to when an individual engages in self-appraisal? These are just some of the interesting questions that are worthy of future investigation.

In terms of the leniency results, it is not clear why a similar pattern of results was obtained for the quantity of proofreading and quality of comments dimensions (i.e., significant three-way interactions), while a different pattern of results was obtained for the quality of proofreading and quantity of comments dimensions. One possible explanation is that comments is a "quality" type of dimension, while proofreading is more amenable to a

"quantity" index. Given this, it would be easier for subjects in the social comparison condition to discern their self-performance relative to others for quantity of proofreading and quality of comments, contributing to the significant three-way interactions.

If nothing else, the disparate results among dimensions for both the leniency and validity measures illustrate the importance of examining dimensions separately, rather than combining dimensions to form a composite index (Edwards, 1993). The nature of the dimensions (i.e., proofreading and comments), as well as the way in which these dimensions were conceptualized and operationalized (i.e., quality versus quantity) appears to be very important in terms of how the variables of appraisal purpose, validation expectation, and social comparison jointly affect self-rating leniency and validity.

Conceptualizing the independent variables. Researchers investigating self-appraisals have operationalized the independent variables under study as dichotomous (e.g., Farh & Dobbins, 1989a; Farh and Werbel, 1986). Following this trend, I dichotomized the variables of appraisal purpose, validation expectation, and social comparison information. More specifically, subjects either performed their appraisals for a reward or not, either expected or did not expect validation, and either received social comparison information or no social comparison information. In terms of ecological validity, however, it is necessary to examine the generalizability of this dichotomization in a real world context. For example, it is realistic to assume that in most organizations where an appraisal is performed for reward purposes, this reward could take many forms. It could be in the form of a bonus, which could vary substantially in terms of the dollar amount, or the reward could

take the form of a promotion, which could be a small step up the corporate ladder or a large one. The same case could be made for both validation expectation and social comparison information. Validation could range from a specific, detailed inspection of ratings to a cursory glance of them. Social comparison information could range from an informal, partial knowledge of the performance of others to publicly posted, detailed information regarding the performance of every co-worker.

In effect, the variables used across most studies have varied in nature. For example, some researchers have used class points as a reward (e.g., Farh & Werbel, 1986), while others have used money (e.g., the present study). What is needed however, is a direct examination of this variability within a single study. Examining variables like reward, validation, and social comparison by varying the quality and/or the quantity of these variables may assist us in consolidating research by helping to explain discrepant findings when ostensibly the same variables have been utilized.

Given that it is now generally accepted that self-rating leniency is largely an issue of motivation (Farh, Werbel, & Bedeian, 1988; Klimoski & Hayes, 1980; Murphy & Cleveland, 1991; Shrauger & Osberg, 1981), perhaps a useful framework with which to consider these independent variables is expectancy theory (Kanfer, 1992; Vroom, 1964). As one moves along the "continuum" of any one of these variables the issues and motivations involved for individuals engaged in self-appraisal may change with respect to their perceived instrumentality, valence, and expectancy. More specifically, in the present study, reward might be associated with valence, social comparison information with instrumentality, and validation expectation with expectancy. Consider the following example: An individual engaged in

self-appraisal may consider the valence of inflating her ratings to be low for the purpose of a \$100 bonus, but quite high when the purpose is for a \$1 000 bonus. If social comparison information suggests that her performance is substandard, she may consider it more instrumental to inflate her ratings as a means of reward attainment than if she is the top performer. Finally, if validation expectation is high, she may expect that inflating her ratings will not result in reward attainment, whereas if validation expectation is low, she might expect that inflating her ratings will result in reward attainment.

In addition to the above, individual differences would likely determine how one person perceives the valence, instrumentality, and expectancy of a given situation compared to another. Research might incorporate moderator variables (e.g., equity sensitivity) to examine how these variables potentially affect the expectancy theory parameters. In sum, it seems important that future self-rating research investigates the nature of the independent variables under study, rather than continuing to treat them all dichotomously. Additionally, a theoretical framework consolidating these variables (such as expectancy theory) would be fruitful for suggesting possible underlying mechanisms which influence the motivation to elevate/deflate self-ratings of performance.

Relative versus absolute processing. The issue of providing social comparison information points to another area of research worthy of further investigation: the issue of relative versus absolute ratings. The self-appraisal literature has been inconsistent in terms of the methodology employed in obtaining self-ratings. More specifically, the research is inconsistent in framing self-appraisals in either relative or absolute terms. In some studies, subjects are instructed to compare themselves to others when rating their

performance (e.g., Meyer, 1980), while in other studies, subjects are simply told to evaluate their performance with no mention of others' performance (e.g., Fox & Dinur, 1988). Some studies have actually used both relative and absolute instructions, without taking this difference into account (e.g., Farh & Dobbins, 1989b). Thus, the instructions given to subjects have varied with respect to using relative or absolute information, which would presumably affect the way in which subjects rate their performance.

Additionally, the rating format utilized across studies has been inconsistent, with some studies providing subjects with comparative anchors (e.g., better than others) (e.g., Fox, Caspy, & Reisler, 1994) and others providing subjects with absolute anchors (e.g., high or low performance). Moreover, some studies have mixed relative instructions and absolute rating formats, and vice versa (e.g., Farh & Dobbins, 1989b).

The above issues represent inconsistencies in the self-appraisal literature which renders the consolidation and interpretation of research findings in this area difficult. Differences in terms of the way in which subjects rate themselves (i.e., relatively or absolutely) may affect the leniency and/or validity of self-ratings. Related to this, a meta-analysis by Heneman (1986) found that when appraisals were framed in relative terms (i.e., employee compared to employee) the average correlation between supervisor ratings and results-oriented criteria was .66, compared to an average correlation of .21 when appraisals were framed in absolute terms (i.e., employee compared to absolute standards). Heneman labeled this variable (absolute versus relative) as rating method, which makes it unclear as to whether or not this included rating instructions, rating format, or both. Although this meta-analysis was conducted on supervisor ratings, it suggests

that the distinction between absolute and relative ratings is important. There is no reason to believe that the same distinction is without consequences for self-ratings.

Some cognitive models of the performance appraisal rating process have considered the issue of relative versus absolute ratings of performance. For example, DeNisi et al. (1984) discussed the implications of rating format as a guide for the types of information raters choose to seek out. Specifically, comparative rating scales may guide raters to seek out more comparative information, while absolute rating scales guide people to seek out performance information specific to individual ratees. Extending this issue further, does the search for one type of information, as influenced by the rating format, preclude attending to, and the encoding of, other types of information?

The question of relative versus absolute ratings also has important implications for how raters process information. For example, do raters have a propensity to process information in a relative or absolute fashion? Moreover, does this processing change whether and the extent to which comparative information is provided? These questions are important to address for both self-ratings and ratings in general. It may well be that in the presence of social comparison information, raters process and rate performance in a relative manner, regardless of whether or not the rating format instructs them to do so.

Training for self-raters. As with previous self-rating research, the self-raters in the present study were untrained. Thus, it is likely that the raters were not operating under a common frame of reference regarding what constituted poor and good performance. The present study included social

comparison information as a possible method of providing raters with some information regarding performance standards. An alternate and more direct method for providing self-raters with a common frame of reference regarding alternate performance levels would be to employ frame-of-reference (FOR) training (Bernardin & Buckley, 1981). FOR training attempts to train raters to share and use common conceptualizations of performance when making evaluations (Woehr, 1994). FOR training has consistently been shown to increase the accuracy of supervisor ratings (e.g., Woehr & Huffcutt, 1994), but as yet has not been applied to training self-raters. This seems to be a logical progression for FOR training research, which has the potential to help reduce self-rating leniency. Indeed, given that self-raters and supervisors may possess different performance standards and thus attend to different aspects of performance (e.g., Hauenstein & Foti, 1989; Schmitt, Noe, & Gottschalk, 1986), the implementation of FOR training for both supervisors and self-raters may help to reduce the rating discrepancy typically found between these two groups. Future research should address these possible applications of FOR training.

In addition to FOR training, self-raters may benefit from rater error training as first described by Latham, Wexley, and Purcell (1975). This training provides raters with information regarding common errors that raters commit (e.g., leniency/severity, central tendency, halo, etc.) and has been found to be effective with supervisor ratings (e.g., Woehr & Huffcutt, 1994)). Perhaps if self-raters were enlightened on the possible errors they might commit (especially leniency), the psychometric quality of their ratings might be improved. It is important to note that this type of rater error training does not make reference to "correct" or "incorrect" rating

distributions, nor does it instruct raters not to commit these errors. Rather, it informs raters of common errors, explains possible reasons why raters would commit these errors, and discusses ways to avoid these errors (Woehr & Huffcutt, 1994). It is then left up to the raters whether and the extent to which they utilize the training information when they construct their ratings. In sum, future research should investigate the application of this rater error training to self-ratings.

Limitations and Conclusions

As with any research, the present study has certain limitations and concerns. First, the use of undergraduates in a contrived laboratory setting is always an issue with regards to the generalizability of results. Presumably somewhat mitigating this concern, however, is that the task included proofreading, something with which students should have considerable experience. Although the task was appropriate in the present context, it is not known how generalizable it is to work organizations. A field investigation using similar variables with an alternate task would certainly lend more credibility to the present results.

Second, based on anecdotal evidence, there is reason to believe that self-efficacy of proofreading may have confounded the results to some degree. Subjects were given the opportunity to make comments regarding their performance and many indicated that they felt they did not perform very well because they were not very good at proofreading. Whether and the extent to which this affected their self-ratings is not known, however, because self-efficacy was not measured. Although previous research has not specifically addressed self-efficacy, the effect of self-esteem on self-ratings has been

investigated. For example, Farh and Dobbins (1989b) found that subjects with high self-esteem tended to be more lenient in their self-ratings than those with low self-esteem. Thus, the results of the present study must be regarded with caution, as there is reason to believe that self-efficacy on the task may have influenced the self-ratings. This is another important area that future research should address.

The fact that only objective scores were used as the standard of comparison (i.e., the study did not include supervisor ratings) may make it more difficult to compare it with other research in the area. As stated in the introduction, however, there is no evidence to suggest that supervisor ratings are the appropriate standard of comparison against which to assess self-ratings. In fact, the extant literature would suggest the contrary. Thus, the exclusion of supervisor ratings in the present study represents a conscious decision to move away from considering supervisor ratings to be an appropriate standard of comparison. In any event, it is believed that the use of absolute objective scores provides a contribution to the literature by assessing self-ratings on a commensurate scale with objective scores. Perhaps this methodology will be adopted in future research as a means of assessing self-rating quality without having to employ supervisor ratings as the standard of comparison.

In light of the results of the present study, it appears that the consolidation of appraisal purpose, validation expectation, and social comparison information provides valuable information regarding self-ratings. That is, combining these variables leads to self-rating effects different from those that would be predicted when they are investigated separately. It appears that relatively speaking, self-raters are capable of rating their

performance in a reasonably accurate manner, but under certain circumstances (i.e., a combination of reward, no validation expectation, social comparison) they choose not to do so.

In closing, this study examined the effects of three variables on the quality of self-ratings, all of which were at the individual level of analysis. Future research should investigate additional variables that may influence self-ratings. For example, cognitive limitations in how individuals process information about their own performance may exist. Indeed, a number of theories (e.g., self-serving attribution bias) suggest avenues for further inquiry into the process of self-rating. Additionally, variables at the group level (e.g., group norms) and at the organizational level (e.g., span of control, type of rating system) may account for significant variability in self-ratings. The investigation and consolidation of some of these additional variables would represent an important step towards the formulation of a meta-theory of the self-rating of performance. Such a meta-theory would be fruitful from a scientific and practical standpoint, hopefully providing a comprehensive picture of the antecedents of self-rating behaviour.

References

- Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.
- Arnold, J., & Davey, K.M. (1992). Self-ratings and supervisor ratings of graduate employees' competence during early career. Journal of Occupational and Organizational Psychology, 65, 235-250.
- Banks, C.G., Murphy, K.R. (1985). Toward narrowing the research-practice gap in performance appraisal. Personnel Psychology, 38, 335-345.
- Bassett, G.A., & Meyer, H.H. (1968). Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.
- Bernardin, H.J., & Beatty, R.W. (1984). Performance appraisal: Assessing human behaviour at work. Boston: Kent.
- Bernardin, H.J., & Buckley, M.R. (1981). A consideration of strategies in rater training. Academy of Management Review, 6, 205-212.
- Bernardin, H.J., & Klatt, L.A. (1985). Managerial appraisal systems: Has practice caught up with the state of the art? Public Personnel Administrator, November, 79-86.
- Borman, W.C. (1974). The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 12, 105-124.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.
- Campbell, D.J., & Lee, C. (1988). Self-appraisal in performance evaluation: Development versus evaluation. Academy of Management Review, 13, 302-314.

Cascio, W.F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cronbach, L.J., Furby, L. (1970). How should we measure "change"- or should we? Psychological Bulletin, 74, 68-80.

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

Dipboye, R.L., & de Pontbriand, R. (1981). Correlates of employee reactions to performance appraisals and appraisal systems. Journal of Applied Psychology, 66, 248-251.

Eder, R.W., & Fedor, D.B. (1989). Priming performance self-evaluations: Moderating effects of rating purpose and judgment confidence. Organizational Behavior and Human Decision Processes, 44, 474-493.

Edwards, J.R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. Personnel Psychology, 46, 641-665.

Farh, J.L., & Dobbins, G.H. (1989a). Effects of comparative performance information on the accuracy of self-ratings and agreement between self- and supervisor ratings. Journal of Applied Psychology, 74, 606-610.

Farh, J.L., & Dobbins, G.H. (1989b). Effects of self-esteem on leniency bias in self-reports of performance: A structural equation model analysis. Personnel Psychology, 42, 835-850.

Farh, J.L., & Werbel, J.D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. Journal of Applied Psychology, 71, 527-529.

Farh, J.L., Dobbins, G.H., & Cheng, B-S. (1991). Cultural relativity in action: A comparison of self-ratings made by Chinese and U.S. workers. Personnel Psychology, 44, 129-147.

Farh, J.L., Werbel, J.D., & Bedeian, A.G. (1988). An empirical investigation of self-appraisal-based performance evaluation. Personnel Psychology, 41, 141-156.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 60, 127-148.

Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Folger, R., & Greenberg, J. (1985). Procedural justice: An interpretive analysis of personnel systems. Research in Personnel and Human Resources Management, 3, 141-156.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. Personnel Psychology, 41, 581-592.

Fox, S., Caspy, T., & Reisler, A. (1994). Variables affecting leniency, halo, and validity of self-appraisal. Journal of Occupational and Organizational Psychology, 67, 45-56.

Furnham, A., & Stringfield, P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of supervisor evaluation. Journal of Occupational and Organizational Psychology, 67, 57-67.

Gilliland, S.W., & Landis, R.S. (1992). Quality and quantity goals in a complex decision task: Strategies and outcomes. Journal of Applied Psychology, 77, 672-681.

Greller, M.M. (1975). Subordinate participation and reactions to the appraisal interview. Journal of Applied Psychology, 60, 544-549.

Harris, M.M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.

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

Heneman, H.G. (1974). Comparisons of self- and superior ratings of managerial performance. Journal of Applied Psychology, 59, 638-642.

Heneman, H.G. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Hoffman, C.C., Nathan, B.R., & Holden, L.M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. Personnel Psychology, 44, 601-619.

Holzbach, R.L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 63, 579-588.

Judge, T.A., & Ferris, G.R. (1993). Social context of performance evaluation decisions. Academy of Management Journal, 36, 80-105.

Kanfer, R. (1991). Motivation theory and industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.). Handbook of Industrial and Organizational Psychology, 1, (pp. 76-170). Palo Alto, CA: Consulting Psychologists Press.

Klimoski, R.J., & Hayes, N.J. (1980). Leader behaviour and subordinate motivations. Personnel Psychology, 33, 543-555.

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

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

Landy, F.J., Barnes, J.L., & Murphy, K.R. (1978). Correlates of perceived fairness and accuracy of performance evaluation. Journal of Applied Psychology, 63, 751-754.

Lane, J., & Herriot, P. (1990). Self-ratings, supervisor ratings, positions and performance. Journal of Occupational Psychology, 63, 77-88.

Latham, G.P., & Wexley, K.N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

Latham, G.P., Wexley, K.N., & Purcell, F.D. (1975). Training manager to minimize rating errors in the observation of behaviour. Journal of Applied Psychology, 60, 550-555.

Levine, E.L. (1980). Introductory remarks for the symposium "organizational applications of self-appraisal and self-assessment: Another look". Personnel Psychology, 33, 259-271.

Levine, E.L., Flory, A., & Ash, R.A. (1977). Self-assessment in personnel selection. Journal of Applied Psychology, 62, 428-435.

Levy, P.E. (1993). Self-appraisal and attributions: A test of a model. Journal of Management, 19, 51-62.

Longenecker, C.O., Sims, H.P., & Gioia, D.A. (1987). Behind the mask: The politics of employee appraisals. Academy of Management Executive, 1, 183-193.

Mabe, P.A., & West, S.G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.

Meyer, H.H. (1980). Self-appraisal of job performance. Personnel Psychology, 33, 291-295.

Meyer, H.H. (1991). A solution to the performance appraisal feedback enigma. Academy of Management Executive, 5, 68-76.

Miceli, M. (1985). The effects of realistic job previews on newcomer behavior: A laboratory study. Journal of Vocational Behavior, 26, 277-289.

Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74, 619-624.

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

Pearce, J.L., & Porter, L.W. (1986). Employee responses to formal performance appraisal feedback. Journal of Applied Psychology, 71, 211-218.

Riggio, R.E., & Cole, E.J. (1992). Agreement between subordinate and superior ratings of supervisory performance and effects on self and subordinate job satisfaction. Journal of Occupational and Organizational Psychology, 65, 151-158.

Roberson, L., Torkel, S., Korsgaard, A., Klein, D., Diddams, M., & Cayer, M. (1993). Self-appraisal and perceptions of the appraisal discussion: A field experiment. Journal of Organizational Behavior, 14, 129-142.

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

Schmitt, N., Noe, R.A., & Gottschalk, R. (1986). Using the lens model to magnify raters' consistency, matching, and shared bias. Academy of Management Journal, 29, 130-139.

Shore, L.M., & Thornton, G.C. (1986). Effects of gender on self- and supervisory ratings. Academy of Management Journal, 29, 115-129.

Shrauger, J.S., Osberg, T.M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. Psychological Bulletin, 90, 322-351.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the Validation and Use of Personnel Selection Procedures (3rd ed.). College Park, MD: Author.

Steel, R.P., Ovalle, N.K. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-685.

Sulls, J.M., & Miller, R.L. (1977). Social comparison processes: Theoretical and empirical perspectives. Washington, DC.: Hemisphere.

Sulsky, L.M., & Balzer, W.K. (1988). The meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.

Teel, K. (1978). 'Self-appraisal revisited'. Personnel Journal, 57, 364-367.

Thornton, G. (1980). Psychometric properties of self-appraisal and job performance. Personnel Psychology, 33, 263-271.

Vroom, V.H. (1964). Work and motivation. New York: Wiley.

Williams, J.R., & Levy, P.E. (1992). The effects of perceived system knowledge on the agreement between self-ratings and supervisor ratings. Personnel Psychology, 45, 835-847.

Woehr, D.J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. Journal of Applied Psychology, 79, 525-534.

Woehr, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology, 67, 189-205.

Zammuto, R.F., London, M., & Rowland, K.M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.