Evaluating Differential Rater Functioning in Performance Ratings: Using a Goal-Based Approach

A dissertation presented to

the faculty of

the College of Arts and Sciences of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Kevin B. Tamanini

November 2008

This dissertation titled

Evaluating Differential Rater Functioning in

Performance Ratings: Using a Goal-Based Approach

by

KEVIN B. TAMANINI

has been approved for

the Department of Psychology

and the College of Arts and Sciences by

Jeffrey Vancouver

Associate Professor of Psychology

Benjamin M. Ogles

Dean, College of Arts and Sciences

Abstract

TAMANINI, KEVIN B., Ph.D., November 2008, Psychology

Evaluating Differential Rater Functioning in Performance Ratings: Using a Goal-Based Approach (223 pp.)

Director of Dissertation: Jeffrey Vancouver

Measuring performance in the workplace is an endeavor that has been the central focus of many applied researchers and practitioners. Due to the limited information that objective data provides to decision makers, subjective data are often used to supplement performance ratings. Unfortunately subjective ratings can be biased. Indeed, rating errors frequently bias ratings and have plagued performance evaluations. Much of the performance appraisal (PA) research has focused on ways of eliminating, detecting, or controlling these rater errors. The results from these areas are mixed and insufficient in providing insights and understanding about how to deal with rater errors.

This research extends and tests a technique called differential person functioning (DPF; Johanson & Alsmadi, 2002) to the detection of rater bias (specifically leniency/severity) during a performance evaluation, as well as test a goal-based approach for performance evaluations. The DPF technique is used to identify the responses for a given individual that are different for different groups of items. The goal-based approach proposes that individuals' pursuit of different goals is what leads to different ratings. Two studies were conducted to examine these phenomena.

The first study was a pilot study to refine the materials and manipulations that were to be used in the main study. Specifically, two different evaluation formats were compared, sex differences were examined, and the manipulation was tested. In the second study (i.e., the main study) the sensitivity and consistency of the DPF technique was compared with two other traditional methods for detecting leniency/severity. Participants completed an actual performance evaluation for

Approved: ____

Jeffrey Vancouver

Associate Professor of Psychology

Acknowledgments

There are several individuals I would like to thank for their help in the successful completion of this dissertation. I would like to thank my committee chair and advisor, Jeff Vancouver, for his guidance, assistance, and advice throughout this entire process. I would also like to thank George Johanson for his valuable insights throughout the project as well as to the other committee members, Paula Popovich, Rodger Griffeth, and Keith Markman for their extremely helpful comments and suggestions to improve the quality of this document. I would also like to offer a special thanks to my family for their continued support for everything I do. Finally, I would like to thank my wife, Heather, who showed an everlasting patience throughout this entire process. I love you and I've finally done it.

Table of Contents

Page

Abstract 3

Acknowledgments 5

Introduction 13

Bias in Performance Appraisal 19

Rating Errors 19

Halo. 21

Central Tendency. 21

Leniency 22

Methods for Addressing Rating Errors 24

Rating Format 25

Graphic Rating Scales. 27

Behaviorally Anchored Rating Scales (BARS). 28

Mixed Standard Scales (MSS). 30

Behavior Observation Scales (BOS). 31

Forced-Choice Ratings 32

Summary. 33

Rater Training 35

Rater Error Training (RET). 36

Frame-of-Reference training (FOR). 36

Summary. 37

Cognitive Approach to PA 38

Rating Process Models 39

Goal-Based Perspective for PA 41

Summary. 45

Methods for Detecting Leniency/Severity 47

Utilizing IRT for Performance Appraisal 49

Differential Item Functioning. 50

Model-Fit Approach (e.g. Person-Fit). 55

Limitations of Current IRT Approaches 57

Differential Person Functioning 60

Current Research 64

Present Studies 67

Pilot Study 69

Method 69

Participants 69

Response Format. 69

Response Instructions. 69

Evaluation Form. 71

Measures 73

Evaluation Process and Format Reactions. 73

Goal Importance Questionnaire 73

Additional Items 73

Procedure 74

Results.......... 76

Reliability Analysis.......... 76

Manipulation Check.......... 76

Tests for Sex Effects.......... 78

Item-type Effects.......... 80

Pilot Study Discussion.......... 81

Main Study.......... 83

Methods.......... 83

Participants.......... 83

Manipulations.......... 83

Response Instructions.......... 83

Measures.......... 85

Evaluation Form.......... 85

Evaluation Process and Format Reactions.......... 85

Goal Importance Questionnaire.......... 86

Additional Items.......... 86

Psychology Department Rating Form.......... 87

Procedure.......... 87

Results.......... 90

Test for Fatigue and Order Effects.......... 90

Manipulation Check/Goal Manipulation.......... 90

Comparison of Detection Methods.......... 93

Classification with the Differential Person Functioning Method. 94

Classification with the Mean Score Method. 98

Classification with Skewness Ratings. 100

Classification Consistency between Methods... 101

Summary 109

Scale Differences 110

Proportion of Biased Raters 110

Evaluation Reactions 112

Goal Questionnaire Relationships 114

Discussion 118

Primary Findings... 119

Additional DPF Findings 125

Additional Goal-Based Findings 129

Summary 133

Personnel Implications.. 134

Conclusions.. 135

References.. 140

Appendix A 159

Appendix B 168

Appendix C 179

Appendix D.. 180

Appendix E 182

Appendix F 183

Appendix G 184

Appendix H 186

Appendix I 187

Appendix J 188

Appendix K 190

Appendix L 197

Appendix M 204

Appendix N 211

Appendix O 218

Appendix P 220

Appendix Q 222

LIST OF TABLES

Table Page

1. Descriptives and Sex Differences across Format Type and Instruction Condition for Overall Evaluation Ratings 79

2. Open-ended Manipulation Check Item Percentages..91

3. Means and Standard Deviations for the Overall Scores for Each Response Condition...93

4. Results of the Classification Consistency Analysis for the DPF, Mean Score, and Skewness Methods for Detecting Rater Bias...105

5. Results of the Classification Consistency Analysis for the DPF, Mean Score, and Skewness Methods for each Type of Bias (i.e., Leniency and Severity)......................107

6. Descriptive Statistics for Evaluating Reaction Items...113

7. Descriptive Statistics for the Goal Importance Questionnaire.................................114

LIST OF FIGURES

Figure Page

1. Generic Item Characteristics Curve 137

2. DPF Analysis Showing Two PCCs for Different Types of Items: Demonstrating DPF 138

3. Theoretical Decision Classification Table 139

Introduction

Defining, understanding, and evaluating performance within a work context is a central issue within industrial and organizational psychology (Arvey & Murphy, 1998; Landy & Farr, 1980). Realizing the importance of performance measurement and measuring it accurately are two distinct matters (Landy & Farr, 1980). Although it is a goal of an organizational decision maker to determine accurate assessments of employees' performances (Murphy & Balzer, 1989), doing so is easier said than done. Often, decision makers assume the most accurate measurement of performance is hard, objective criteria (e.g., absences, accidents, or tardiness). However, these are commonly deficient measures that do not adequately capture an individual's overall performance. Performance (as an ultimate criterion) is a complex construct that is difficult to completely capture. Deficiency occurs when the measurement of the performance criteria is incomplete. Indeed, Landy and Farr (1983) note several aspects that lead to this deficiency in objective data.

First objective indices tend to have low reliability. For example, in terms of absences, the observation period may not be stable across measures or external factors (i.e., sick leave policies) may influence the reliability of absence measures. Second, objective measures are only available for a limited number of jobs. For example, it does not make sense to look at tardiness from those who may not have a predetermined work day, or even work from home on a frequent basis (e.g., consultants, contractors, etc.). Finally, the changing nature of work often makes objective measures inappropriate for measuring work performance. Technological advances make outputs more dependent on those technologies than on individual performance. Because the goal of a performance appraisal is to choose criteria that optimize the assessment of job success, keeping overall deficiency to a minimum is imperative (Riggio, 2003).

To compensate for the deficiencies in objective data, most ratings of individual performance depend on subjective indices (Guion, 1965; Murphy & Cleveland, 1995). Unfortunately the subjective data often leads to contaminated/biased ratings (i.e., rating errors). Indeed, according to Holzbach (1978; p. 578), "Rater bias, in its various forms and manifestations, is perhaps the most serious common drawback to performance ratings." Because subjective ratings can be contaminated (biased), they lose the accuracy that decision makers desire (Borman, 1979, Landy & Farr, 1980). Hence, the dilemma; if objective data are deficient, and subjective data contaminated, how should performance be evaluated?

Despite the realization that subjective indices may yield biased ratings, organizations have had no choice but to continue to use them because there is no other alternative. Indeed, subjective appraisals are found in 90% of organizations (Bernthal, Sumlin, Davis, & Rogers, 1997) and influence decision making processes (Bernardin & Villanova, 2005). This heavy use has influenced researchers and practitioners to seek a "cure" for dealing with biased ratings (Landy & Farr, 1980; Murphy & Cleveland, 1995; Saal, Downey, & Lahey, 1980). For example, research has examined how different rating formats (e.g., graphic rating scale vs. behaviorally anchored rating scales), different rater characteristics (e.g., peer vs. supervisor), different ratee characteristics (e.g., race, sex), different rater training programs (e.g., frame of reference training), and different statistical controlling techniques influence the occurrence of various rating errors. Much of this research was based on the assumption that individuals unknowingly commit rating errors. In turn, errors are assumed the result of unconscious (i.e. automatic) information processing processes that might be overcome by "raising the consciousness" of raters through techniques such as error training. However, some researchers claim that there are instances in which individuals are aware of the biases in the ratings they give (Murphy & Cleveland, 1995).

The evidence that biased ratings are due to the deliberate, "volitional" distortion of performance ratings has been growing (Bernardin & Beatty, 1984; Bernardin & Villanova, 1986; Murphy & Cleveland, 1995; Tziner, Murphy, & Cleveland, 2005). There have been speculations as to why individuals may intentionally distort their responses, including: 1) performance appraisal purpose, 2) organizational goals, and 3) rater goals (Murphy & Cleveland, 1995). However, there have been few empirical studies that have attempted to provided evidence of a goal-based (i.e., motivational) aspect behind the occurrence of rating errors.

If individuals are intentionally distorting their responses, then the aforementioned approaches will remain insufficient for adequately understanding rating errors. Indeed, the goal-based approach, in which individuals provide ratings based on the goals they are pursuing, is the only current perspective that examines rater errors as an intentional process (Murphy & Cleveland, 1995). Because of this, rather than using techniques to control errors (e.g., format, training), it may be better to utilize methods to detect those who are committing errors and deal with their ratings accordingly. Unfortunately, the limitations with the performance appraisal research are not isolated to interventions (e.g., format changes, training, etc). Indeed, there has been an unrealized opportunity to utilize newer statistical techniques that could provide better insights and explanations about why rater errors occur. These techniques typically focus on identifying ratings that fit a certain response pattern. Once those ratings are identified, then that information is utilized to make decisions regarding the usefulness (i.e., reliability and validity) of those ratings.

Depending upon the type of error that one is attempting to detect, there are various statistical procedures that may be used (e.g., mean correlation among performance dimensions, over ratees [halo], mean ratings over ratees and dimensions [leniency], etc.). Even though these methods are consistently used, some have argued that there is still ambiguity concerning the detection of these rating errors, in part, because the incorrect unit of analysis has been utilized (Murphy & Balzer, 1989). For example, there has been a considerable amount of performance appraisal research that has examined leniency by examining differences between groups (e.g., peer ratings vs. subordinate ratings). However, this approach assumes that groups, not the individuals within the groups are lenient or severe. Additionally, mean ratings across ratees, although predominantly used, may not provide the most accurate information on how or why rating biases arise. The problem is that the ratings are confounded with the raters. For example, it could be that raters are applying biases only to a subset of ratees, but without a fully crossed design (i.e., all raters rate all ratees), this cannot be determined. Because of this, a technique that does not confound the rater with the ratings would be more ideal.

An alternative to the traditional methods of understanding and detecting rating errors is to examine person-fit models. Person-fit models use item response theory (IRT) and differential item functioning (DIF) to assess specific latent trait IRT models that represent rater effects (Wolfe, 2004). According to this person-fit approach, if an individual does not fit a model, then there is evidence that the individual is responding in a biased manner (e.g., his or her ratings are lenient or severe). Although these techniques have provided some useful information regarding the examination of rater effects, in that they demonstrate the usefulness of utilizing IRT models in conjunction with non-IRT based functions (i.e., DIF) for detecting various errors, there are still limitations to consider.

For example, the person-fit modeling approach is similar to the traditional methods discussed previously, in that it assumes a typical response pattern. Although this person-fit approach may identify biased raters, it merely provides information about individuals and little, if any, information about groups of items. Information from the item level is not being captured; therefore even if a rater is identified as giving biased ratings, there is no understanding as to why. Just as with test bias, we should not ignore the item level. Indeed, it is the interaction between the individual and the items that should be the focus of our attention. For example, if an individual is demonstrating differential functioning as a rater, the properties of the item could then be examined to determine if the effects are due to the items themselves or possibly to some other factor (e.g., goals). Ideally, we would like to be able to cross our levels of analysis and obtain information about both individuals and items that would allow us to determine who is giving biased ratings and why they are doing so.

Fortunately there is a technique that may allow for a more sensitive examination of both individuals and item properties simultaneously. This technique is called differential person functioning (DPF: Johanson & Alsmadi, 2002). Rather than determining which items are “acting” differentially for different groups (e.g., peers vs. subordinates), DPF is a technique that can be used to determine if the responses for given individuals are different across different groups of items (e.g. focal vs. referent). Because of the fact that DPF takes both items and persons into account, it may be a more sensitive (and appropriate) technique than the traditional methods (i.e., mean differences, skewness) for identifying biased raters. Specifically, utilizing the DPF technique is not just a matter of specificity (i.e., one rater for one ratee as opposed to multiple ratings averaged across multiple ratees for a given rater), but it is more sensitive in that the information about individuals allows for the detection of biased raters and the information about the items allows for an understanding of why he/she is giving those ratings.

The DPF technique has yet to be used in performance appraisal research; hence it is my intention to utilize this technique to detect rater errors during a performance appraisal situation. Specifically, I extend the DPF technique to detecting rater effects (specifically leniency/severity) from a performance evaluation measure. Unlike the other research that has attempted to detect rater effects with simulated data (i.e., person-fit models, IRT-base approaches), I will use field data. By using the more sensitive DPF technique, I hope to demonstrate a larger effect when compared to the commonly used methods for detecting leniency. Additionally, I attempt to provide some empirical support for the goal-based, motivational approach that has been proposed (i.e., Cleveland & Murphy, 1992). Although some evidence exists for the goal-based conceptualization of rater errors, it is weak. Specifically, I will test the goal-based theory by manipulating rater goals (e.g., administrative decisions vs. feedback) as well as item properties. In addition to providing evidence for a goal-based perspective, this test serves as a validation of the DPF technique for detecting lenient raters. In the following sections of this paper common rater errors, approaches to detecting and dealing with rater errors, the differential person functioning approach, and the methods utilized in the paper are discussed.

Bias in Performance Appraisal

Rating Errors

As long as organizations continue to rely on rating instruments to evaluate the performance of employees, the quality of ratings will continue to be of interest to both managers and researchers (Tsui & Barry, 1986). It is important to know whether performance ratings provide an accurate reflection of performance for those being rated. Performance appraisal (PA) has traditionally been viewed as a measurement problem, which has focused on various issues including the reduction of test and rater bias (Murphy & Cleveland, 1991). Indeed, rater bias is considered a substantial source of error within psychological research (Hoyt, 2000). Because of this, there is an inherent need for criteria that can be used to assess the quality of ratings, focusing much of the PA research on the search for “better,” more accurate, techniques for measuring job performance (Murphy & Cleveland, 1991).

The most common approach used to examine the quality of performance ratings is to examine the psychometric characteristics/properties of the ratings themselves (Borman, 1991, Cleveland & Murphy, 1995). According to Murphy and Cleveland (1995) these measures of the psychometric quality of ratings are classified into three broad groups: 1) traditional psychometric criteria (e.g., reliability, validity); 2) indices of rater errors that reflect response biases on the part of the raters; and 3) direct measures of the accuracy. Of these, the rater error approach has been the most common. Rater error approaches assume accuracy is a function of the presence or absence of rating errors (Murphy & Cleveland, 1995). Likewise, many believe that rater errors tend to undermine the reliability and validity of the information obtained (Bannister, Kinicki, DeNisi, & Hom, 1987). Hence, the most common method for evaluating ratings involves the assessment of rater errors (Landy, 1986).

Based on a comprehensive review of the literature, Saal and colleagues (1980) identified the major categories of rater errors: 1) halo, 2) central tendency, and 3) leniency (or severity). Research that has examined rater errors has taken many different perspectives. As such, there are numerous operational definitions of each type of error. To further complicate the matter, there are different statistical methods of detecting each type of error, according the operational definition that is used. Below I will review each of the three errors as well as the typical definitions that are used for each error.

Halo. Halo refers to a rater's tendency to give similar evaluations to separate aspects of an individual's performance, even though the dimensions are clearly distinct (Thorndike, 1920). Typically, halo is defined in one of two ways: 1) a rater's tendency to allow overall (global) perceptions to distort the ratings on specific aspects of a ratee's performance, or 2) a raters unwillingness to discriminate among separate aspects of an individual's performance (Saal et al., 1980). The first definition tends to agree with the belief that raters commit halo unintentionally, therefore there are statistical methods to control for such errors (see Ritti, 1964). However there are several researchers who have shown that this approach to control for halo tends to do more harm than good and should not be used (Harvey, 1982; Hulin, 1982; Murphy, 1982).

According to Murphy and Cleveland (1995) the second definition suggests that individuals intentionally distort their ratings so that the correlations among dimensions correspond to the conceptual similarity among dimensions. This definition is more in line with the current notion behind rater errors. The issue with this definition is that one reason why ratings on separate dimensions may be correlated is that the behaviors being rated really are correlated (valid (true) halo). It is invalid (illusory) halo that is a result of the intentional distortion on the part of the rater, therefore the rating error that is occurring.

Central Tendency. Central tendency refers to a rater's unwillingness to assign extreme (i.e., high or low) ratings. This is an error in which a rater assesses a disproportionately large number of ratees as performing in the central part of a distribution of rated performance, in contrast to their true level of performance

(Muchinsky, 2006). The assumption is that the true distribution of performance is assumed to be normal and the true variability of performance is considered "substantial" (Murphy & Cleveland, 1995). When the variability of the ratings is small, there is range restriction. When the range restriction falls around the center of the scale, then central tendency is believed to be occurring. If a rater is committing this error, one can imply that they view everyone as "average", because only the middle part of the evaluation scale is utilized. Many times central tendency occurs when raters are supposed to rate aspects of an individual's performance unfamiliar to them.

Leniency. Leniency typically refers to the tendency of raters to "rate well above the midpoint of the scales used" (Kneeland, 1929; p. 356), as indicated by average ratings over all ratees (Saal, et. al., 1980). The assumption in this case is that the true mean level of performance corresponds to the scale midpoint. The notion behind this error is that a rater may give ratings that are higher than warranted by actual performance (leniency) or ratings lower than warranted (severity). Leniency (as with central tendency) is a distributional error in that the restriction of range in scores around the upper end of the scale (high mean ratings) imply leniency. There is much speculation (especially within performance appraisal research) as to why raters give lenient/sever ratings (e.g., inaccurate frame of reference or norms, PA purpose, etc). Indeed, inflation is one of the most frequently cited problems associated with performance ratings (Bernardin & Orban, 1990; Ilgen & Feldman, 1983; Murphy & Cleveland, 1995). The appraisal process for the military and civil service are examples of domains where the pervasiveness of leniency in ratings often renders and entire appraisal system worthless (Bernardin & Orban, 1990).

Similarly, Hide (1982) noted that there are often “vast quantities” of inflated reports that lead to severe consequences when using performance ratings.

Lenient ratings can lead to a variety of outcomes that can severely influence decision making. Specifically, lenient ratings are a source of problems when an organization wants to terminate an employee because of poor performance (Bernardin & Cascio, 1988), as well as when personnel decisions are based on comparisons of individuals to some standard (Bernardin & Orban, 1990). Similarly, Murphy and Cleveland (1995) provide a detailed discussion of several consequences as a result of inflated ratings. These include: 1) consequences for the ratee – pay, promotion, etc.; 2) consequences for the rater – manager looks better with higher performing employees; 3) avoidance of negative reactions – reduce confrontations with employees; and 4) maintaining the organization’s image.

Just as with halo and central tendency, leniency could be the result of a rater’s unwillingness to give accurate ratings. Because intentional distortion is a possibility, the traditional methods for dealing with leniency may not appropriate in many situations. As such, the motivational (i.e., goal-based perspective) may be helpful in attempting to understand why raters give lenient (e.g., inaccurate) ratings. One purpose of this study is to provide empirical support for this perspective.

Whereas the majority of rater error research has focused on these three main errors, there have been other errors discussed within the literature: logical error (Newcomb, 1931); contrast error (Murray, 1938); proximity error (Stockford & Bissell, 1949); similar-to-me (Latham, Wexley, & Pursell, 1975); the first impression error

(Latham, et. al., 1975); and systematic distortion (Kozlowski & Kirsch, 1987). Due to the lack of research surrounding these errors, they will not be the focus of the remainder of this paper. Within the performance appraisal literature, it is has been noted that leniency (i.e., inflated ratings) is the most serious problem that needs to be dealt with due to the implications lenient or severe ratings may have on personnel decisions (Ilgen & Feldman, 1983; Landy & Farr, 1980; Murphy & Cleveland, 1995). Interestingly though, leniency may not be an "error" at all, but rather a behavior that allows a rater to obtain rewards and avoid punishments (Murphy & Cleveland, 1995). From this perspective, there are many understandable reasons for giving inaccurate (typically inflated) ratings, and more importantly, relatively few reasons for giving accurate ratings. As such, applied researchers have been focused on finding ways to eliminate and/or reduce lenient raters. Below, the techniques that were developed for this purpose are reviewed.

Methods for Addressing Rating Errors

Over the last 80 years, there have been many attempts to understand and deal with rater errors (Murphy & Cleveland, 1995). Over that time researchers have taken several approaches. Much of the early work regarding the issue of rater errors, focused on the development and comparison of different rating formats. Rater training focused on reducing rating errors and improving observation skills has also received substantial attention (Ilgen, Barnes-Farrell, & McKellin, 1993). Results from format research are somewhat mixed, and although there is evidence that training does reduce certain rating errors, there is a common theme to both perspectives. They perceive the rater as committing errors unknowingly; therefore changes to the environment should alleviate the occurrence of errors. Because of the consistently mixed results from both the format research and the training research, the focus began to shift away from structural changes to process changes. In general, there was a belief that cognitive characteristics of raters (e.g., rater characteristics, ratee characteristics, etc.) held the most promise for understanding the rating process (e.g., Feldman, 1986; Landy & Farr, 1980). More recently, a motivational approach has begun to make some headway because it addresses the issue of why individuals provide certain ratings (e.g., rating errors). Specifically, researchers believe that the goals of the rater, and/or goals of the organization, will influence the types of ratings that individuals will give (Murphy & Cleveland, 1995). Although, research based on the cognitive approach and the goal-based approach have been more helpful at providing answers to issues regarding rater errors there has been limited success at best. Each of these areas of research and their results are reviewed below.

Rating Format

Much of the early work dealing with PA had focused on the development of many different types of rating formats to be used for both research and practice. As noted by Borman (1991), it has been compelling for researchers to believe that there are characteristics of the rating formats themselves that play a role in determining the accuracy of ratings. Indeed, there is an enormous amount of research devoted to the efforts of exploring the potential effects of rating formats on rating errors. According to Murphy and Cleveland (1995), if the number of studies devoted to the rating scale format were counted, it would appear as if this were the most important issue in PA, dating back to the pioneering work of Paterson (1922) and his development of the graphic rating scale.

Most of the popular methods typically require raters to provide some judgment of performance based upon some absolute criterion (e.g., goal) or the performance of others (Berdardin & Beatty, 1984). Either way, raters are being asked to make performance-based decisions based on human judgment. As such, there is potential that rating errors may occur. Because of this, much of the research on scale formats has attempted to determine what formats are superior (i.e., which ones result in the fewest rater errors) (e.g., Bernardin, 1977; Borman & Dunnette, 1975; Borman, 1979). For example, research has examined specific characteristics of the rating scales, such as: the number of response categories (e.g., Bernardin, LaShells, Smith, & Alvares, 1976), types of anchors (e.g., Smith & Kendall, 1963), the process of assigning values to anchors (e.g., Barnes & Landy, 1979; Silverman & Wexley, 1984), as well as the psychological processes involved when using different formats (Murphy & Constans, 1987, 1988).

Borman (1991) lists 12 different types of rating formats that have been examined in both research and practice (e.g., forced choice, critical incidents, behaviorally anchored rating scales , etc; for an extensive review of formats see Bernardin & Beatty, 1984; Whisler & Harper, 1962). Although, research on formats has been extensive and long lasting, much of the current PA research has paid little attention to the question of which format is best (Murphy & Cleveland, 1995). This drop off in interest is mainly due to the results of a review by Landy and Farr (1980). Based on their search of the literature, they concluded that formats had only a minimal effect on the quality of ratings. Additionally, they ultimately concluded that no one format was consistently better than the others. According to Murphy and Cleveland (1995), the results of the Landy and Farr review called for a "moratorium" on the research dealing with scale format. The drastic decline in dealing with rating formats does seem justified. As noted previously, rating formats assume that rater errors are systematic and therefore can be dealt with via environmental changes (i.e. structural changes). It is apparent, however, that rater errors may not be systematic if they are based on the goals one is pursuing. Therefore structural changes will continually be inadequate for dealing with rater errors. Although research on this topic has fallen by the wayside, it is still useful to briefly examine some of the features and findings (especially those related to rating errors) of several of the more widely used formats in PA to gain an understanding of the progression of rater error research.

Graphic Rating Scales. This is the simplest scale format, in which raters recorded their judgments about specific aspects of a ratee's performance. Developed by Paterson (1922) to free the rater from quantitative judgments and allowed the rater to make fine discriminations (Landy & Farr, 1980). These types of scales consist of trait labels with varying types and number of adjectives (e.g., very poor, very good, intermediate). Unfortunately, there is little structure to this type of format. Indeed, much of the research dealing with graphic rating scales compared different variations of the scale itself. For example, Madden and Bourdon (1964) examined the physical arrangement in which the rating scale definition and levels were presented to raters. Specifically, they varied the position of the high end of the scale (top vs. bottom; right vs. left), spatial orientation (horizontal vs. vertical), segmentation (segmented vs. unbroken), and numbering (1-9 vs.

-4 to +4) simultaneously. They found that varying the physical appearance of a graphic rating scale showed a significant main effect across all of the conditions.

Although the simplicity of the scale is its main advantage, it is also its main disadvantage. It often lacked clarity and definition of the categories and often resulted in very different standards for evaluating the same behaviors (Murphy & Cleveland, 1995). Additionally, as noted by Ryan (1958), the subjective and arbitrary nature of the graphic scales often led to leniency as well as halo, eliminating the usefulness of the ratings. As a result, several other scale formats were developed to solve the problems with defining performance dimensions. One of which, was the Behaviorally Anchored Rating Scales.

Behaviorally Anchored Rating Scales (BARS). Developed by Smith and Kendall (1963), BARS accounted for much of the research on PA in the late 60s and 70s. BARS differ from graphic scales in that the anchors that appear at different intervals are examples of actual behavior rather than adjectives modifying trait labels or simple numbers. Specifically, this “behavioral expectation scaling” used behavior examples of different levels of performance to define the dimension being rated as well as the performance level on the scale. Essentially, it is the rater’s job to compare the observed behaviors of a given ratee with the behavioral anchors on the scale to assign a rating on a particular dimension.

According to Bernardin and Smith (1981), the BARS method is more than just a format; it is a system that requires acute attention to detail. In terms of direct methods of performance rating, the BARS system commands the most attention (Landy & Farr, 1980).

There has been a good deal of research that has attempted to assess the effectiveness of the BARS in relation to traditional graphic rating methods with varying results. Whereas Campbell, Dunnette, Arvey, and Hellervik (1973) compared these two methods and concluded that the BARS format yielded less method variance, less halo, and less leniency in ratings; Borman and Vallon (1974) found that the BARS technique yielded ratings that had better reliability and rater confidence in ratings, but that the graphic scales resulted in less leniency and better discrimination among ratees.

Burnaska and Hollmann (1974) compared three different formats. The first was the standard BARS format, the second format consisted of the same dimensions and definitions as the BARS scale, but used adjectival anchors instead of behavioral anchors, and the third was a traditional graphic rating scale. They found that leniency and composite halo were present in all three formats, although the BARS method reduced leniency and increased the amount of variance that was attributed to ratee differences. Similarly, Borman and Dunnette (1975) compared the standard BARS format to rating scales that had identical dimension labels and definitions, but with numerical anchors, and also traditional graphic rating scales that had trait labels and numerical anchors. Even though the standard BARS format exhibited less halo and leniency, and had higher reliability, they concluded that the differences in format were negligible since those differences only accounted for 5% of the rating variance.

In another study by Keaveny and McGann (1975), using ratings of college professors, the behaviorally anchored scale resulted in lower halo than the graphic rating scale, but the two scales did not differ in terms of leniency. The general conclusion of these authors was that neither format could be judged superior to the other. Similarly, Bernardin, Alvarez, and Cranny (1976) found that summated rating scales (a variation of the traditional graphic rating format) resulted in less leniency though maintaining greater interrater agreement when compared to BARS ratings. In a follow-up study, Bernardin (1977) used item analysis procedures to choose anchors for a BARS format and found no differences in the two types of scales.

Although use of this scale has been somewhat justified empirically (e.g., Landy & Gion, 1970), there have been several negative findings. Specifically, the process of developing BARS can be time-consuming and expensive (for a detailed description of the BARS technique see Schmitt & Chan, 1998), and there seems to be problems with identifying anchors for the central portion of the scale (Landy & Farr, 1980; Murphy & Cleveland, 1995). Additionally, according to Murphy and Cleveland (1995), the belief that BARS are more objective than a graphic rating scale and that defining performance in terms of actual behaviors would yield more accurate ratings has not been supported in research (e.g., Murphy & Constans, 1987). Nonetheless BARS continues to be one of the most predominant scales used. In addition to the feeling of personal investment from both the raters and ratees, BARS are also highly regarded because of their usefulness as a performance feedback tool for developing employees (Schmitt & Chan, 1998).

Mixed Standard Scales (MSS). Introduced by Blanz and Ghiselli (1972), the MSS had several appealing features. It also incorporated behavioral examples, but they utilized a different response format than BARS. Specifically, the MSS utilized three behavioral statements; one reflecting relatively effective performance, one representing midlevel (average) performance, and a third representing low level performance. Statements regarding behavior are randomized and presented to the raters. Raters are then supposed to indicate whether a ratee's performance is "better than", "as good as," or "worse than" the behavior described in a statement. Just as before, research was conducted to compare this newer format with those being currently employed. Specifically, Saal and Landy (1977) conducted a study comparing the MSS format to a graphic rating scale and a BARS. They found that the MSS resulted in lower halo than both the graphic scale and the BARS, but its reliability was "exceptionally" low. A lack of empirical support for the usefulness of MSS scales over others (e.g., BARS and graphic rating scales), as well as the complexity of the scoring system, has resulted in little use of MSS scales in the field.

Behavior Observation Scales (BOS). A final variation on the use of behavioral examples in evaluating performance is the BOS (Murphy & Cleveland, 1995). A BOS uses items that are similar in style to MSS item, but instead of asking for an evaluation of each ratee, this type of scale asks raters to describe how frequently particular behaviors occurred over a certain period of time. According to Murphy and Cleveland (1995), many of the proponents of this format claim that it removes much of the subjectivity that is present in evaluative judgments. Contrary to this, research has found that the cognitive processes involved in responding to BOS are just as subjective as evaluative judgments (Murphy & Constans, 1988; Murphy, Martin, & Carcia, 1982). Indeed Murphy and Constans (1988) found that behavior frequency ratings were more subjective than both trait ratings using graphic rating scales or overall judgments. Whereas the orientation of these scales appears to be an advantage, there are researchers who do not advocate for the use of this type of scale (e.g., Murphy & Cleveland, 1995). These researchers believe that the downfall of BOS is that raters rely on their overall, subjective evaluations to guide their responses; therefore it is more relevant to utilize a scale in which judgments of behavior are utilized instead of frequency of behavior.

Forced-Choice Ratings. An alternative to the direct rating schemes discussed already is forced-choice. Unlike the other formats, forced-choice scales were developed using the perspective that raters sometimes deliberately distort ratings. Deliberate distortion typically results in rating errors (esp. leniency). A forced-choice scale is a checklist of statements that are grouped together according to certain properties. The rationale underlying this approach is that the statements that are grouped have equal importance. Raters are forced to choose, from each group of statements, a subset that is the most descriptive of each ratee (Landy & Farr, 1983). Rather than using anchors (behavioral or otherwise), a forced-choice format derives a performance score based on a priori scale values assigned to the descriptors. Note this format is premised on the notion that raters do not know the a priori scale values when making their ratings.

Indeed, one of the main advantages of forced choice over the other "direct" rating formats is its resistance to leniency. According to Landy and Farr, this is mainly due to the fact that the rater did not know the preference and discrimination indices of the different descriptors. The property was demonstrated in a study by Lovell and Haner (1955) who found that even when raters were told to deliberately make ratees look good, there was little leniency present in the resulting ratings. There have been numerous other studies that have compared forced-choice formats to graphic rating scales and found that forced-choice formats demonstrated higher convergent validity with performance measures (Staugas & McQuitty, 1950), less leniency (Taylor, Schneider, & Clay, 1954), and less range restriction (Cotton & Stoltz, 1960). Similarly Sharon and Bartlett (1969) utilized student evaluations of college instructors and found that several variations of graphic rating scale all showed "significant" signs of leniency, whereas the force-choice format was uniformly resistant to leniency bias

Unfortunately, forced-choice formats are not widely used because they are seen as somewhat "constricting". There is evidence that raters prefer to use other formats because forced-choice formats do not allow them to know if they are rating their best people high and their worst people low (Bernardin & Beatty, 1984). In addition there is anecdotal evidence that there are ways to fake a forced choice appraisal (Popovich, 2007) by thinking of a high performing employee during all evaluations; however, there are no reported cases of this in the literature. Because of these drawbacks, the use of forced-choice formats is sparse even though it tends to eliminate the occurrence of rater errors (specifically leniency).

Summary. In general, the results of the comparisons of different formats are ambiguous. According to Cozan (1959), even though the forced choice formats demonstrated more validity and less range restriction than graphic rating formats, the cost of change to the organization is not beneficial. Indeed, he suggested that the traditional graphic scales be utilized over the forced-choice (mainly because of its simplicity). Whereas the findings comparing forced-choice and graphic rating scales appears to be relatively clear, the findings associated with behaviorally anchored scales (i.e., BARS, MSS) was much less clear.

Based on reviews of these different methods, the findings suggest that the psychometric superiority of the BARS is questionable at best (Borman, 1991). Although some studies showed that BARS had better psychometric qualities (in terms of rater errors and reliability) (e.g., Borman & Vallon, 1974, Campbell, et. al., 1973), other studies found partial support (e.g., Keaveny & McGann, 1975), and yet others found no differences in format at all (e.g., Bernardin, 1977; Bernardin, Alvarez, & Cranny, 1976). Indeed, based upon the findings it seems difficult to justify the increased investment in the BARS development process. According to Bernardin and Orban (1990), one reason for the lack of consistent findings may be due to variations that exist in the organizational context and/or rater characteristics that were present when these formats were compared. Rather than focusing on which scale is better, Friedman and Cornelius (1976) suggest that superior scales are a result of psychometric rigor during development rather than characteristics associated with the behavioral anchors.

As noted previously, much of the research dealing with the examination of rating formats has tapered off, although not halted (e.g., Stark & Drasgow, 2002). Hence, despite the widespread use of these various judgmental indices of performance, there has been a constant dissatisfaction with these measures on the part of both researchers and practitioners (Landy & Farr, 1980). Indeed, the main cause for this dissatisfaction is the vulnerability of these measures to both intentional and inadvertent bias (i.e., rating errors). As mentioned previously, this is not surprising because most format changes are merely structural changes and fail to account for the intentional distortion of responses. Due to the limitations and inconsistent findings from format research, the focus of attention for PA research subsequently turned to training that could be given to raters in the hopes of reducing the occurrence of rating errors.

Rater Training

Training raters to improve the accuracy of their performance evaluation has been the focus of numerous studies in the PA literature (Woehr & Huffcutt, 1994). Indeed, despite any empirical work at the time, Driver (1942) believed that training raters on performance evaluations was a fundamental step in the rating process. Rater-training research has focused on such aspects as reducing rating errors (e.g., Latham, Wexley, & Prusell, 1975), increasing accuracy (e.g., Woehr & Huffcutt, 1994), and providing raters with a common frame of reference (Sulsky & Day, 1992). In all cases, the rater training programs were designed to influence ratings by educating raters about key cognitive and observational demands of the rating process. Even though the research on this topic is substantial, the results are mixed. Indeed, although initially developed to help deal with the issues associated with format changes (i.e., inadvertent response distortion), rater training is an environmental (i.e. structural) change that fails to take intentional distortion into account. For instance, if a rater is going to give higher ratings to someone because he/she wants that person to get a raise, then no amount of rater error training will prevent that. Within the PA literature, there have been two main training programs that have been utilized and studied in the endeavor to generate more error free ratings (Borman, 1991; RET, FOR).

Rater Error Training (RET). Rater Error Training is a method that alerts raters to certain types of errors, either psychometric (e.g. leniency, halo, etc.) or perceptual (e.g., similar-to-me). This training program is often conducted with a lecture or demonstration of a particular error and a plea to avoid such errors when making performance evaluations (Bernardin & Buckley, 1981). Although most of the initial research utilizing RET found that training raters did reduce rating errors between trained and untrained raters (e.g., Vance, Kuhnert, & Far, 1978), the effect was typically only short term (Bernardin, 1977). Indeed Wexley, Sanders, and Yukl (1973) found that only extensive training was effective in reducing rating errors; a finding that was corroborated by several other studies (e.g., Latham, Wexley, & Prusell, 1975; Bernardin & Walter, 1977).

Even though rater training concentrated on the avoidance of typical rating errors (e.g., leniency, halo, etc.) there was very little evidence that knowledge of such errors showed any reduction of such errors in actual ratings (Landy & Farr, 1980). Because of this, the emphasis in rater training research and application has shifted away from training that is designed to reduce rater errors, and toward methods that increase a rater's ability to observe, recall, and classify behavior (Bernardin & Pence, 1980; Pulakos, 1984).

Frame-of-Reference training (FOR). Frame-of-reference training (Bernardin & Pence, 1980) attempts to demonstrate to raters that performance is a multidimensional construct and therefore familiarize them with the actual content of each performance dimension. The essence behind FOR is to provide raters with a common set of standards for evaluating their subordinates. The training program itself typically involves matching behavioral exemplars to performance dimensions and then maps those exemplars to a common evaluative rating scale. According to Murphy and Cleveland (1995), there has been consensus that FOR is better at allowing raters to accurately observe, recall, and classify behavior. FOR training helps raters become properly calibrated, so that ratings for individual dimensions have roughly equivalent meaning for all raters.

Research has provided evidence that FOR is a viable approach for teaching raters how to make distinctions between alternative levels of work performance (e.g., Day & Slusky, 1995). Overall, the research dealing with FOR training has consistently shown it to be effective in increasing performance rating accuracy (e.g., Athey & McIntyre, 1987; Bernardin & Pence, 1980; Day & Sulsky, 1995; Pulakos, 1984; Woehr, 1994).

Summary. Research has supported the notion that training raters to focus on not committing rating errors (i.e., RET) is successful in reducing various psychometric errors (e.g., Latham, Wexley, & Prusell, 1975). Rather than focusing on rating errors, FOR training has been shown to increase rating accuracy (e.g., Pulakos, 1984). Additionally, FOR has become more popular recently because of the commonalities it shares with cognitive modeling methods. Instead of utilizing one training method or another, several researchers have suggested that RET and FOR be used in conjunction with each other (e.g., Borman, 1991; Stamoulis & Hauenstein, 1993).

Just as with the work dealing with rating formats, the focus on rater training techniques has somewhat subsided, mainly due to the realization that structural changes were inadequate for dealing with rater errors. According to Borman (1991), researchers need to go beyond basic manipulations of format and training to understand the sequence that raters go through when making evaluative ratings. Indeed, two seminal papers by Landy and Farr (1980) and Feldman (1981), shifted the focus of PA research from research on formats and training to the understanding of the rater as a decision maker who processes social cues. Because of this “cognitive revolution” (Sulsky & Keown, 1997), PA researchers have spent a considerable amount of time and effort examining the process of performance ratings from a cognitive perspective.

Cognitive Approach to PA

Just as with other areas of cognitive research, PA researchers have focused more attention on how information processing errors affects ratings (Funder, 1987). Indeed, research has supported the notion that biases in the encoding and retrieval of information can lead to a variety of errors (e.g., Higgins & King, 1981; Ilgen & Feldman, 1983; Murphy, Balzer, Lockhart, & Eisenman, 1985). Although, this cognitive approach examines ratings (and subsequently rating errors) from a different perspective, researchers are still interested in figuring out ways to reduce rater errors (Murphy & Cleveland, 1995). Even though it is commonly known that humans have a limited capacity for processing information (March & Simon, 1959; Schneider & Shiffrin, 1977), researchers believed that evaluating another individual’s performance is a relatively simple one for raters to do well if they know how (implying bias arises from unintentional sources). Because of this, the main avenue for understanding the cognitive aspects of the rating process have utilized rating process models (e.g., Landy & Farr, 1980), which examine the rating sequence along with various factors that are believed to affect the process (i.e., rater and ratee characteristics). By combining findings regarding characteristics with social psychological concepts (such as attribution theory and personal construct theory), further conceptualizing and studying of the rating process could be done (Borman, 1991).

Rating Process Models. Evaluating performance from a rater's perspective can be construed as a process of cognitively processing information in order to make judgments or evaluations (Ilgen, et al., 1993). There are two traditional kinds of rating process cognitive models with somewhat different emphasis. The first type of process model focuses mainly on aspects of the rating sequence (i.e., observing, encoding, storing in memory, retrieving, judgment, rating) along with various factors that are hypothesized to influence the rating process (e.g., DeNisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980). The second type of model elaborates on the encoding step in the rating process and in particular considers categorization in processing performance-related information (e.g., Feldman, 1981; Ilgen & Feldman, 1983). Much of the current cognitive research draws on these two main types of models for hypotheses and interpretations of experimental findings (Murphy & Cleveland, 1995).

The first type of model attempts to specify the cognitive steps that take place during the rating process. The notion is that performance information is sought, encoded, and stored in "memory bins." Before evaluations are made, a rater makes judgments about various external influences on performance and how performance is indicative of a given ratee. These types of models see the rater as an active seeker of performance information (DeNisi, et al., 1984). According to Landy and Farr (1980), highly filtered information forms the basis of performance ratings, in that the actual behavior of a given ratee is influenced by several factors before “emerging” as a performance judgment.

The second type of process model (e.g., Feldman) incorporates the same cognitively-based sequencing process described in the previous model but also emphasizes several other features. First, these models elaborate on the categorization process (specifically where encoding takes place). The notion is that as raters become barraged with performance-related information about a ratee, they will simplify the information by categorizing it into dimensions (Borman, 1991). In addition to specifying the categorization process, these types of models distinguish between automatic and controlled attentional processes. Ilgen and Feldman (1983) make the point that when the patterns of ratee behaviors conform to previous impressions, that behavior is automatically categorized. This automatic categorization process implies that individuals are unaware of how they encode and store information. The issue with this approach is the same as those with the format changes and rater training. Specifically, most of the time individuals are not unaware of the ratings they give during a performance evaluation, but rather they intentionally give certain ratings.

Regardless of the particular process model that is utilized, research on the appraisal process focuses on the judgment component of ratings. Unfortunately, the criteria for evaluating judgments (especially in conjunction with process models), is somewhat vague. Traditionally, accuracy was indexed indirectly by examining various psychometric biases (i.e., rating errors). Accuracy was conceptualized as the absence of rating errors. More recently, many have utilized actual accuracy measures. A meta-analysis by Murphy and Balzer (1989) though showed no relationship between error measures and accuracy measures. As such, much of the research that has utilized cognitive process models has attempted to examine how various components affect the accuracy of judgmental ratings. For an extensive review of much of the literature that has utilized a cognitive approach see Ilgen and colleagues (1993).

Although these models have helped to clarify cognitive research in PA, according to some (e.g., Murphy & Cleveland, 1995), they have; 1) paid insufficient attention to the context in which appraisals occur, 2) failed to identify issues of concern for researchers and practitioners, and 3) failed to illustrate the links between the concerns of PA researchers and the practice of performance evaluation. Although cognitive process models are effective in depicting the ways in which unintentional biases arise, it did little to increase our understanding of intentional biases. Because of this, Murphy and Cleveland developed a goal-based perspective that views raters as active agents pursuing specific goals that introduce biases into performance ratings.

Goal-Based Perspective for PA

According to some researchers, it is possible that rater errors and other psychometric deficiencies in rating are not the result of limitations in a rater's capacity (i.e., errors), but rather reflect the effects of strategic decisions on the part of the raters (e.g., Murphy & Cleveland, 1995; Murphy, Cleveland, Skattebo, & Kinney, 2004). Specifically, Cleveland and Murphy (1992) have suggested that raters pursue different goals when they are conducting performance appraisals, and these goals may lead them to give ratings that appear to be psychometrically unsound. According to this approach, a rater who is lenient (i.e., gives high ratings), may not be making a judgmental error, but rather may be making a calculated decision that it is advantageous (usually for them) to give someone a higher rating (e.g., it may make supervisors look better because they have better performing subordinates).

According to Murphy and Cleveland, the cognitive appraisal process treated contextual variables as nuisance variables (if at all), whereas their goal-based approach treats appraisals as a communication and social process in which contextual variables are key. The central assumption is that appraisal outcomes are the result of a rater's goal-directed behavior, which is shaped by the organizational context in which ratings are occurring. Although this approach views goals as a main source of rating biases, they not are the sole source (Murphy et al., 2004). Indeed, the research based upon process models has provided evidence that there are many aspects to the rating process including personalities as well as cognitive processes of raters (e.g., Landy & Farr, 1980).

Although, there have been reviews of the literature that supports this approach (see Cleveland & Murphy, 1992; Murphy & Cleveland, 1991; 1995), there has been little empirical work examining this model. Murphy et al. (2004) were the first to conduct research that tested this goal-based perspective. They utilized teacher evaluations to determine if raters, who were pursuing different goals, were giving different ratings. They found significant multiple correlations, both within classes as well as in an analysis of the pooled sample (in which differences in the mean ratings were controlled for prior to estimating the predictive value of goal ratings; incremental $R^2 = .08$). Additionally, the measures of goal importance that were obtained after the raters had observed the ratee's performance, accounted for variance not accounted for by measures of goals obtained at the beginning of the semester (incremental R^2 = .07). According to the researchers, these results supported their belief that raters who were pursuing different goals tended to give different ratings, even when they had observed the same performance.

Although the Murphy et al. (2004) study did provide some evidence for the goal-based perspective, there were several issues with this research. First, the correlational analyses, although a longitudinal design, do not provide conclusive evidence for causation. For example, as noted by Murphy and colleagues (2004), it is possible that some students chose a particular class or instructor based upon their reputation (e.g., easy grader), which could influence the rating goals of interest. For example, students may seek out instructors who are lenient graders; therefore they (i.e., the raters) may report different rating goals from those who do not know anything about their instructor. Despite the apparent evidence for the goal-based perspective, the lack of experimental manipulations does not allow for an adequate examination of this phenomenon.

More recently, Wong and Kwong (2007) attempted to extend the work of Murphy and colleagues (2004). Their main issue with the previous work was that raters often rate more than one ratee; therefore there are two main features that should be utilized to examine rating patterns. The first (in their view) is the mean ratings, or the mean level of all ratings from a rater, and the second is the discriminability, which refers to the dispersion of ratings of different ratees from a given rater. This information was not available from Murphy et al. (2004), because each rater only rated one person. As noted by Wong and Kwong, the impacts of rater goals on mean rating could be different from the impact on discriminability. For example, a rater who wants to maintain harmony within a group may increase mean ratings, while decreasing discriminability. According to Murphy and Cleveland (1995), the discriminability index is particularly important for making personnel decisions (e.g., promotion, salary, etc.). Another issue that Wong and Kwong wanted to address concerned the differential influences of different rater goals. The idea was that different rater goals should yield qualitatively different rating patterns. The correlational design of Murphy et al. (2004) did not allow for this determination. Indeed, they were not able to determine whether raters gave more lenient or severe ratings when they pursued one goal versus another. By manipulating rater goals (identification, harmony, fairness, motivating) and requiring respondents to rate a group of ratees, Wong and Kwong were able to gain a more comprehensive understanding of the impacts of rater goals on rating patterns.

Specifically, Wong and Kwong (2007) utilized undergraduates who were working in 14 groups of 7 to 8 members on a project examining human resource practices in firms. The projects consisted of three main deliverables during the course of a semester. The first was a project proposal (October), the second was an oral presentation (November), and the third was a written report (December). Additionally, the participants were asked to provide peer evaluations at the semester midpoint and at the end of the project. They were told that these evaluations would be used to adjust final project grades. Rater goals were manipulated within participants within the same project. There were 16 possible combinations of condition orders and the authors tested for order main effects and interactions with other factors and found that condition order (which rating goal they were presented with first) did not have a main effect and did not interact with other variables of interest at either the mid-semester evaluations or the end-of-semester ratings. First, in the identification goals condition, raters were required to give ratings that identified strengths and weaknesses. In the harmony goal condition, raters were instructed to maintain group harmony. Then in the fairness goal condition, they were instructed to give ratings that fairly and accurately reflected individual contributions to the group. Finally, in the motivating goal condition, raters were asked to give ratings that would motivate group members.

Two separate analyses were conducted to examine the effects of goal conditions on mean ratings and discriminability. They found significant main effects for both mean ratings and discriminability. Specifically, they found higher mean ratings and lower discriminability overall when pursuing a harmony goal compared to pursuing an identification goal. Additionally, they found that pursuing a fairness goal resulted in higher mean ratings and lower discriminability during the project, but increased mean ratings and had no effect on discriminability after the project had ended when compared to pursuing an identification goal. These results provide further evidence for a goal-based perspective when examining performance evaluations. As noted by Wong and Kwong, giving goal instructions to raters may lead them to give ratings consistent with those instructions; therefore they suggest using them to address some of the problems observed with performance appraisals (e.g., rater errors).

Summary. The cognitive approach to understanding the performance appraisal process provided a much needed extension from methods of *dealing* with rating errors, toward an understanding of why they occur and what factors might influence those decisions. Although, this approach was a step in the right direction, some still felt that other contextual factors (i.e., goals) needed to be addressed as well (Murphy & Cleveland, 1995). According to Murphy and Cleveland's approach, raters are aware of the goals around them and make decisions (e.g., performance evaluations) to achieve those goals. This motivational, goal-based perspective has provided some promising results among the few empirical studies that have looked at the question. In particular, the results of the Wong and Kwong (2007) study were very promising. Their results beg the question of how researchers and practitioners will deal with the likelihood that ratings are influenced by different goals. To that end, the approach being presented in this paper will attempt to answer that question.

Regardless of whether formats are manipulated, training is given to raters, or the cognitive and motivational aspects of the performance appraisal process are examined, it is important to know whether the ratings that are given provide an accurate reflection of the performance of the individuals being rated. There have been many conceptualizations of criteria for evaluating ratings (e.g., psychometric, rating accuracy, rater errors; for a review see Murphy & Cleveland, 1995). The most common is an indirect, inverse measure of accuracy through the examination of the rater errors described earlier (e.g., leniency). The purpose of this paper is to present an alternative to detecting rating errors. In particular, leniency has been considered the most serious problem that has plagued performance appraisals (Murphy & Cleveland, 1995), and therefore is the focus of this paper. Because of this, the methods reviewed focus on the traditional methods for evaluating leniency/severity, and the limitations associated with those methods.

Methods for Detecting Leniency/Severity

Although discussions of the most common rater errors date back to the 1920s and 1930s, there have been a variety of conceptual and operational definitions of these various rating criteria (Saal, et al, 1980). As such, there are several different methods for evaluating any given rating error. For instance, there have been three typical methods for conceptualizing leniency or severity. By far the most popular (and most common) approach deals with the simple comparisons of average dimension ratings from the scale midpoint (e.g., Bernardin, Alveres, & Cranny, 1976). According to this approach, lenient ratings are reflected by mean ratings that exceed the midpoint, whereas severe ratings are reflected in mean ratings that fall below the midpoint. As noted by Saal and colleagues, the most common analysis of this approach is a basic comparison of group means for a given rater across ratees. Thus, if Rater X gives ratings of multiple ratees, and the mean rating is 5.0 (on a 7 point scale) and Rater Y also rates multiple ratees and has a mean rating of 4.1, Rater X is labeled as a lenient rater (Murphy & Cleveland, 1995).

Although mean ratings are the most popular and common method for determining leniency, they are averaged across all ratees, and therefore they may not be capturing the true nature of lenient ratings. Additionally, even if Rater X is designated as giving lenient ratings, there is no evidence that allows for an understanding of reasons behind such ratings. Because scores are averaged across all ratees, mean differences do not allow one to determine which ratees Rater X is being lenient with and why. As such, an approach that detects leniency on an individual level should be utilized.

A second, far less popular approach according to Saal and colleagues (1980) is based on a Rater X Ratee X Dimension analysis of variance. A significant Rater main effect, which accounts for a “sizable” proportion of rating variance, is evidence of leniency or severity (e.g., Friedman & Cornelius, 1976). Finally, a third approach that has been rarely used, examines the degree of skewness that characterizes frequent distributions of dimensions of ratings across ratees (e.g., Landy, Farr, Saal, & Freytag, 1976). Significant skewness is thought to reflect leniency or severity depending on the direction.

As noted by Murphy and Cleveland, although multiple operational definitions of a construct are usually considered desirable, it is not necessarily true regarding rater errors. Indeed, in a meta-analysis, Murphy and Balzer (1989) found that the average correlation between alternate measures of the same rater error was, r = .08. If this is indeed the case, it would suggest that there should be methods for choosing among the alternative methods. Unfortunately this is not the case and there appears to be no clearly defensible way of choosing one operational definition over another (Murphy & Cleveland, 1995). It may even imply the construct is unfounded. Alternatively, it could imply that none are adequate for the task. Specifically, the methods that have been used to examine rater errors may not be the best methods available.

The notion behind performance appraisal is that a rater rates a given ratee. The methods that are currently used to examine the quality of ratings (i.e., detect rater errors) do so from an aggregated perspective. In the case of leniency, mean ratings are the most common, but as noted by Wong and Kwong (2007), there is more going on than the mean scores can capture. Specifically, they suggest that discriminability (dispersion of scores) be utilized as well. As with many of the other methods (e.g., Rater X Ratee X Dimension ANOVA), scores are averaged across ratees, as well as raters. These analyses do not allow decision makers to understand who is committing errors on whom simultaneously. Additionally, none of the traditional methods that have been utilized allow for the understanding of reasons behind biased ratings. In particular, if a rater is lenient or severe, neither researchers nor practitioners know why. It seems apparent that the techniques that are currently employed are not sensitive enough to capture information that allows decision makers to determine who is giving biased ratings and why.

Recently, alternative techniques, based on an item response theory (IRT) approach, have been utilized to detecting rater errors. Item response theory is a statistical modeling approach (as opposed to an objective measurement approach) that has proven to be useful for psychological and cognitive measurement (Drasgow & Hulin, 1990; Embertson & Yang, 2006).

Utilizing IRT for Performance Appraisal

As noted by Drasgow and Hulin (1990), there are multiple organizational issues that can be addressed by an IRT approach. Within performance appraisal, there have been two main conceptual perspectives and subsequent methods for applying IRT. The first perspective deals mainly with the issue of measurement equivalence. This perspective examines two groups of raters and attempts to determine the degree to which ratings are directly comparable. With this approach a traditional IRT model is used to generate item parameters (usually item difficulty (b) and item discrimination (a)) for each of the groups that are being compared (see Hambleton, Swaminathan, & Rogers, 1991 for an overview of IRT models). To determine rater equivalence, the difficulty and discrimination parameters are compared with an IRT-based differential functioning approach, which will be discussed shortly (Raju, van der Linden, & Fleer, 1995).

The second perspective utilizes IRT-based models to examine rater effects. Specifically, this research tends to focus on the development and utilization of latent trait approaches for detecting rater effects (i.e., errors) (Wolfe, 2004). According to a latent trait approach, the probability of responding in a certain way is a function of some latent trait (e.g., ability) that is underlying performance (Crocker & Algina, 1986). The notion behind this approach is that rater errors are believed to be systematic, and therefore, they are detectable as patterns in the ratings assigned by raters. This approach is considered a model-fit (or person-fit) approach. By examining magnitudes of different parameters, various rater effects (i.e., leniency/severity) can be detected in different raters. Both the differential functioning approach and the model-fit approach are described below.

Differential Item Functioning. According to Hambleton, Swaminathan, and Rogers (1991), test fairness is one of the most highly charged issues surrounding testing. These notions of test fairness are parallel with the literature that has taken an IRT perspective to investigate item bias and differential functioning. In particular, an IRT perspective examines the relationship between an individual's item performance and the set of traits underlying item performance. This relationship can be described by a monotonically increasing function called an item response function (IRF), typically referred to as an item characteristic curve (ICC) (Hambleton, et al., 1991). An ICC focuses on how the probability of a “correct” response to an item is related to an individual’s ability (or some underlying trait) and the item’s properties. Typically, ICC functions (Figure 1) take an “S” shaped curve, which closely resemble normal ogive, or logistic curves (Crocker & Algina, 1986).

When tests are used with different groups of respondents (e.g., peers vs. subordinates); there may be instances in which items on that test function differently for members of each group. Groups of respondents who have different ICCs are believed to be responding differentially (e.g., the probability that one group member gets the item correct is different than the probability that a member from another group gets the item correct when both members have the same ability). As noted by Schmitt and Chan (1998), an item is biased only when “equally able” members of the two different groups have unequal probabilities of getting a particular item correct. In such a situation, differential item functioning (DIF) is said to exist. Although some argue that an item shows DIF when the groups being compared (i.e., focal and referent groups) differ in their mean performances on an item, this does not take into account the possibility that other variables, such as real ability differences between groups, may be responsible for significant differences (Hambleton et al., 1991). As such, mean ratings demonstrate impact, but they should not be the determining factor for bias, whether item bias or rater bias (i.e., leniency/severity). Rather, when groups have different probabilities of getting an item correct then the item is considered biased. It is important to note that DIF analyses examine differences on an item across groups of respondents.

As noted by Schmitt and Chan (1998), there are many different ways of detecting DIF (see Camilli & Sheppard, 1994, for a review). When utilizing DIF techniques, two well-defined groups are required. In the performance appraisal literature, peer and subordinate are common examples when attempting to determine measurement equivalence. Once groups are defined, there are two general techniques for identifying DIF. The first technique is based on the estimates of the latent parameters and the second technique is based on the observed scores (Embretson & Yang, 2006; Hambleton et al., 1991).

The latent parameters method uses the item and ability parameters from the different groups to detect DIF. According to this approach, the person and item parameters are estimated for each group first (through the application of an IRT model – 1PL, 2PL, or 3PL) and then they are placed on the same metric. If, after placing them on a common scale, the ICCs are identical, then the area between the curves should be zero, but if they are not zero, then DIF is present (Rudner, Getson, & Knight, 1980). The general idea is that there should be no significant differences in the item parameters for the comparison groups.

From this approach there are several ways to identify DIF (Cohen, Kim, & Wollack, 1996). The item parameters can be compared using a chi-square test (e.g., Lord, 1980). A non-significant chi-square indicates that the item parameters are not different for the various groups. Another approach uses area measures to compare the expected scores for examinees at the same level of the latent trait from the different groups (Raju, van der Linden, & Fleer, 1995). A common method for measuring the distance between ICCs is to use differential functioning of items and tests (DFIT; Raju et al., 1995). DFIT is a technique that compares the expected scores [ΣP(θ)*test length] for examainees at the same level of the latent trait from the focal and referent groups (Raju et al., 1995). In a performance appraisal context, measurement equivalence is determined when there are no significant differences between the expected scores for the groups being compared (e.g., peer vs. subordinate). Failing to reject the null hypothesis would indicate that there were no differences in the probability of responding for the two groups, therefore their ratings are equivalent.

As noted previously, there are very few studies that have utilized IRT approaches (specifically DFIT) to determine the equivalence of different types of raters (using the same inventory). Maurer, Raju, and Collins (1998) utilized confirmatory factor analysis (CFA) along with DFIT to examine the measurement equivalence of ratings between peer ratings and subordinate ratings. They used IRT models to develop the person (i.e., ability [θ]) and item parameters (i.e., difficutly [b], discrimination [a]). The parameter estimates were calculated for each group and then compared using DFIT statistics. This procedure essentially compared the difficulty and discrimination parameters for each group and found that there was not a significant difference between peer ratings and subordinate ratings.

Facteau and Craig (2001) used a similar methodology to test for invariance of the rating instrument that was used across self, peer, supervisor, and subordinate raters. The multiple group confirmatory factor analysis indicated that the instrument used was invariant across all rater groups, and the IRT analysis provided some evidence of DFIT, but was limited to three items and was, according to the researchers, trivial in magnitude. Facteau and Craig concluded that the instrument they used (i.e., multisource appraisal form) was invariant across groups, thereby supporting the practice of directly comparing the ratings from these sources.

Additionally, Barr and Raju (2003) utilized three different IRT models to examine rater equivalence in a multiple-source feedback instrument. By using data from managers who responded to the Benchmarks survey (Lombardo & McCauley, 1990), Barr and Raju found that the traditional DIF approach provided the most information about the rater's conception of the ratee's ability. This approach examined rater effects (leniency/severity) examining rater shift parameters. This approach merely examined group equivalence across various subscales within the Benchmarks survey.

An issue that is common to the IRT approaches described above is that they require the *estimation* of item parameters for all groups that are being compared (Hambleton et al., 1991). Because estimating parameters is not ideal, the second approach (i.e., observed score method) uses the actual responses of individuals in the different groups to detect DIF. The general idea behind this approach is that there should be no relationship between group membership and the response to an item after controlling for the trait. According to Hambleton and colleagues, the most popular method for detecting DIF is the Mantel-Haenszel method (Holland & Thayer, 1988). This relationship is assessed through the use of an odds ratio. The odds ratio reflects the odds that an individual in the focal group will agree with a given item when compared to an individual in the reference group (or vice-versa). The Mantel-Haenszel method has not been utilized within the performance appraisal literature. Therefore this paper also extends the performance appraisal literature by using the Mantel-Haenszel method for detecting rater effects.

Model-Fit Approach (e.g. Person-Fit). The general purpose of the model-fit (i.e., person fit) technique is to identify individuals whose response patterns are inconsistent with their estimated level of a latent trait, as measured by the whole test (Drasgow & Hulin, 1990). Assuming that the model fits the data, there is a family of appropriate measures that can be used to identify the specific individuals whose response patterns do not fit the model (see Drasgow, Levine, & McLaughlin, 1991 for a review). Within the literature examining rater effects (i.e. rater errors) the Multifaceted Rasch Rating Scale Model (MRRSM) is typically used to generate data. The MRRSM uses the log of the odds (i.e. logit) of observing one rating scale category versus the next lower category (assuming a polytomous scale) using parameter estimates that represent the raters and items. Mathematically, this model can be expressed with the following equation:

$$LN\left(\frac{P_x}{P_{x-1}}\right) = \theta_n - \lambda_r - \delta_{ik}\,, \tag{1}$$

where P_x is the probability of an individual being rated x on some domain by a rater [note: the IRT models being used for this approach utilize polytomous scales (i.e., Likert-type); therefore x represents some category (e.g., Strongly Agree), while x-1 represents the previous category (e.g., Agree)], P_{x-1} is the probability of an individual being rated x-1 on some domain by a rater, θ is the location of the individual (n) being evaluated on the underlying continuum, λ is the severity of a rater (y), and δ is the difficulty parameter item (i) of the domain being evaluated. By examining the relative magnitude of the λ estimates for a particular set of ratings, raters who are being lenient or severe relative to the pool of raters can be identified. Standard errors can be estimated for each parameter, and Wald statistics can be computed to identify raters who deviate from a group mean.

$$\chi^2_{Wald} = \left(\frac{\lambda_r}{SE_{\lambda_r}} \right)^2 \quad (2)$$

Additionally, model-based expected values can be computed for each rater-by-measurement object-by-item combination. The observed ratings can also be compared to expected ratings to determine various rater effects (for a detailed discussion of the formulas and rationale see Engelhard, 1994). The residuals of observed ratings from these model-based expectations (X_{nir} - E_{nir}) can be used to identify rater effects (Wolfe, 2004). The general notion is that departures in the data from model-generated expected values indicate potentially misfitting raters.

There have been several studies that have utilized these model-fit approaches to examining various rater effects. In a two-part paper Myford and Wolfe (2003; 2004) introduced the idea of using the many facet Rasch measurement (MFRM) approach for detecting and measuring rater effects. The first paper (2003) provided the background and context for using the MFRM, whereas the second paper (2004) utilized a special *Facets* program to study several rater effects (e.g., leniency/severity, central tendency, randomness, and halo). In each case, data was generated that would simulate raters behaving (i.e. rating) in various ways (e.g., lenient, severe, halo, etc.) and then utilized their model to detect which raters were committing these errors. In another study, Wolfe (2004) utilized a similar latent trait model-fit approach to examine several rater effects. Again data was generated and a model was used to determine which raters did not fit the proposed model. In both cases the research that was conducted to demonstrate how this approach could be utilized to identify and examine various rater effects used simulated data. Although Myford and Wolfe discussed the other methods for detecting rater errors (e.g., mean differences, Rater X Ratee X Dimension ANOVA), they merely sought to demonstrate the usefulness of the MFRM. So, even though using a multi-faceted Rasch model or some other IRT approach for examining and detecting rater effects may be promising, there has been no empirical evidence demonstrating its usefulness. Additionally, this approach merely identifies those individuals who do not fit a particular response pattern, but provides no explanation as to why they are responding in such a manner. Although a step toward progress, these approaches add little to the understanding of rater effects.

Limitations of Current IRT Approaches

Within performance appraisal research, there are two approaches to utilizing IRT. The first approach uses IRT models and a DIF analysis to examine the differences that exist between the ratings of different groups to determine their degree of equivalence. Unfortunately this approach does not have much bearing on the detection of rater errors. The second approach uses IRT models from a model-fit (person-fit) perspective to determine if a rater is committing various rating errors. Despite the potential advantages of the IRT approaches for detecting rating errors in performance appraisals, the empirical work (at least to date) is vague in terms of the degree to which potential advantages can be achieved. As was mentioned above, the various studies that have utilized IRT techniques have been able to determine rater equivalence as well as identify scoring patterns that indicate rater effects, but these results still do not help to address the central issue regarding rater errors. Specifically, not only who is giving biased ratings, but why. I argue that this is due to potential limitations to these approaches. These limitations are discussed below.

The first limitation deals with the type of information that can be gained and therefore used from a practical sense. The DIF approach (and DFIT) does a good job of providing information about items and how groups of raters respond to those items, but provides very little information about individuals. From a performance appraisal context, evaluations are done by individuals, not groups. Therefore, the information may be useful when determining what ratings one should collect, but again, in most instances ratings on a single target are done by individuals (i.e., a peer, a supervisor, a subordinate), not averaged ratings from groups of individuals (360 degree assessment is an exception). Although even with 360 degree assessment, it is useful to examine each individual’s ratings separately. Even though different raters (e.g., peers, self, subordinate) have different perspectives, it would still be helpful to be able to determine if a rater is introducing biases into an evaluation. Additionally, the research using DIF (and related concepts) has yet to be utilized to address issues related to rater errors. Even if one determines that various sources are/are not equivalent, there is no guarantee that they are not committing biased ratings. Additionally, there is no understanding as to why they may be committing those errors, such as item properties.

The model-fit approach provides information about each individual, but little (if any information) about the items that those individuals are responding to. In a performance appraisal context, we can determine if the response pattern of a given individual is inconsistent with a model estimate for a latent trait in relation to an entire inventory. The fit indices that are used only indicate that, given the data that was generated, the estimated parameters are inappropriate. However, a model-fit approach does not provide any insights as to why there is a lack of fit. Merely detecting aberrant response patterns is inadequate for effectively understanding rater errors. Indeed, this approach is a sophisticated "mirror" of the traditional techniques used previously. Additionally, from a practical standpoint, the research using the model-fit approach is based upon simulated data. Researchers have yet to apply such models to actual data to determine if their techniques are still viable.

Another issue with a model-fit approach is that researchers may encounter an inability to link item properties to aberrant responding. Because inconsistent patterns may not occur across an entire inventory, one may not be able to determine if the response pattern is due to a rater committing an error, or if it is actually due to some other reason. Indeed, as an example, Zickar and Drasgow (1996) found that the item content played a role in which items were responded to dishonestly on a personality inventory. Although, a model-fit approach allows for the identification of individuals who may potentially be committing errors, the inability to link item properties does not allow us to understand why they are distorting their responses. It seems apparent that more research is needed, which attempts to link the item to the functioning of that item.

From a decision making standpoint, it would be ideal to have a procedure that combines information on the properties of each individual as well as the properties of each of the items. This is not only true for issues related to performance evaluation, but also for issues related to test fairness. Indeed, as noted previously, the literature on item bias and differential item functioning parallels the literature on test bias (Schmitt & Chan, 1998). Just as in personnel issues of test bias, the combination of this information would allow for the detection of those who are committing rating errors (such as leniency), and why they are doing it. Fortunately, a recent approach has been developed that combines information from items and individuals, and also links the item properties to response patterns. This approach, called differential person functioning (DPF; Johanson & Alsmadi, 2002), was developed mainly for educational assessment, but has also been utilized for organizational issues (e.g., response distortion on personality inventories; Scherbaum, 2003).

Differential Person Functioning

Differential person functioning (DPF; Johanson & Alsmadi, 2002) was developed to enhance the diagnostic assessment in which individuals' scores between groups of items are narrowed by classifying the scores on item difficulty. Specifically, the primary purpose is to determine if individuals (i.e., raters) have different response patterns between groups of items. Inherent in this approach is that information about both individuals as well as items is combined. Additionally, DPF allows for the linking of item properties to various response patterns. Although Johanson and Alsmadi note that DPF is similar to person-fit, they note that it is “reasonably unrelated” to model-fit and their differences are addressed below. Specifically, DPF overcomes certain limitations of both IRT and traditional approaches for detecting rater errors such as leniency.

As noted by Johanson and Alsmadi (2002), the DPF technique is an extension of DIF, therefore they are very similar conceptually. Whereas both methods examine how item responses differ in relation to individuals, DIF focuses on how different person groups (peers vs. subordinates) respond to an individual item, and DPF focuses on how an individual responds to different groups of items (e.g., focal vs. referent). The DPF approach essentially takes the matrix of data for a DIF analysis and transposes (i.e., rotates) it so that the differential functioning of an individual is examined rather than the differential functioning of an item. Therefore, instead of examining item characteristic curves, a person characteristic curve (PCC) is examined. Figure 2 shows a hypothetical situation in which a given rater rated an individual on two different types of items (focal and referent). Because the two curves are not identical there is some degree of differential functioning. If the distance between the curves is great enough then this rater is considered to be responding differentially.

Additionally, as noted previously, DPF is similar to a person-fit approach in that both attempt to identify when individual’s estimates do not fit a given response pattern. The main difference lies in that DPF does not provide information about an entire test (as does a person-fit approach), but rather separate groups of items (e.g., focal vs. referent). Fit measures merely determine if estimated parameters are appropriate given a set of simulated data. This information does not help to understand why they are appropriate/inappropriate. Because DPF allows for the link between individuals and item properties, the results are more interpretable than person-fit measures.

In terms of dealing with rating errors, such as leniency/severity, DPF also provides several advantages over traditional methods. As noted previously, the most common method of detecting leniency is through the use of mean scores. Although some have argued that mean differences between groups is enough to indicate differential functioning, the predominant view is that differential functioning can only be determined when there is a different response probability for those at the same level of some underlying trait (e.g., ability) (Camilli & Shepard, 1994). As noted by Johanson and Alsmadi, mean differences merely show impact. It is not unusual to find impact, but not have differential functioning (Johanson & Alsmadi, 2002). Indeed, in many instances impact simply reflects actual differences between focal and reference groups. For example, person impact might be used to describe a person who simply agrees to a different extent with, say, instructor items or course items (an overall mean difference), but when the probability of responding to instructor versus course items is different after they have been conditioned by some overall measure of item agreement (similar response probabilities) then they are responding differentially. Hence, this would suggest that mean differences are an inappropriate method for detecting true rater errors.

From an empirical standpoint, the most often recommended method to detect DIF (and subsequently DPF) is the Mantel-Haenszel chi-square procedure discussed previously. The Mantel-Haenszel procedure (Dorans, 1989; Mantel & Haenszel, 1959) is a non-IRT based statistical procedure that examines the relationship between two variables in a 2 X K frequency table, while controlling for the level of a third variable, where K represents the number of subgroups of items. For each level of K, a 2 X 2 frequency table is formed by crossing the person's response with group membership. An overall odds ratio is then compared from the comprehensive K X 2 X 2 table as a measure of effect size. In terms of a DPF analysis, the relationship between the type of item and the response for a given individual (after controlling for overall agreement) is being examined. Significant chi-square values indicate that an individual is responding differentially over the different groups of items. Because this is used to detect differentially functioning persons, this analysis is done for each person.

There have been only a handful of studies that have utilized the DPF approach. Johanson and Osborn (2004) used DPF to examine the differential responding of individuals to positively and negatively worded items in an attempt to detect respondent aquiescence. Based upon their analyses, they were able to determine which individuals were acquiescing to different inventories and then remove them for analytic purposes. Indeed, when an item displays DIF it is removed from an inventory, likewise if individuals are displaying DPF, they too could be removed to allow for a more accurate interpretation of the analyses. Scherbaum (2003) utilized DPF to detect response distortion (i.e., faking) on personality inventories. This research expanded DPF to be used with polytomous item scoring and compared DPF with other traditional approaches to detecting faking. He found that DPF was a more sensitive technique than other methods for detecting response distortion on personality inventories and had comparable levels of accuracy to other measures. Additionally, this research provided an explanation as to why individuals were distorting. In this sense, the reasoning behind response distortion was due to the instructions that were given. Specifically, individuals were told to try to “fake good”, while others were told to be as honest as they could. In a way this is similar to giving each of the raters a different goal to pursue. As such, Scherbaum was able to determine who was faking and who was not. Similarly, Scherbaum (2005) used DPF to detect differential responding in biodata items. This study was able to identify individuals who were responding to the biodata inventory differentially as a function of item attributes (i.e., verifiable vs. non-verifiable).

Finally, Johanson and Alsmadi (2002) were the first to publish research on DPF, in which they examined the differential responding of students on the mathematics section of the California Achievement Test. They provided several examples of students who were responding differentially to demonstrate the potential utility of using the DPF technique. The DPF technique has not been applied to the problems of rater errors, nor has it been utilized in its traditional dichotomous approach for such a phenomenon.

Current Research

The current research is an attempt to direct the performance appraisal literature in a new direction. In particular, the purpose of this research is two-fold. First and foremost, the ability of the DPF method for detecting both the incidence and nature (i.e., reason) of rater bias will be examined. By comparing the sensitivity of the DPF method to detecting bias across experimental conditions with the sensitivity of other traditional approaches, and by comparing the consistency of classifying biased raters between methods, the effectiveness of the DPF method can be tested.

In terms of detecting rater errors, leniency is targeted given that it is considered the most serious rater error (Murphy & Cleveland, 1995). Due to the structure of the data (multiple raters and one rate), a comparison of DPF to all of the traditional techniques for detecting leniency (e.g., Rater X Ratee X Dimension ANOVA) cannot be done; however, the mean score method and skewness ratings can be calculated and compared with the DPF method in this case.

Secondly, this research will add to the small but growing literature demonstrating the effect of rater goals on rating tendencies. By having participants respond under different instruction sets, the effects of different goals can be tested. This experimental manipulation allows for a test of the sensitivity of each detection method for detecting bias between conditions. Specifically, the incidence of leniency will be examined in regard to the goals that raters were assigned (i.e., Administrative vs. Feedback). This manipulation will also provide a method to validate the DPF method. Without an effective manipulation, it is difficult to determine if the method was effective or not. In order to determine if the methods were accurate, the goal manipulation must work. Because of this I hypothesized the following:

H1: Raters pursuing an administrative goal will give higher ratings than raters who are pursuing a feedback-related goal or a control.

H2: A differential person functioning analysis will be more effective at detecting lenient raters than traditional methods, (e.g. mean scores, rater skewness) in terms of sensitivity.

As mentioned in the description of DPF, this procedure is unique because it takes into account information about individuals as well as item properties. Just as with a DIF analysis, focal and referent groups are needed, but instead of focal and referent groups of people, a DPF analysis uses different groups of items.

Because the purpose of this research is to examine rater errors, it is imperative that actual performance ratings be utilized. Specifically, teaching evaluations were used in this research. Student's ratings have been used as performance evaluations in numerous PA-related studies (e.g., Centra, 1976; Murphy, Balzer, Kellam, & Armstrong, 1984; Murphy Cleveland, Skattebo, & Kinney, 2004), and are still the primary mechanism that is used to assess both instructor and course effectiveness (Barnett & Mathews, 1998). Indeed, many evaluations contain items related to the instructor as well as the course (Aleamoni & Hexner, 1980). Because of this, instructor items and course items served as the focal and referent groups for the DPF analysis. Research has shown that raters tend to give higher ratings on items measuring instructor effectiveness (i.e., leniency) and lower ratings on items measuring course effectiveness (e.g., Aleamoni & Gary, 1980; Aleamoni & Hexner, 1980; Kidd & Latif, 2004; Phipps, Kidd, & Latif, 2006). Because of this I hypothesized the following:

H3: Instructor ratings will be higher than course ratings across all conditions.

Finally, the goal-based perspective suggests that researchers should attempt to determine what goals are relevant during a given evaluation period as a way to determine the motivational factor behind a given set of ratings. Indeed, researchers have noted that when ratings are to be used for administrative decisions like tenure and pay, raters tend to be more lenient, whereas when ratings are to be used for training and development (i.e., feedback), raters tend to be more severe (e.g., Aleamoni & Hexner, 1980; Bernardin, Orban, & Carlyle, 1981; Centra, 1976; Murphy & Cleveland, 1991; 1995; Sharon & Bartlett, 1969; Taylor & Wherry, 1951). Because of this, I hypothesized the following:

> **H4:** The proportion of differentially lenient raters (as determined by DPF) will be higher for raters who are pursuing an administrative goal than those who are pursuing a feedback-related goal or no explicit goal (i.e., the control group).

Additionally, there are several individual difference variables that will be examined. These variables include liking of the instructor, sex, and the goals they reportedly pursued during the evaluation. Given the lack of theories regarding individual differences and rating biases, these will be exploratory in nature.

Present Studies

To that end, two experimental studies were conducted where individuals completed performance evaluation forms under different instructions (i.e., rating goals) (e.g., rate for pay, promotion, tenure; rate for development; rate for evaluation). The first study was a pilot study that was used to determine the reliability of the newly developed evaluation forms, which would help to determine which format should be used, check for sex differences (both mean differences and DIF differences), and check the effectiveness of the goal manipulation. Based on these results, the main study instruments, materials and procedure were modified. The second study was designed to test the hypotheses stated above. These two studies and their results are presented below.

Pilot Study

Method

Participants

For the pilot study, 137 undergraduate students from Ohio University were recruited from introductory psychology and statistics courses (except for one of the introductory psychology courses that was to be used for the Main Study) during the Fall Quarter 2007. There were 39 males (28.5%) and 98 females (71.5%). Students completed an online evaluation form in exchange for one experimental research credit.

Manipulations

Response Format. Participants received one of two different evaluation formats, dichotomous responses (Appendix A) or Likert-type responses (Appendix B) with identical item content. Even though the Mantel-Haenszel (MH) procedure (e.g., Dorans, 1989; Holland & Thayer, 1988; Mantel & Haenszel, 1959) utilized dichotomous scoring, it was believed that respondents may prefer more than two options. Even if the Likert-type response format was chosen, the responses would be dichotomized. This would allow the procedure to be used appropriately with dichotomized data.

Response Instructions. The response instructions given to the participants were manipulated as a way of manipulating rater goals (e.g., Wong & Kwong, 2007). Participants were presented with one of three different response instructions (i.e., goal manipulation) and were instructed to provide ratings according to those instructions. The instructions were presented after introducing the study. All participants were told that they were piloting a new evaluation instrument and that they were not required to participate. However, each participant received one of three instructions sets, or goal conditions. One goal condition indicated that ratings would be used for administrative purposes (i.e., pay, promotion, tenure). The specific instructions for the administrative condition were as follows:

> "You are being asked to fill out this form in order to provide the Psychology Department with a performance evaluation for this course and instructor. Your responses provide information that the department will use for tenure, promotion, and salary considerations for this instructor. These evaluations are a crucial aspect when making tenure, promotion, or salary decisions within the department. Careful responding is important to make evaluation results informative and useful. Thank you!"

A second goal condition indicated that ratings would be used for feedback purposes for the instructor to change the course or themselves. The specific instructions for the feedback-related condition were as follows:

> "You are being asked to fill out this form in order to provide the Psychology Department with a performance evaluation for this course and instructor. Your responses provide information that will be used by the instructor to improve the course as well as themselves as instructors. If needed your feedback may be used by the instructor to use/develop different approaches that allow him/her to provide a better educational experience. Likewise, if required, the course may be reevaluated to more appropriately address its objectives. Constructive feedback is

> greatly appreciated and indeed the instructor’s job and/or reputation is not at stake, so please respond to each question as honestly as possible. Thank you!”

Finally, the third condition was meant to serve as a control condition and presented the standard instructions that the department of psychology uses for evaluations. The specific instructions for the control condition were as follows:

> “You are being asked to fill out this form in order to provide the instructor and the Psychology Department with a performance evaluation for this course. Your responses provide information which the instructor may use to improve himself/herself or the course. In addition, the department uses the information in tenure, promotion, and salary considerations. Careful responding is important to make evaluation results informative and useful. Thank you!”

Evaluation Form. Participants responded to an online evaluation questionnaire. These evaluation questionnaires consisted of 43 items related to the instructor (e.g., the instructor is knowledgeable in the field) as well as 40 items related to the course (e.g., the course was well organized). There has been research that indicates an acceptable sample size for a differential functioning analysis is as few as 50 items (Fidalgo, Ferreres, & Muniz, 2004). In such instances, the alpha level needs to be adjusted to allow for significance at the .20 level. In this case, an approximate number of 40 items per focal and referent group were developed (total of 83) with the anticipation that some may be eliminated because of poor functioning.

The item content of the evaluation form is based on extensive research that has been conducted to determine appropriate content domains for evaluating teaching effectiveness (e.g., Kapel, 1974; Feldman, 1996). Although there has been research that has utilized as many as ten content domains (Flowers & Hancock, 2003) and as few as five content domains (Kapel, 1974), the majority of research has relied on seven main content domains. According to this research, the relevant content domains that should be covered for evaluating instructor performance include: 1) Intellect/Knowledge, 2) Motivation/Learning/ Stimulation of Interest, 3) Preparation and Organization, 4) Student Development, 5) Presentation, 6) Personality, and 7) Evaluation (Damron, 1996; Feldman, 1996; Flowers & Hancock, 2003; Kapel, 1974; Phipps, Kidd, & Latif, 2006). In addition to instructor domains, there has also been research that has determined specific course domains that are relevant when evaluating course effectiveness (Feldman, 1978). These five content domains consist of: 1) Organization, 2) Course Level/Difficulty, 3) Goals/Objectives, 4) Subject Matter, and 5) Evaluation (Feldman, 1978). Based upon both the instructor and course content domains, items that have been used in existing evaluation forms as well as additional items were utilized to develop the evaluation forms that were used. The same items were utilized for two different response formats (i.e., dichotomous; Appendix A, and Likert-type, Appendix B), thereby resulting in six different formats. Each participant was asked to respond to only one format (i.e., between subjects). There were 7 instructor items that were reverse scored and all items were recoded so that higher scores demonstrated higher agreement (i.e., more desirable scores).

Measures

Evaluation Process and Format Reactions. A short questionnaire was designed to gauge rater's attitudes towards the questionnaire format that was presented. Although forced-choice formats have been shown empirically to reduce leniency, there has been noted dissatisfaction with the format from the rater's perspective, thereby leading to a lack of use. This questionnaire did not contain any forced choice items, but was rater reactions questions to determine if the format and process used were acceptable alternatives to the current methods and procedures. Six questions were designed to determine the degree to which rater's were satisfied with this evaluation process and format (Appendix C).

Goal Importance Questionnaire. This is a 19-item questionnaire that was developed by Murphy, Cleveland, Skattebo, and Kinney (2004) to assess the goals that raters pursue when evaluating instructors. Each item is rated on a 5 point Liker-type scale (1 – Strongly Agree to 5 – Strongly Disagree). This questionnaire was utilized to determine if raters were pursuing other goals in addition to the one assigned during the experiment (Appendix D). If there were certain goals that raters indicated as particularly influential in their rating process, they would be considered for use as the main manipulation for the main study. Additionally, the use of this questionnaire would allow for an examination of possible relationships that exist between ratings and goals.

Additional Items. Participants were also asked several additional questions at the end of the evaluation form. Specifically, one item asked them to indicate how their ratings would be used. This was an open-ended item and served as a manipulation check for the goal manipulation. Participants were also asked whether or not they liked the instructor, their sex, as well as the grade they expected to earn in the course.

Procedure

Participants were recruited through the psychology experiment subject pool beginning in the third week of the Fall 2007 term. The Pilot Study was administered online and students in five psychology 101 (Introduction to Psychology) courses and four psychology 120 (Fundamental Statistics) were targeted. Participants were assigned to one of six versions [i.e., Administrative: Likert-type (1) and dichotomous (2); Feedback: Likert-type (3) and dichotomous (4); and Psych. Department: Likert type (5) and dichotomous (6)] based on the first digit of their university email account. Participants were sent a message indicating that they matched criteria specific to this study and indicated that they could participate in the online study to receive one research credit. By opening the provided URL, they were directed to their respective version of the evaluation form.

Once directed to the online evaluation form, participants were presented with a description of the purpose of the study (Appendix E). Specifically, they were told that the study was designed to develop and refine a new evaluation instrument that may be used for evaluating instructors. They were instructed to consider their respective course (either PSY 101 or PSY 120) when completing the evaluation, and to answer all questions. Participants were then asked to indicate that they understood the purpose of the study.

Next participants were presented with a consent form (Appendix F) and were asked to indicate that they had read and understood the consent form and wished to continue. Finally, participants were presented with their assigned instruction set and were asked to indicate that they had read and understood the instructions for the evaluation. All participants indicated that they 1) understood the purpose, 2) had read and understood the consent form, and 3) had read and understood the instructions.

Participants then responded to the materials. The order of the measures was kept constant. They first completed the evaluation form and additional questions, then the evaluation process and format reactions questionnaire, and finally the Goal Importance Questionnaire. Finally, before participants were asked to submit their responses, they were asked to indicate how long it took them to complete the entire evaluation process. Across instruction sets, the dichotomous format took 20.71 minutes on average and the Likert-type format took 19.89 minutes on average to complete. Upon submitting their responses, participants were thanked for their participation in the study and were awarded their research credit.

Results

Reliability Analysis

The internal consistency was examined for both formats. The traditional Mantel-Haenszel procedure requires dichotomous data; however there was a possibility that raters would prefer a Likert-type format versus a dichotomous format. Because of this, both formats were used and the internal consistency was calculated for each. Results indicated that the overall reliability for the dichotomous version was α = .88, and for the Liker-type format was α = .94. Because the evaluation form consisted of two distinct types of items, the internal consistency of each subscale was also examined. Results indicated that the internal consistency for the instructor item scale was α = .77 for the dichotomous format and α = .85 for the Likert-type format. The internal consistency for course item scale was α = .82 for the dichotomous format and α = .93 for the Likert-type format.

Manipulation Check

The effectiveness of the response instruction (i.e., goal) manipulation was also examined. Although, previous research has indicated that using different instruction sets yield different response patterns, there was some concern that the manipulation was not strong enough. This manipulation check was done in two ways. Although the participants were presented with different instruction conditions and were asked if they had read and understood the instructions, an open-ended item asking participants what their ratings would be used for was included at the end of the evaluation form and served as a manipulation check. An examination of this open-ended item by the researcher indicated several trends. Of those students who responded to the open-ended item (90 out of 137 – 66%), 71% (i.e., 24 out of 34) of those in the feedback-related condition indicated that the ratings would be used for constructive feedback or improvement, whereas only 14% (3 out of 22) of those in the administrative condition correctly indicated that the ratings would be used for pay, promotion, tenure decisions, etc.

In addition to the content analysis, an ANOVA was conducted to examine possible mean differences. An overall ANOVA across formats indicated no significant differences between any goal conditions, $F(2, 134) = 0.25$, p >.05, partial $\eta^2 = .004$. Additionally, there were no significant differences between any of the goal conditions for both the dichotomous format, $F(2, 83) = 1.72$, p >.05, partial $\eta^2 = .040$, and for the Likert-type format, $F(2, 48) = 0.60$, p >.05, partial $\eta^2 = .024$. In addition to the overall test, several planned comparisons were performed to examine instruction group differences. Although results indicated no significant differences between goal conditions, an examination of the mean scores indicated that the responses in the administrative condition, ($M = 249.61$, $SD = 16.75$) were lower than both the feedback-related condition, ($M = 253.56$, $SD = 25.25$) and the control (department) condition, ($M = 252.15$, $SD = 26.10$). These results corroborate the results from the examination of the open-ended item. It seemed apparent that the administrative condition was too subtle. Because of this, the instructions were modified to help make the goal manipulation more salient (i.e., stronger).

An examination of the Goal Importance Questionnaire also revealed some interesting results. Specifically, although participants indicated that they were pursuing multiple goals during the evaluation, there were two that had the highest means for both formats and for each condition. Participants' two highest rated goals were: 1) Convey my satisfaction with the instructor's performance (item #4), and 2) Rate instructor fairly (item #2). Indeed, analyses of variance conducted using the ratings from these two goals as the dependent variables and the instructions as the independent variable indicated that there was no significant difference between any of the conditions for item #2, $F(2,133) = 0.56, p > .05, \eta^2 = .01$, nor for item #4, $F(2,133) = 0.35, p > .05, \eta^2 = .01$. This provided evidence that the respondents were, for the most part, pursuing the same goals, which support the findings of a non-significant overall ANOVA.

Tests for Sex Effects

Although, there has been prior research that has indicated no sex differences, this instrument had never been utilized before and therefore sex differences were examined. To determine if the sex of the participants was affecting the responses to the evaluation form, an independent samples t-test was performed. Although females tended to give higher overall ratings, ($M = 253.08$, $SD = 24.90$) than males, ($M = 249.97$, $SD = 21.76$), there was no significant difference between the two, $t(135) = -0.68$, p >.05. There were also no significant differences between males and females on either the instructor scale, $t(135) = -0.18$, p >.05, or the course scale, $t(135) = -1.23$, p >.05. Likewise, there were no significant sex differences for either format, or any of the instruction conditions. These results can be seen in Table 1.

Additionally, a DIF analysis was conducted to examine sex differences on individual items. For these analyses males were used as the focal group and females were used as the referent group. Overall there were only 5 items that showed significant DIF across all three conditions and both formats (using α = .20 to determine significance).

Table 1

Descriptives and Sex Differences Across Format Type and Instruction Condition for Overall Evaluation Ratings.

Format	Condition	Male			Female			t-test
		N	M	SD	N	M	SD	
	Administrative	4	241.50	4.43	15	241.40	6.29	0.03
Dichotomous	Feedback	10	236.70	8.88	23	238.91	7.53	-0.74
	Department	10	237.60	11.57	24	237.29	7.08	0.10
	Administrative	3	264.67	20.98	6	268.00	20.50	-0.23
Likert-type	Feedback	4	263.00	14.09	18	279.56	26.09	-1.21
	Department	8	274.25	26.51	12	279.25	28.11	-0.40

Upon an examination of these items, the researcher determined that these items did not contain sex-related biased wording (e.g., This instructor motivated me to do my best; This course was well organized), and the occurrence of 5 DIF items was within the probability of chance. Because of this, these items were not removed from the evaluation form.

Item-type Effects

As noted previously, research has indicated that instructor items tend to be rated higher than course items within evaluation forms. A related samples t-test indicated that instructor items were rated significantly higher than course items across both formats, *t*(137) = 20.36, p <.001, as well as within both the dichotomous format, *t*(85) = 22.25, p <.001, and the Likert-type format, *t*(50) = 14.48, p <.001.

Additionally, a DIF analysis was conducted to determine if the DPF procedure would be sensitive to the goal manipulation. Specifically, two DIF analyses were conducted on the instructor items and the course items. It was believed that the instructor items should show DIF (with the manipulation as the group variable); whereas the course items should not show DIF. Results indicated that no items showed significant DIF within either scale. These results are not surprising given that they are consistent with the previous results from the manipulation check. Those results indicated that the participants did not recognize the goal manipulation, specifically the administrative manipulation; therefore it is not surprising that there was no DIF within the instructor items. Also, this is consistent with the results from the Goal Importance Questionnaire, in which all participants indicated that they were pursuing the same goals.

Pilot Study Discussion

There were several goals of the Pilot Study. First the reliabilities of the two formats were examined to determine which should be used. Based on the internal consistency, Cronbach's alpha, the Likert-type format demonstrated better reliability overall, as well as for both the instructor and course scales. Because of this, it was decided that the Likert-type format would be used for the main study. Although the Mantel-Haenszel procedure requires dichotomous data, these Liker-type responses can (and have been) dichotomized. The methods and rationale used to dichotomize scores are discussed in the following sections dealing with the differential person functioning analysis (i.e., the main study).

A second purpose was to examine possible sex differences. Although previous research has indicated no sex differences during instructor/course evaluations, this is the first time this particular evaluation form had been used. The results from the univariate statistics as well as the DIF analyses indicated that there were no sex differences for the evaluation form overall as well as for any particular items. Although there were a few items that demonstrated significant DIF, there were only 5 which is within the margin of chance occurrence. They were kept in the evaluation form and all 83 items were utilized in the main study.

Third, a manipulation check was performed for the instruction conditions. It seemed apparent from an examination of the open-ended manipulation check item that the administrative condition was too subtle. Additionally, a lack of significant differences indicated that the participants did not provide substantially different ratings across the three instruction conditions. Similarly, the DIF analyses that were conducted on the instructor and course items showed similar results. Ideally, one would like to have seen no DIF items within the course scale and mostly DIF items in the instructor scale when using the manipulation as the grouping variable. However, this was not the case because there were no items that showed significant DIF.

As noted previously, the students were initially told that the purpose of the study was to refine a newly developed evaluation form. Although, participants were also presented with an instruction set afterwards, it was understandable that there were non-significant results. The one trend that was alarming was that the mean overall evaluation ratings were lower in the administrative condition than either of the other two. Based on the research cited, ratings in this condition should have been the highest. Again, this finding was corroborated in the administrative condition participants' inability to correctly identify the purpose of the ratings, as determined from the open-ended item. Although, it was believed that the main study would provide a stronger context for the rating process, the administrative instructions were re-written to help make this manipulation stronger.

Main Study

Methods

Participants

For the main study undergraduate students from a large section (380 students) of PSY101 during the Fall Quarter 2007 served as participants. The main study consisted of 280 undergraduate students from Ohio University, representing a 73.4% response rate to the evaluation request. There were 118 males (42.1%) and 162 females (57.9%). Students participated in this study under the belief that it was part of a newly developed evaluation program initiated by the psychology department. Although participants were given credit for their participation and received a debriefing form, it was essential that they believed these evaluations would be used for decision making.

Manipulations

The same measures and manipulations that were used in the pilot study were also used in the main study. Modifications were made and several measures were added as well. Those that remained unchanged from the pilot study are only listed, whereas any changes, modifications, or additions are discussed in detail.

Response Instructions. The response instructions given to the participants were manipulated as a way of manipulating rater goals (Wong & Kwong, 2007). Participants were presented with one of three different response instructions (i.e., goal manipulation) and were instructed to provide ratings according to those instructions. The instructions were presented after the purpose of the experiment was described. Specifically, participants received one of three instructions sets, or goal conditions. One goal condition indicated that ratings would be used for administrative purposes (i.e., pay, promotion, tenure). Based on results and feedback from the pilot study, the instructions for this condition were modified to help increase the strength of the manipulation. The specific instructions for the administrative condition were as follows (the bolded section was added to the instructions used in the pilot study):

> "You are being asked to fill out this form in order to provide the Psychology Department with a performance evaluation for this course and instructor. Your responses provide information that the department will use for tenure, promotion, and salary considerations for this instructor**. This is particularly critical at this time as your instructor, Dr. Popovich is being considered for promotion.** These evaluations are a crucial aspect when making this promotion decision. Careful responding is important to make the evaluation results informative and useful. Thank you!"

The second goal condition indicated that ratings would be used for feedback purposes for the instructor to change the course or themselves. The same instructions that were used in the pilot study were again utilized for this study (see pilot study manipulations). Finally, as in the pilot study, the third condition was meant to serve as a control condition and presented the standard instructions that the department of psychology uses for evaluations (see pilot study manipulations).

To test for order and fatigue effects, the evaluation forms were counterbalanced within each instruction condition such that there were four forms: 1) instructor items first then course items; 2) instructor items then course items, but all items were in reverse order; 3) course items first then instructor items; 4) course items first then instructor items, but all items were in reverse order.

Measures

Evaluation Form. Participants responded to an evaluation questionnaire that utilized a Likert-type response format ranging from 1 (Strongly Agree) to 4 (Strongly Disagree). The evaluation questionnaires consisted of 43 items related to the instructor (e.g., the instructor is knowledgeable in the field) as well as 40 items related to the course (e.g., the course was well organized). The scores on the evaluation form demonstrated satisfactory internal consistency (i.e., α) for this administration. Specifically, the reliability across the three experimental groups was .98.

For the DPF analyses, the Mantel-Haenszel method requires dichotomously scored data, therefore the responses were recoded into a binary format. To dichotomize the data, the response options indicating high levels of the trait were coded as "1" and the remaining options were coded as "0". Specifically, the mean item score, across all items, was used as a cut point (the distribution of item means was relatively normal – median = 3.21, therefore the mean was chosen as a cut point because of its stability). Those values that fell above 3.19 (4 and 5) were coded as 1, whereas those coded 3 and below were coded as 0. The dichotomized instrument also demonstrated satisfactory internal consistency, α = .98.

Evaluation Process and Format Reactions. This questionnaire was used to gauge raters' attitudes towards the questionnaire format and process that was presented. Although forced-choice formats have been shown empirically to reduce leniency, there has been noted dissatisfaction with the format from the rater's perspective, thereby leading to a lack of use. This questionnaire did not contain any forced choice items, but asked raters to provide their reactions to the format and process to determine if it was an acceptable alternative to the current methods and procedures. Six questions were designed to determine the degree to which rater's were satisfied with this evaluation process and format (Appendix C).

Goal Importance Questionnaire. This is a 19-item questionnaire that was developed by Murphy, Cleveland, Skattebo, and Kinney (2004) to assess the goals that raters pursue when evaluating instructors. Each item is rated on a 5 point Liker-type scale (1 – Strongly Agree to 5 – Strongly Disagree). This questionnaire was utilized to determine if raters were pursuing other goals in addition to the one assigned during the experiment (Appendix D).

Additional Items. Participants were also asked several additional questions. Specifically, one item asked them to indicate how their ratings would be used. This was an open-ended item and served as a manipulation check for the goal manipulation. The responses from this open-ended item were categorized by the researcher. Specifically, responses were examined and there were 5 main theme or "buckets" that were apparent and responses were classified into: 1) promotion (administrative); 2) feedback; 3) psychology department (i.e., both administrative and feedback); 4) general evaluation (e.g., course and/or instructor evaluation); and 5) other (e.g., I don't know, experiment, etc.). Participants were also asked whether or not they liked the instructor, their sex, as well as the grade they expected to earn in the course.

Psychology Department Rating Form. Because this session also served as the instructor's quarterly evaluation for the course, the department's standard evaluation form was included along with the open-ended response questions given to all students enrolled in courses within the psychology department (Appendix G). Additionally, the psychology department's standard open-ended items for feedback were also included (Appendix H). Responses from those participants who received the "control" condition were analyzed and provided to the department as the instructor's quarterly evaluation. Open-ended sheets were also given to the instructor for feedback purposes.

Procedure

Because the purpose of this study was to examine rater bias during a performance evaluation, steps were taken to deceive participants as to the true nature of the evaluation (i.e. a research study). The instructor of this PSY101 "mega section" (e.g., ~ 400 students) presented the following text to the student's of the course aloud and then posted it as an announcement on an online course website:

> "This year the psychology department is conducting teaching evaluations differently for the "mega section" of Psychology 101. This new method is considerably longer than the previous method and will take more class time; therefore you will receive 1 research credit for your participation if you choose to participate. Because of this, you may save one credit for this evaluation. Hence if you participate in this performance evaluation you will only have to obtain 3 additional credits to fulfill your research requirement. You may still obtain up to 6

> additional research credits by participating in studies or evaluating a research article (as described in the syllabus)."

Following this initial information at the beginning of the quarter regarding the process, the students were also sent a reminder via email and an announcement was posted on Blackboard, an online university course management system (Appendix I), during the 6th week of the quarter (2 weeks before the typical course evaluation). As the instructions indicate, the participants would be given an index card and asked to write their name and university email address. These would be collected when they turned in their evaluation form and used to give participants credit on the experimental website.

In accordance with IRB approval for both the pilot and the main study, to ensure a more realistic experience, no consent form was given, rather an explanation of the study, benefits, potential risks, and contact information was given to all participants in a debriefing form, which they received upon completion of the evaluation packet (Appendix J). Upon entering the class room, students received a lecture during the first 15 minutes of class. Because the pilot indicated that participants who took the Likert-type form spent an average of 19.89 minutes to complete the evaluation, I wanted to make sure they would have enough time to complete all components of the evaluation packet, which consisted of a cover sheet presenting them with one of the three instruction sets, the evaluation form and additional questions, the evaluation process and participant reactions questionnaire, the Goal Importance Questionnaire, and then the psychology department evaluation form and open-ended items. The order of the measures was consistent for all participants.

All students were instructed to wait to begin the evaluation until the experimenter had read all directions to the class. After all evaluation packets were passed out, the experimenter read the following script:

> "You are being asked to fill out this form in order to provide a performance evaluation for this instructor/course. Your responses will be analyzed and may be used for various personnel and administrative decisions. It is extremely important that you read your specific instructions and respond accordingly. As Dr. Popovich mentioned earlier in the year, due to the lengthy nature of this evaluation, you will receive 1 research credit if you choose to participate. Because of this, I am going to ask that you print your name on these index cards (which were passed out before the evaluations and instructions are given) and turn them in when you turn in your evaluation. Your name will not be linked to your responses in any way, they are merely going to be used to give you credit on the experimental website."

After this script was read, the students were instructed to begin the evaluation and bring their completed packets to the front when they were finished. Upon turning in a completed evaluation packet, each participant was given a debriefing form (Appendix J) to explain the study, the purpose of the deception, and information regarding their rights as a research participant.

Results

Test for Fatigue and Order Effects

The impact of the evaluation form order was examined to determine if there was any fatigue effects associated with the evaluation form. If there was a fatigue effect, counter-balancing would have distributed the effect equally across the conditions and an order effect would detect it. The evaluation form was counter-balanced such that there were four different "versions" (i.e., 1. instructor items first then course items; 2. instructor items reversed then course items reversed; 3. course items first then instructor items; 4. course items reversed then instructor items reversed). An analysis of variance was conducted for each of the response conditions. The results indicated that there was no significant difference between any of the four forms for the administrative condition, $F(3, 90) = 1.31, p > .05, \eta^2 = .04$, the feedback-related condition, $F(3, 91) = 0.37, p > .05, \eta^2 = .01$, and the control condition, $F(3, 87) = 0.38, p > .05, \eta^2 = .01$. Given the non-significant results, as well as the very small effect sizes, it appears there were no order/fatigue effects in the responses of the participants.

Manipulation Check/Goal Manipulation

Just as in the pilot study, the effectiveness of the response instruction manipulation was examined. First, the open-ended items were examined as a function of condition. Table 2 contains the results from a content analysis of the open-ended response items for each condition conducted by the researcher. The responses that participants listed were categorized into one of five buckets: 1) promotion (administrative); 2) feedback; 3) psychology department (i.e., both tenure and feedback); 4) general evaluation (e.g., course or instructor evaluation; and 5) other (e.g., I don't know, experiment, etc.). Within the administrative condition, of those who responded, 57.1% indicated that their ratings would be used for pay, promotion, and tenure decisions. No one in the feedback condition indicated that the ratings would be used for these purposes. Meanwhile, 59.7% of those who responded in the feedback-related condition effectively indicated that their responses would be used as constructive feedback. Only 8.5% of individuals in the administrative condition indicated the ratings might be used for feedback. Overall, these results confirm the administrative and feedback instructional manipulations were effective.

Table 2

Open-ended Manipulation Check Item Percentages

	Condition					
	Administrative (N=94)		Feedback (N=95)		Psych. Dept	
(N=91)						
Response Category	N	%	N	%	N	%
Promotion	40	57.1	0	0.0	11	15.9
Feedback	6	8.5	46	59.7	18	26.0
Department	0	0.0	0	0.0	6	8.7
Evaluation	14	20.0	21	27.3	30	43.5
Other	10	14.3	10	13.0	4	5.8

Within the department instruction set conditions, the findings were more ambiguous. Specifically, of the participants who responded to the open-ended item, 16%

indicated that their ratings would be used for administrative purposes (e.g., pay, promotion, tenure), 26% indicated their results would be used for some type of constructive feedback (e.g., for instructor to improve herself and/or course), only 9% indicated that their ratings would be used for both administrative and constructive purposes (i.e., the psychology department instructions), 30% indicated they would be used for an evaluation, and 6% indicated something other than the previous responses (e.g., I don't know). It seems apparent that the response pattern for the psychology department instruction condition was erratic and unpredictable. Indeed it was composed of a combination of participants pursuing administrative, feedback-related, and general evaluation goals.

Second, the mean level of responses in each condition was examined. If the manipulation led the participants to respond differently in each condition, there should be significant differences between the response instruction conditions on the total evaluation score. Also, this analysis served as a test of the goal-based perspective. Specifically, as Hypothesis 1 states, r*aters pursuing an administrative goal will have significantly higher ratings than raters who are pursuing a feedback-related goal or a control.* The means and standard deviations for each item within each condition are presented in Appendix K. The means and standard deviations for the total scores across conditions are presented in Table 3. To test for potential differences, an ANOVA was performed with the instruction manipulation as the independent variable and the evaluation score as the dependent variable.

Table 3

Means and Standard Deviations for the Overall Scores for Each Response Condition.

					Scale				
		Instructor			Course			Total	
Condition	N	M	SD	N	M	SD	N	M	SD
Administrative	94	141.99	17.99	94	124.29	17.07	94	266.28	33.89
Feedback	95	137.71	14.90	95	119.73	14.19	95	257.43	27.74
Department	91	143.52	15.73	91	124.34	15.73	91	267.86	29.64

Overall there was a significant ANOVA, $F(2, 277) = 3.18$, $p < .05$, $\eta^2 = .02$. To test Hypothesis 1, a series of planned comparisons were performed to examine group differences between the three conditions. Results indicated that ratings from the administrative instruction condition were significantly higher than ratings from the feedback-related instruction condition, $t(277) = 1.99$, $p < .05$, $\eta^2 = .020$; and that ratings from the psychology department instruction condition were significantly higher than the ratings from the feedback-related instruction condition, $t(277) = 2.33$, $p < .05$, $\eta^2 = .032$, but there were no significant differences between the administrative instruction condition and the psychology department instruction condition, $t(277) = 0.35$, $p > .05$, $\eta^2 = .001$. These results provide partial support for Hypothesis 1.

Comparison of Detection Methods

To test Hypothesis 2, *a differential person functioning analysis will be more effective at detecting lenient raters than traditional methods, (e.g. mean scores, rater* *skewness) in terms of sensitivity;* a series of analyses was conducted. Specifically, the percentage of raters each method classified as lenient, severe, and not biased was calculated. Next classification indices (e.g., the estimated probability of a consistent classification and Cohen's Kappa; see below for a detailed explanation) were calculated for each method and used in conjunction with an index of the sensitivity of each method for detecting bias. The results from all of these analyses provided a test for Hypothesis 2 and the resulting conclusion are presented at the end of this section.

Classification with the Differential Person Functioning Method. To determine which individuals were functioning differentially across the instructor and course items (i.e., rating in a differentially lenient or severe manner), differential person functioning (DPF) analyses were conducted. Specifically, the Mantel-Haenszel (MH) procedure was used to identify which raters were functioning differentially (i.e., differentially lenient or severe) across the experimental items. This analysis also provided information that was used to test Hypothesis 2. In particular, based on the number of biased (lenient or severe) raters, the relationship between the type of bias and the experimental condition provided an examination of the sensitivity of this method for detecting bias.

As noted previously, the MH procedure examines the relationship between two variables in a 2 x K frequency table (where K = the number of response options), controlling for the level of a third variable. The relationship is assessed through an odds ratio, and the odds ratio over all levels of the stratification (i.e., grouping) variable is a measure of the effect size. Specifically, the Educational Testing Service (ETS) has identified three categories of differential functioning that can be determined through a simple conversion (Camilli & Sheppad, 1994; Clauser & Mazor, 1998). The odds ratio for each 2 x K frequency table is used to calculate ETS's delta value using the following equation:

$$\Delta_{MH} = -2.35(\ln\alpha) \qquad (3)$$

where ln is the natural log and α is the odds ratio from the MH procedure. As indicated by Clauser and Mazor (1998), individuals who have a non-significant chi-square and an absolute value of Δ_{MH} that is less than one are considered "A" raters and are not considered a problem in that they demonstrate no differential responding. Individuals who have a significant chi-square and the absolute value of Δ_{MH} is greater than 1, but less than 1.5 are considered "B" raters because they show a small to moderate degree of differential responding. Finally, individuals who have significant chi-square and an absolute value of Δ_{MH} that is greater than 1.5 are considered "C" raters, because they demonstrate a large degree of differential responding.

The DPF analysis in this study utilized the traditional MH procedure (Mantel & Haenszel, 1959), which is approximately distributed as a chi-square with one degree of freedom. As such, the null hypothesis is that the common odds ratio is 1.0. Odds ratio values that are significantly greater than 1.0 indicate that the group coded as "1" has a higher odds of success than the group coded as "0", and values significantly less than 1.0 indicate that the group coded as "1" has a lower odds of success. Significant values are interpreted as evidence of DPF.

To perform the MH procedure with these data to detect DPF, the data had to be dichotomized and transposed. The scores from the Likert-type evaluation form were dichotomized by using the mean item score across all raters and items (*M* = 3.19). Those responses rated as 3.19 and higher were recoded as 1.0 and those responses rated 3 or below were recoded as 0. In this analysis, the item group (instructor vs. course) and the rater's responses (0 vs. 1) were crossed to form a 2 x 2 frequency table. According to Johanson and Alsmadi (2002), when the MH is used with transposed data (as it is in a DPF analysis), an overall mean item score over persons can be used to form the levels of the stratification variable. The stratification (i.e., grouping or blocking variable) is essential for assessing differential functioning. When dealing with an attitudinal scale (such as in this case), differential person functioning means that the level of agreement for a particular rater on one group of items is different than that for the other group of items; within subgroups of items that have similar overall scores across raters (Johanson & Osborn, 2004). To make comparisons within homogenous subgroups (of items in this case) on an overall attitude means using the mean or sum of item responses as a covariate, or blocking variable. When a stratification variable is developed in the described manner many of the traditional methods for detecting differential functioning can be used (e.g., Mantel-Haenszel; Johanson, & Alsmadi, 2002). As such, the total (i.e., summed) score across raters was used to develop the stratification variable in this study.

There is research that has attempted to determine the "best" number of stratification levels (e.g., Donoghue & Allen, 1993). Indeed, this research has even noted that there are numerous methods that one may choose from to pool score levels to assure that there is an adequate level of matching between the focal and referent groups within a given stratification level. Although "thin" matching (i.e., a high number of stratification levels) is the most desirable situation and provides the best ability to detect differential functioning, it may result in sparse cell frequencies. “Thicker” matching allows all, or at least, most of the data to be used. The premise is to use the number of stratification levels that allow matching of focal and referent items within each.

The stratification variable in each analysis in this study was the composite item score (i.e., sum across all raters). Levels of the stratification variable were collapsed (i.e., thickened) to produce sample sizes of at least 4 for each item type (i.e., instructor or course) within each stratification level, thereby resulting in six categories of total item scores that were used to conduct the DPF analysis.

The results of the differential person functioning analysis using the Mantel-Haenszel procedure are presented in Appendix L. As can be seen in Appendix L, 72 of the 280 participants (25.7%) were identified as differentially functioning at a significant level and had an effect size in the ETS B or C categories. Of these 72 differentially functioning raters, 50 of them were classified as lenient raters because they had a positive delta value (17.9%), where as the other 22 were classified as severe raters because they had negative delta values (20.9%). There were 16 lenient raters (17.0%) and 7 severe raters (7.4%) in the administrative instruction condition, 15 lenient raters (16.8%) and 9 severe raters (9.5%) in the feedback instruction condition, and 19 lenient raters (20.9%) and 8 severe raters (8.8%) in the psychology department instruction condition. A chi-square indicated that there was no significant relationship between the incidence of bias (i.e., no bias, leniency, severity) and the instruction conditions (e.g., administrative, feedback, psychology department), $\chi^2(4)=1.16$, $p > .05$, as well as between the administrative and feedback conditions only, $\chi^2(2)=0.277$, $p > .05$. These results suggest that the DPF method was insensitive to conditions. Specifically, the number and type of biased raters, was not different across the three conditions. Thus, this analysis indicated a lack of support for Hypothesis 2.

Classification with the Mean Score Method. In addition to determining the effectiveness of the DPF method at detecting bias, Hypothesis 2 required a comparison of the DPF method to the more traditional methods. In this section, the mean score method is examined. The next section will examine the skewness method. As described previously, the typical use for this method is for a given rater across ratees. And leniency (of a particular measure) is typically determined by examining the percentage of ratees who are rated over the scale midpoint (e.g., 60% of the ratees are rated higher than the scale midpoint). When individual raters are compared with this method, there are no specific criteria for determining leniency/severity (e.g., number of standard deviations above/below the mean to determine leniency/severity). Rather raters are compared relative to each other (e.g., Rater X is more lenient than Rater Y) or relative to the scale midpoint (e.g., Rater X has mean rating of 4.3 where as the scale midpoint is 3 on a 5 point scale, then Rater X is lenient). In this case a 4-point scale was used; therefore the scale midpoint is 2.5. Using the scale midpoint assumes that true performance is normally distributed around the scale midpoint (i.e., mean performance).

Unfortunately, Murphy and Cleveland (1995) pointed out that true performance is almost always unknown; therefore using the scale midpoint is often inadequate and leads to inaccurate results. Based on an analysis comparing each rater's average score on the instructor items to the scale midpoint, 269 of the 280 raters (96%) were classified as lenient raters. Alternatively, the mean (3.19) could be used. To justify this, the distribution of item means was examined and the results indicated that distribution was approximately normal with 3.19 as the mean (3.21 was the median). Because the ultimate goal is to use a value that represents the mean of true performance, the mean of the item means is more appropriate than the scale midpoint. Indeed, this value seems to offer a better representation of the mean score method. Appendix M presents the results of an analysis with the mean score method.

As can be seen in Appendix M, 199 of the 280 participants (71.1%) were identified as biased raters because there was a significant difference between their scores on the instructor items and the grand mean. Of these 199 biased raters, 122 of them (61.3%) were classified as lenient raters because they had significantly higher ratings than the grand mean, whereas the other 80 (40.2%) were classified as severe raters because they had significantly lower instructor ratings than the grand mean. There were 44 lenient raters (44.8%) and 27 severe raters (28.7%) in the administrative instruction condition, 31 lenient raters (32.6%) and 31 severe rater (32.6%) in the feedback instruction condition, and 47 lenient raters (51.6%) and 21 severe raters (23.1%) in the psychology department instruction condition. A chi-square indicated that there was no significant relationship between the incidence of bias (i.e., no bias, leniency, severity) and the instruction condition (e.g., administrative, feedback, psychology department), $\chi^2(4)=7.89$, $p > .05$, as well as between the administrative and feedback conditions only,

$\chi^2(2)=4.30$, $p > .05$. Just as was the case with the DPF method, the traditional mean score method was also insensitive to condition.

Classification with Skewness Ratings. Another approach to detecting leniency/severity is by assessing the degree of skewness in ratings (Landy, Farr, Saal, & Freytag, 1976). Skewness scores were calculated for each rater and the results can be seen in Appendix N. To determine the significance of the skewness statistics, the scores were transformed into z-scores (Tabachnik & Fidell, 1996). The significance test for skewness is tested against the null hypothesis of zero. Specifically, the standard error of the skewness statistics is approximately:

$$s_s = \sqrt{\frac{6}{N}} \tag{4}$$

where N is the number of cases (83 items in this instance). The obtained skewness value is compared with zero using the z distribution, where:

$$z = \frac{S - 0}{s_s} \tag{5}$$

and S is the obtained skewness statistic. According to Tibachnik and Fidell (1996), a conservative value for alpha should be used (i.e., .01), therefore the critical value for determining significance was z = ± 2.58, which translated into an obtained skewness statistic of ± .694.

As can be seen in Appendix N, 121 of the 280 participants (43.2%) were identified as biased raters because they displayed significant skewness scores. Of these 121 biased raters, 114 of them (40.7%) were classified as lenient raters because they demonstrated significantly negatively skewed ratings, whereas the other 7 (2.5%) were classified as severe raters because they had significantly positively skewed ratings. There were 39 lenient raters (41.5%) and 3 severe raters (3.2%) in the administrative instruction condition, 34 lenient raters (35.8%) and 3 severe rater (3.2%) in the feedback instruction condition, and 41 lenient raters (45.1%) and 1 severe raters (1.1%) in the psychology department instruction condition. A chi-square indicated that there was no significant relationship between the type of bias (i.e., no bias, leniency, severity) and the instruction conditions for all three conditions (e.g., administrative, feedback, psychology department), $\chi^2(4)=4.28$, $p > .05$, as well as between the administrative and feedback conditions only, $\chi^2(2)=0.664$, $p > .05$. Thus, all three methods were insensitive to condition thereby indicating a lack of support for Hypothesis 2.

Classification Consistency between Methods

Although none of the methods showed sensitivity to the manipulated goals, a supplemental analysis that can provide insights deals with the degree of agreement between the various methods, also called decision, or classification, consistency. Decision consistency deals with the extent to which the same decisions can be made from two different sets of measurements (Crocker & Algina, 1986). The agreement among the three techniques, in regards to their decisions about the presences or absence of bias was examined. If the three methods resulted in the same overall decisions, then the consistency would be high, whereas low agreement would result in low decision consistency.

To test for consistency, two Classical Test Theory (CTT) indices were used. Each index is based on a theoretical decision framework (see Figure 3). The first CTT index that was used was the estimated probability of a consistent classification (Crocker & Algina, 1986). Specifically, this index compares the decisions made (e.g., bias vs. not bias) for two different methods (e.g., DPF vs. mean scores). The results provide evidence about the likelihood for arriving at the same decision using two different methods. High probabilities indicate a high likelihood that each method will arrive at the same decision, whereas low probabilities indicate less likelihood of arriving at the same decision.

This index is the sum of the probabilities for a "biased" decision on each measure and a "not biased" decision on each measure. Mathematically, this is represented with the following equation:

$$\hat{P} = \hat{P}_{11} + \hat{P}_{00} \tag{6}$$

where $\hat{P}_{11}$ is the estimated probability of a "biased" decision on each measure and $\hat{P}_{00}$ is the estimated probability of a "not biased" decision on each measure. Values close to 1.0 indicate a high degree of consistency, whereas values close to 0.0 indicate a low degree of consistency.

In addition to examining the estimated probability of a consistent decision, Swaminathan, Hambleton, and Algina (1974) also suggest the use of Cohen's Kappa. Although similar to $\hat{P}$, it takes consistency due to chance into account. Specifically, Cohen's Kappa compares the decisions made (e.g., bias vs. not bias) by two different methods (e.g., DPF vs. skewness) after removing the consistency attributable to chance alone. As Crocker and Algina (1986) noted, if the probability is high, it is very likely that the same decision would be reached using either method, even after adjusting for chance, whereas if the probability is low, it indicates that the consistency in the decisions may be due to chance. Mathematically, Cohen's Kappa (i.e., Kappa) is represented with the following equation:

$$\kappa = \frac{P - P_c}{1 - P_c} \tag{7}$$

where P_c is the chance probability of a consistent decision (i.e., chance consistency), and is calculated by using the formula:

$$P_c = P_{1.}P_{.1} + P_{0.}P_{.0} \tag{8}$$

where $P_{1.}$, $P_{.1}$, $P_{0.}$, and $P_{.0}$ represent the column and row totals from the classification table in Figure 3.

All possible comparisons between the three bias detection methods were made and $\hat{P}$ and κ were calculated for each. These analyses were performed using 1) all of the participants across condition and regardless of the type of bias (i.e., leniency vs. severity), 2) within each of the three response instruction conditions regardless of the type of bias, 3) across all participants for each type of bias (i.e., leniency or severity), and 4) within each response condition for each type of bias.

The results of the classification consistency analyses for overall bias (i.e. leniency and severity) across all of the response conditions, as well as within each response condition are presented in Table 4. As can be seen in Table 4 below, the highest levels of consistency in terms of overall bias (both leniency and severity simultaneously) were found between the mean score method and the skewness method across all conditions, as well as within each response condition. Additionally, the consistency in a bias decision above chance was moderate between the mean score method and the skewness method, ranging from 13.2% (Administrative) to 32.2% (Psychology Department) consistency above chance.

Table 4

Results of the Classification Consistency Analysis for the DPF, Mean Score, and Skewness Methods for Detecting Rater Bias (N=280)

Condition	Method		1	2	3
	1. DPF	*P*	----		
		κ	----		
Across All Conditions	2. Mean Score	*P*	0.418	----	
		κ	0.034	----	
	3. Skewness	*P*	0.511	0.600	----
		κ	-0.045	0.245	----
	1. DPF	*P*	----		
		κ	----		
Administrative	2. Mean Score	*P*	0.404	----	
		κ	0.055	----	
	3. Skewness	*P*	0.479	0.543	----
		κ	-0.102	0.132	----
	1. DPF	*P*	----		
		κ	----		
Feedback	2. Mean Score	*P*	0.432	----	
		κ	0.012	----	
	3. Skewness	*P*	0.526	0.610	----
		κ	-0.064	0.270	----
	1. DPF	*P*	----		
		κ	----		
Psych. Department	2. Mean Score	*P*	0.418	----	
		κ	0.030	----	
	3. Skewness	*P*	0.528	0.648	----
		κ	0.024	0.322	----

When comparing the DPF method with the mean score method in terms of overall bias, there was moderate classification consistency (ranging from .432 to .404), although the consistency above chance was low, ranging from 1.2% (Feedback) to 5.5% (Administrative).

Finally, although the classification consistency was also moderate between the DPF method and the skewness method, the consistency above chance was only positive in one case (2.4% - Psychology Department), whereas it was negative in all other cases. This indicates that most of the observed agreement is likely due to chance alone.

When comparing the DPF method with the mean score method, as well as when comparing the skewness ratings with the mean score method, the moderate consistency ratings were achieved because of equally similar classifications for those who were biased versus not biased. When comparing the skewness method with the mean score method across all conditions, the consistency was due to high agreement in identifying individuals who were biasing their responses. Conversely, when comparing the DPF method with the skewness method, the consistency ratings were achieved because of similar classifications for individuals who were not biasing their responses.

The results of the classification consistency analyses for each type of bias (i.e., leniency or severity) across all of the response conditions, as well as within each response condition are presented in Table 5. As can be seen, the results indicate that the highest levels of consistency and agreement were found between the mean score method and the skewness method, as well as between the DPF method and the mean score method across all conditions, as well as within each response condition.

Table 5

Results of the Classification Consistency Analysis for the DPF, Mean Score, and Skewness Methods for each Type of Bias (i.e., Leniency and Severity). (N=280)

Condition	Method		1	2	3
	1. DPF	*P*	----	0.675	0.889
		κ	----	-0.017	-0.040
Across All Conditions	2. Mean Score	*P*	0.650	----	0.707
		κ	0.237	----	0.001
	3. Skewness	*P*	0.543	0.686	----
		κ	-0.038	0.356	----
	1. DPF	*P*	----	0.702	0.894
		κ	----	0.066	-0.047
Administrative	2. Mean Score	*P*	0.650	----	0.702
		κ	0.298	----	0.010
	3. Skewness	*P*	0.543	0.691	----
		κ	-0.031	0.376	----
	1. DPF	*P*	----	0.621	0.874
		κ	----	-0.055	-0.050
Feedback	2. Mean Score	*P*	0.726	----	0.663
		κ	0.282	----	0.001
	3. Skewness	*P*	0.568	0.684	----
		κ	-0.072	0.229	----
	1. DPF	*P*	----	0.703	0.901
		κ	----	-0.067	-0.020
Psych. Department	2. Mean Score	*P*	0.560	----	0.758
		κ	0.137	----	-0.021
	3. Skewness	*P*	0.517	0.714	----
		κ	-0.026	0.430	----

Note: Consistency ratings for Leniency are on the bottom of the diagonal and consistency ratings for Severity are on the top of the diagonal.

In regards to leniency (under the diagonal), the consistency above chance was moderate to considerable between the DPF method and the mean score method, ranging from 13.7% (Feedback) to 29.8% (Psychology Department), as well as between the skewness method and the mean score method, ranging from 22.9% (Feedback) to 43.0% (Psychology Department). However, this was not the case between the DPF method and the skewness method. When comparing these two methods, Kappa was negative for all comparisons, thereby indicating that most of the observed agreement was likely due to chance alone. In all cases, the consistency ratings were achieved because of similar classifications for individuals who were not biasing their responses. This is most evident in the agreement ratings for severity in all cases. Although the P values indicate that there were moderately high levels of agreement between the methods (e.g., .5-.7), the agreement over and above chance was either very low (e.g., .001) or negative. As mentioned previously, this indicates that most of the observed agreement is likely due to chance alone.

The classification consistency between the different bias detection techniques was modest at best. Overall, there were 26 individuals who were identified as “biased” raters across all conditions with all three methods. Of those 26, 17 individuals were identified as being lenient raters across all three methods, however there were zero individuals identified as being severe raters with all three methods. The other 9 individuals were identified as severe by the DPF method and lenient by both the mean score method and the skewness method. Of the 50 raters who were identified as lenient raters with the DPF method, 37 of them were also identified as lenient raters with the mean score method;

however, the mean score method identified 122 raters overall as being lenient. As discussed previously, this high number of “lenient” raters (44% from the total sample) may be the result of measuring “impact” (Dorans, 1989) rather than biased responding. Another possible explanation is that, although the Mantel-Haenszel procedure can be conducted with relatively small samples (e.g, as low as 50; Fidalgo, Ferreres, & Muniz, 2004), it has been noted that only those individuals with more extreme differential functioning (i.e., DPF) will consistently be detected as statistically significant (Dorans & Holland, 1993). Because of this, it seems that those raters identified as biased by the DPF method had a higher degree of bias (either leniency or severity), than those identified by the mean score method.

In the Administrative condition, 85.1% of the raters were identified as being biased raters (from the total sample, 75.0% were lenient and 40.0% were severe) by at least one of the three methods. In the Feedback condition, 78.9% of the raters were identified as being biased raters (from the total sample, 72.0% were lenient and 53.3% were severe) by at least one of the three methods. Finally, in the Psychology Department condition, 84.6% of the raters were identified as being biased raters (from the total sample, 81.8% were lenient and 37.7% were severe) by at least one of the three methods.

Summary. Chi-square analyses indicated that none of the bias detection methods showed greater sensitivity to the manipulated goals (i.e., experimental conditions). Because this was true for the DPF method, the findings indicated a lack of support for Hypothesis 2. Additionally the classification consistency was examined between the three detection methods. The results indicated that the highest consistency was between the mean score method and the skewness method. Although consistency was relatively high with the skewness method, the negative Kappa indicates that the results were likely due to chance alone.

Scale Differences

Although the mean score method has been used to detect biased raters, another use of this method is to compare the rater means for the different conditions (e.g., Bernardin, Alvarez, & Cranny, 1976), and determine which format (e.g., course scale vs. instructor scale) produces more leniency. To test Hypothesis 3; *instructor ratings will be significantly higher than course ratings across all conditions,* the two scales that encompassed the evaluation form were examined. An independent samples t-test indicated that instructor ratings were rated significantly higher than course ratings, $t(81) = 3.46$, $p < .001$. This result provides support for Hypothesis 3

Proportion of Biased Raters

Another test of the goal-based approach that was intended to serve as a validation for the DPF method's effectiveness was to examine the proportion of biased raters within each condition. Even though there were mean differences (i.e., Hypothesis 1), it was believed that the proportion of differentially lenient raters would be higher for raters in the administrative condition than those in the other conditions (i.e., Hypothesis 4). Previously, the relationship between the type of bias and the experimental condition was examined (i.e. sensitivity as determined by chi-square tests). Those analyses indicated that all three detection methods were insensitive to the experimental conditions, thereby providing an omnibus test for this hypothesis. Hypotheses 4 can be considered more a "planned comparison" analysis in regards to this relationship. Although there were no significant differences in terms of the omnibus evaluation, just as with an analysis of variance, that should not preclude us from examining the "planned comparisons." Because of this, to test Hypothesis 4 a series of chi-square statistics were computed.

Results indicated that for the DPF method there was not a significant difference in the proportion of either lenient raters, $\chi^2(1)=0.052$, $p > .05$, or severe raters, $\chi^2(1)=0.250$, $p > .05$, between the administrative and feedback conditions. There was also no significant difference between the administrative and psychology department conditions for either lenient raters, $\chi^2(1)=0.449$, $p > .05$, or severe raters, $\chi^2(1)=0.112$, $p > .05$; nor were there differences between the feedback and psychology department conditions for either lenient raters, $\chi^2(1)=0.806$, $p > .05$, or severe raters, $\chi^2(1)=0.026$, $p > .05$. Although there were mean differences between the groups, there was no difference in the proportion of biased raters between any of the conditions. These results indicated a lack of support for Hypothesis 4.

The same analyses were also conducted for the other traditional methods to compare with the results of the DPF method. Results indicated that for the mean score method there was a significant difference in the proportion of lenient raters, $\chi^2(1)=3.967$, $p < .05$, however there was no significant difference in the proportion of severe raters, $\chi^2(1)=0.339$, $p > .05$, between the administrative and feedback conditions. There was no difference between the administrative and psychology department conditions for both lenient raters, $\chi^2(1)=0.433$, $p > .05$, and severe raters, $\chi^2(1)=0.767$, $p > .05$. When comparing the feedback and psychology department conditions, there was a significant difference in the proportion of lenient raters, $\chi^2(1)=6.903$, $p < .05$, however there was not a significant difference for severe raters, $\chi^2(1)=2.107$, $p > .05$.

Finally, results indicated that for the skewness score method there was no difference in the proportion of either lenient raters, $\chi^2(1)=0.648$, $p > .05$, or severe raters, $\chi^2(1)=0.000$, $p > .05$, between the administrative and feedback conditions. There was also no difference between the administrative and psychology department conditions for either lenient raters, $\chi^2(1)=0.240$, $p > .05$, or severe raters, $\chi^2(1)=0.957$, $p > .05$; nor was there a difference between the feedback and psychology department conditions for either lenient raters, $\chi^2(1)=1.658$, $p > .05$, or severe raters, $\chi^2(1)=0.936$, $p > .05$.

Although Hypothesis 4 was not supported in terms of the DPF method, the results indicated significant differences in the proportion of lenient raters between the administrative and feedback conditions for the mean score method. As proposed previously, this may be due to the sample size (i.e., number of items) used for the DPF method. Although the number of items used in this study (along with the adjustment in alpha level) was adequate to provided sufficient power; when smaller sample sizes are used, only those raters with the most extreme bias are detected. This may account for the lack of support for the DPF method.

Evaluation Reactions

As noted previously, participants responded to several items regarding their reactions to the evaluation format that was utilized. There were no significant differences between any of the three instruction conditions for any of the items, although there were several items for which there were sex differences. Descriptives are presented in Table 6.

Table 6

Descriptive Statistics for Evaluation Reaction Items

	Total(N=280)		Males(N=118)		Females(N=162)		
Item	M	SD	M	SD	M	SD	t-test
Instructor Items							
1. The evaluation is a fair way of evaluating an instructor's level of performance.	3.01	0.62	2.94	0.64	3.06	0.60	1.53
2. The psychology department should adopt this evaluation form for future instructor evaluations.	2.96	0.64	2.87	0.69	3.02	0.59	1.90
3. You can control the ratings you give (you can give high or low ratings when they are appropriate).	3.18	0.59	3.08	0.53	3.24	0.63	2.19*
4. This type of evaluation gave you the ability to give objective ratings of your instructor.	3.09	0.65	3.03	0.59	3.13	0.69	1.32
5. Please rate the level of difficulty in using this method.	3.35	0.62	3.22	0.63	3.44	0.59	3.05**
6. Please rate your level of satisfaction in using this evaluation method.	2.97	0.56	2.92	0.46	3.01	0.62	1.43

*$p < .05$; **$p < .01$

In highlighting these results, 84.7% of the respondents felt that this evaluation method was a fair method for evaluating an instructors performance, 79.6% felt the psychology department should adopt this evaluation form for future ratings, 90.1% felt that they were in control of their ratings, 87.1% felt that this format gave them the ability to give objective ratings, 94.6% rated this method as easy or extremely easy, and 85.8% were satisfied with the method (12.9% were extremely satisfied).

In terms of sex differences, as Table 6 indicates, females indicated that they felt significantly more in control of their ratings and felt the degree of difficulty was significantly easier than did males. Overall, participants gave favorable ratings to this evaluation method. These results help to support the use of this method over other methods that may eliminate bias, but are seldom used because of various constraints (i.e., forced-choice formats).

Goal Questionnaire Relationships

In addition to the goal manipulation, participants were also given a goal questionnaire to determine if there were additional goals that they were pursuing during the rating process. Descriptives are presented in Table 7. Appendix O presents the correlational results for participant's total score with each of the goal questionnaire items. As can be seen in Appendix O, there were differences in the self-reported goals within

Table 7 (continued on page 115)

Descriptive Statistics for the Goal Importance Questionnaire

Item	N	M	SD
1. Identify areas in which the instructor might need improvement.	279	3.68	0.84
2. Rate my instructor fairly.	279	4.16	0.67
3. Identify areas where the instructor needs more training.	279	3.37	0.93
4. Convey my satisfaction with the instructor's performance.	278	4.09	0.66
5. Identify area that the instructor should focus on improving.	279	3.56	0.93
6. Indicate where the instructor fell short in terms of performance.	279	3.49	0.96

7. Give my instructor a rating that she or he will realize is based on performance, rather than my judgment of him/her as a person.	279	4.00	0.86
8. Identify my instructor’s strengths and weaknesses.	279	3.95	0.76
9. Highlight my instructor’s performance so that his or her success is visible to his or her department head.	279	3.90	0.82
10. Improve my instructor’s confidence.	279	3.34	1.03
11. Make it clear to my instructor that there is room for improvement.	279	3.39	0.92
12. Identify my instructor’s performance deficiencies.	279	3.34	0.89
13. Challenge my instructor to improve his or her performance.	279	3.44	0.90
14. Clarify expected performance levels to the instructor.	279	3.53	0.84
15. Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.	279	3.81	0.79
16. Indicate where instructor has exceeded performance expectations.	278	3.83	0.82
17. Encourage the instructor’s current level of performance.	279	3.91	0.81
18. Encourage the instructor to improve performance.	279	3.41	0.94
19. Motivate the instructor.	279	3.51	0.94

each of the conditions in addition to several interesting findings. Within the administrative condition, there were 11 different goals that were reported as influencing a rater that ranged in their focus (e.g., Rate instructor fairly, Encourage current performance, Identify strengths and weaknesses, Improve confidence). In the feedback condition there were only two items that were significantly related to overall evaluation score (i.e., Indicate where instructor has exceeded performance expectation and encourage the instructor to improve). Other goals that would be expected to be related within this condition were not, and were even in the opposite condition in some instances (e.g., Identify strengths and weaknesses, Make it clear that there is room for improvement, Identify areas that instructor should focus on improving). Finally, there were nine goals within the psychology department that were significantly related to overall score. Interestingly, seven of those nine were the same goals identified within the administrative condition.

As the analysis of variance for Hypothesis 1 indicated, there were significant differences between all of the conditions except for the Administrative condition and the Psychology Department condition. Although the Psychology Department condition was intended to serve as a control condition, it appears that respondents were pursuing very similar goals and therefore it is not unreasonable that their ratings (and even incidence of errors) were similar. To that end, multiple ANOVAs were conducted to examine whether self-reported goals were a function of condition. Specifically an analysis of variance was conducted for each item in the Goal Importance Questionnaire by instruction condition. These results are summarized in Appendix P.

As can be seen in Appendix P, there were only two items for which there was a significant difference between at least two of the instruction conditions (item 9: Highlight my instructor's performance so that his or her success is visible to his or her department head, and item 15: Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.) Specifically, for item 9, responses in the Psychology Department condition were significantly higher than those in the Feedback condition. Similarly for item 15, the responses in the Psychology Department condition were significantly higher than those in both the Administrative condition and the Feedback condition. More meaningful from these results is the extent to which there are not significant differences for all other comparisons. This again suggests that it is not unreasonable to expect a lot of individual differences regarding the goals that raters pursue, despite the instructions given to them.

Additionally, as can be seen in Appendix Q, the relationship between various goal-related items and the incidence of leniency both within condition and across conditions are less than impressive. Although there are several goals that were related to the incidence of leniency in the Administrative condition, there were none that were consistent across all three detection methods, and the types of goals that were related are sporadic (e.g., Rate instructor fairly, Highlight performance for department head, Identify strengths and weaknesses). The results of these correlational analyses do little to support the empirical results discussed previously, particularly in relation to the goal-based approach. They do however, provide further confirmation that the raters in this study were pursuing similar goals regardless of the instructions they were given (i.e., condition), and it seems apparent from the correlations in Appendix Q (specifically the correlations across conditions), that raters were pursuing multiple goals during the ratings process.

Discussion

Two studies (i.e., a pilot and the main study) were conducted to determine if differential person functioning (DPF) could be used to detect rater bias within performance evaluations. The purpose of the pilot study was to examine the psychometric properties of two different evaluation formats (i.e., dichotomous vs. Likert-type) and determine which was more appropriate, as well as to test the strength of the experimental manipulation. The results of the pilot study indicated that a Likert-type response format was more reliable than a dichotomous response format and that the wording of the experimental conditions needed to be edited to provide more context (i.e., more strength). Results found no sex differences, consistent with previous research. In particular, the items within the evaluation form did not demonstrate differential item functioning (DIF), therefore all items were retained and used in the main study. These pilot results were used to strengthen the main study.

There were two key purposes for the main study. The primary aim was to determine if the differential person functioning (DPF) technique could be used to identify individuals who were giving biased (i.e., lenient/severe) ratings. This method was also compared to traditional techniques that have been used to detect leniency/severity within the performance appraisal literature (i.e., mean scores and skewness ratings). The other aim of the main study was to examine a goal-based perspective as to why individuals give different ratings (Cleveland & Murphy, 1992). In addition to providing a direct test of the perspective, this aspect also served as a method to validate the DPF technique as a viable means for detecting lenient raters. Although the DPF technique could be used to identify raters as differentially lenient or differentially severe, the effectiveness of the method could not be determined without a successful goal manipulation. Specifically, to test the DPF method’s effectiveness, the goal-based approach needed to be supported first. Because of this, a discussion interpreting the primary results is presented first, and then a discussion of some secondary issues (e.g., classification consistency, goal questionnaire findings) afterwards. Finally, a discussion of some practical implications is presented.

Primary Findings

As noted, the goal-based perspective regarding performance ratings was examined to provide empirical support for the approach as well as to validate the DPF technique. The results indicated several interesting findings. First, consistent with other studies (e.g., Wong & Kwong, 2007), the results provided support for the hypothesis that raters’ goals (manipulated by instructions given) influenced ratings. Indeed, those given “Administrative” instructions (i.e., basis of ratings was pay, promotion, tenure) provided higher ratings than those given “Feedback” instructions (i.e., basis of ratings was constructive feedback for improvement), and those pursing the standard instructions (i.e., both administrative and feedback aspects) provided ratings higher than both the Administrative and Feedback groups. These differences in mean ratings provided further support for the goal-based perspective.

In regards to examining the effectiveness of the DPF technique, the results did not provide support for the hypothesis that a differential person functioning analysis will be more effective at detecting lenient raters than traditional methods (e.g. mean scores, rater skewness). In particular, the results indicated that the DPF technique was insensitive to the goal manipulations. There were differences in average ratings that clearly indicated support for the goal-based approach, yet results also indicated that the proportion of lenient raters (as identified by the DPF method) was not different between the conditions. With differences in average ratings between the goal conditions, the incidence of leniency should have been different between conditions, and yet a chi-square analysis indicated that there was no difference. Because the goal manipulation served as a validation of the DPF method, it seems evident that the technique was ineffective at detecting bias. The question then is whether the DPF method was "given a fair chance".

To help determine if the DPF method received a fair shot at detecting leniency, it was compared to the more traditional leniency detection techniques. Specifically, the DPF method was compared to the mean score method and skewness ratings. The results indicated that both the mean score method as well as skewness ratings were also insensitive to detecting bias between conditions. So, although the DPF method was ineffective at detecting bias, it seems as though the traditional methods were just as ineffective. Initially, these results seem to suggest that all of the methods examined here should be used with caution when attempting to detect leniency/severity in performance evaluations; however there were several issues that may have "hindered" the DPF technique's ability to detect bias.

One issue to consider regards the nature of the statistical technique used in the DPF method. Specifically, the DPF method relied on the Mantel-Haenszel procedure, which compares the probability of responding to one set of items to the probability of responding to another set of items (i.e. instructor items vs. course items). These item types were used because research has found that raters tended to provide inflated ratings for instructor items and lower ratings for course items (e.g., Aleamoni & Gary, 1980; Aleamoni & Hexner, 1980; Kidd & Latif, 2004; Phipps, Kidd, & Latif, 2006), and indeed results from this research supported those findings (i.e., Hypothesis 3). Additionally, it was believed that these item-types would be susceptible to the goal manipulation (i.e., Administrative vs. Feedback). In particular, it was postulated that instructor items would be rated more leniently and that those with the Administrative instructions would be more likely to demonstrate lenient ratings. Essentially, the course items were used as the "base rate" items for determining DPF. It may be the case, however, that this comparison was not the most appropriate and that other types of items may provide a better test for using the DPF technique to detect leniency/severity. For example, Scherbaum (2003) used item response theory to determine which items in an inventory were "fakeable" and which ones were resistant to response distortion and then used the DPF technique to detect faking on a personality inventory. Another possible avenue for future research could be to develop items that are either summative or formative in nature. Summative items tend to measure performance on a more macro scale and may yield more inflated ratings; whereas formative items are more micro in scale and may yield lower ratings. The DPF technique could be used to determine bias in a similar method as was used in this research.

A second issue to consider is that of statistical power for the DPF method. As mentioned previously, for a DPF analysis, the sample size is the number of items. In the analyses that were performed in this study, the number of items was relatively small (N=83). Although research supported the use of sample sizes as low as 50 (Fidalgo, Ferreres, & Muniz, 2004), these are still not ideal. Additionally, the items were split into the two groups (instructor and course) with 43 and 40 items respectively. Even though having equal numbers of items within each group is ideal, this low number still yields much lower power than traditional differential functioning analyses that use sample sizes of 1,000 or more.

Having said that, it is important to consider the ecological validity of this evaluation. Specifically, there is likely a practical limit to the length of a performance evaluation form. Having a performance evaluation instrument that contains 1,000 items or more would be daunting to complete. In fact, it was the case that the 83 item instrument used in this research was a considerable expansion of typical teaching evaluation forms (most instruments are 20 items or less). Even though participant reactions regarding this evaluation method were favorable, it would be hard to believe that raters (e.g., students) would approve of an evaluation instrument containing upwards of 1,000 items.

A third issue to consider deals with the stratification of the grouping variable. The strata were created such that there were at least 4 items from each group in each stratum, thereby resulting in 6 levels. Again, a higher value would have been more desirable, as previous research has indicated that "thinner matching" (i.e., more levels) allows for better detection of differential functioning (e.g., Donoghue & Allen, 1993). Although several methods were initially examined to create strata (e.g., total score; percent of total sample – deciles, sextiles, quintiles; equal intervals, etc.), in each case the resulting strata did not have representation from both the referent and focal groups. Although a chi-square can be calculated, the results are not accurate. Because of this, the cutpoints were created to ensure adequate representation (4 items) from both the referent and focal groups in each stratum. Regardless, it is likely that the relatively small number of items (e.g., 83) my have constrained the possibility of identifying individuals as giving either differentially lenient or differentially severe ratings. Because of this, an individual had to demonstrate considerable differences in their responding on the few items within each stratum.

A final issue that could have impacted the "fairness" of the test of the DPF method for detecting bias is related to the goal manipulation. It may have been the case that the reason the DPF method was unsuccessful was because of the wording in the instructions. Specifically, the instructions that were presented in each condition indicated that the ratings would be used by the Psychology Department as a performance evaluation for the "course and instructor". If the instructions, particularly in the Administrative condition, indicated that the instructor items would be used to evaluate the instructor and the course items would be used to evaluate the course independent of the instructor, then the effect may have been larger. It is likely that the instructor items were not rated differently than the course items within a given condition. Because differential functioning can only be present when the different sets of items are rated differently, it is possible the instructions hindered the DPF method's ability to detect bias.

Another way the manipulation could be made stronger is by using different classrooms (each with the same course and instructor) where each received a different set of instructions. In this study, a mega section of psychology 101 was used and the instructions that were read were generic and simply directed students to their respective instruction set. A situation in which an administrator is able to read the instructions aloud may make those "goals" more salient, and thereby produce stronger results.

With that in mind, it is important to realize that although using two different classes may result in a stronger manipulation, different classrooms could also introduce different confounds. It is likely that different classes develop different "personalities" (e.g., cultures) that could affect the ratings. For example, ratings could be based on how students observe how other students are treated. Different classes have different people who may warrant different treatment. Similarly, changes in the presentation method of the manipulation may have resulted in a stronger manipulation. Future research could compare the incidence of rater bias when instructions were read by the participant, read to by the administrator, presented via video, etc. Research in the area of presentation modality could help guide such studies.

It seems apparent that there were several factors that may have influenced the DPF method's ability to effectively detect biased raters. Although these issues reveal possible avenues for future research, there are also practical considerations that need to be taken into account for similar research endeavors. In addition to examining the effectiveness of DPF, as well as the traditional approaches, this research also allowed for an examination of other aspects that have implications for performance appraisals in both research and practice. Specifically, other findings have implications for bias detection methods, as well as the goal-based approach. These findings are somewhat separate from the primary findings and will be discussed individually below.

Additional DPF Findings

As the previous discussion indicated, the DPF method was ineffective at detecting bias. However, the traditional approaches for detecting leniency/severity in ratings were also examined and the results indicated that they too were ineffective at detecting bias. In addition to examining the incidence of leniency in each condition, the decision (i.e., classification) consistency was also calculated to examine the extent to which each of these methods identified the same raters as "biased". The classification indices indicated that the highest level of agreement was between the mean score method and the skewness scores. There was also a relatively high level of agreement between the DPF method and the mean score method for detecting leniency, with nearly 24% of that agreement above and beyond chance. Although there was relatively high agreement between the skewness method and the DPF method, the negative Kappa values indicate that the results were likely due to chance alone. This finding is particularly interesting and deserves some further discussion.

As can be seen in Tables 4 and 5, the skewness ratings had relatively high consistency with both the mean score method and the DPF method, however; the negative Kappa values indicate that the consistency with the DPF method is likely due to chance alone. Interestingly though, is that there was relatively high consistency between the mean score method and the skewness method. Indeed, nearly 36% of that consistency was above and beyond chance. Although there was a great deal of alignment between the mean score method and the skewness method, the results should be examined with caution. As noted by Murphy and Cleveland (1995), skewness ratings assume that true performance should be normally distributed around the response scale. This is an assumption that may not be the best to make in this situation. The “ratee” being evaluated was a senior level faculty member who had extensive experience teaching the course. It is not unreasonable to believe that the instructor’s true performance level is well above the scale midpoint, constraining the available scale. Indeed the mean item score across all raters was 3.19 on a 4 point scale, as opposed to the scale midpoint of 2.5. This implies that the skewness method is not calibrated properly in terms of detecting bias when true performance is not the center of the rating scale. Because the mean score method is relying on extreme scores (3.19 vs. 2.5) as the reference point it is not unreasonable to suggest that the mean score method and the skewness method are not independent measures. Indeed, extreme means are likely to lead to skewed distributions around the rating scale; therefore, it is reasonable to see high agreement between the two methods.

In terms of the mean score method, it seems relevant to reintroduce a previous argument made in this paper. Specifically, several researchers have argued that mean differences alone should not constitute bias (e.g., Dorans, 1989; Schmitt & Chan, 1998). Indeed, these researchers have noted that mean differences are not enough to determine bias (at least in terms of item bias), and merely demonstrate “impact.” From a test development standpoint this is true for items (e.g., Schmitt & Chan, 1998); therefore the same argument is legitimate for individuals. As Johanson and Alsmadi (2002) note, there may be situations in which an item (or individuals in this case) demonstrates a mean difference between focal and referent groups (i.e., impact), but does not show differential functioning, therefore it does not demonstrate bias. Based on this argument, decision makers should avoid relying on mean differences alone to determine rater bias.

Another concern is related to the generalizability of the results to other evaluation forms, as well as to more applied settings with more traditional raters and ratees. Because there was only one evaluation form used to examine the decision consistency of rater bias, it is not possible to determine if the consistency levels are unique to this particular evaluation form. Other evaluation forms could potentially be more or less susceptible to rater bias. Realistic experimental manipulations (i.e., administrative and feedback) were used to increase the generalizability and realism of the results and although the context for this study was an actual performance evaluation for an instructor, the equivalence of this type of rater bias to rater bias that occurs in employment contexts is not known. Indeed, students and instructors do not have the same didactic relationship that a supervisor and subordinate may have. For example, students typically have no future interaction with instructors. Because of this, there is no fear/unwillingness to provide honest (i.e., harsh) ratings. Indeed, as Murphy and Cleveland (1995) note, one of the reasons supervisors fail to rate subordinates accurately (i.e., provide low ratings) is because of a fear of confrontation, or a desire to maintain harmony. Additionally, instructor evaluations are more of an upward evaluation process. Even though upward ratings do occur in certain applied situations (e.g., teachers rating principals), the majority of employee evaluations are a downward process (i.e., supervisor rates subordinates).

Regardless of the differences between this educational (i.e., student) sample and typical applied work setting samples, there are still universal issues that affect both types of samples and settings. In terms of this research, the notion that students are considering multiple issues (i.e., goals) when providing ratings is no different than the multiple goals any manager or supervisor in an organization need to consider when giving ratings (i.e., organization goals, personal goals, ratee goals, etc.). Moreover, the students were led to believe the ratings would be used. Indeed, the standard instruction, standard format evaluation ratings were used for this instructor. Thus, the study and what was asked of the sample was not artificial or contrived.

Although the sample used in this research has some unique characteristics, it is important to realize though, that the purpose of this research was not to establish generalizability, but rather it was to determine the feasibility of the DPF technique for detecting biased raters. In such cases, generalizability can be considered a lesser concern (Mook, 1983; Sackett & Larson, 1990), although future research should attempt to establish generalizable results.

Although participant reactions to the DPF methodology (e.g., collection procedures) were positive and other advantages have been offered over traditional methods (e.g., bias vs. impact, distributions of performance), the results clearly indicated that all of the methods tested here failed to effectively discriminate lenient raters. It seems apparent that more research is needed to develop other methods of detecting leniency, as these three techniques were unsuccessful. As discussed in the review of this literature, there are more recent person-fit models and IRT-based models that have been proposed as methods for detecting rater errors (e.g., Wolfe, 2004). A future study could use the data from this research, but instead utilize an IRT model, whether a polytomous model such as Samejima's (1969) Graded Response Model or some other model, to determine which items are susceptible to lenient ratings (similar to Scherbaum, 2003). Rather than splitting the items according to instructor and course items, they would be examined in terms of susceptibility to leniency (or not). Not only would this allow item parameters to be estimated, but it could also allow more items to be added to an inventory. This could help to deal with the power issues associated with the current research and would allow for a comparison of different formats as well. The same process could be used for whatever rater error a researcher were interested in. If halo were of interest, instruments that have items that are susceptible to halo and ones that are not could be utilized.

Additional Goal-Based Findings

In addition to experiencing the goal manipulation, all participants were administered the Goal Importance Questionnaire and there were several interesting findings that became evident from an analysis of this questionnaire. As can be seen in Appendices O and Q, correlations between the total scores and the incidence of leniency with each item suggests that there were multiple goals that were being pursued simultaneously both within and across conditions. Indeed, as Appendix P shows, a series of ANOVAs indicated that there were no differences between the different conditions in regards to each item. Even though the participants were presented with a set of instructions as a way to manipulate rater goals, similar to previous research (i.e., Wong &

Kwong, 2007), it was difficult to rule out the possibility that raters would continued to pursue their own set of goals. Although the instructions were intended to make a particular goal salient, just as Murphy and Cleveland (1995) noted; individuals pursue multiple goals during a performance evaluation and as Austin and Vancouver (1996) note, behavior is often influenced by several goals, which may or may not be compatible with each other.

The correlational results from the Goal Importance Questionnaire items support the notion that these raters may have been pursuing multiple goals, and the instruction condition comparisons indicate that it is reasonable to expect a lot of individual differences regarding the goals that raters pursue. Goals may operate sequentially or simultaneously, and the same behavior may be part of several distinct actions. Because of this, using goals to classify behavior in terms of discrete actions may be more difficult than initially considered. Specifically, each rater's behavior may actually reflect a complex set of goals. Although Murphy and Cleveland (1995) have noted that individuals may pursue different goals, it appears that this issue is more complex than simply assigning raters to different response conditions. Indeed, it could be the case that goals arise from the evaluation process itself. For example, someone decides to give feedback because he or she feels the instructor needs it based on how the course was taught. Thus a ratee's behavior could influence a rater's goals.

Also, something else to consider is that if multiple goals are often pursued and if raters are not even aware of all of the goals they are pursuing, simply asking them what they are trying to accomplish (i.e., the open-ended manipulation check item) may not be sufficient to capture the goals involved. In addition, it may even be the case that multiple goals cancel each other out in terms of affecting ratings. If all raters were rating all the same goals, then the goal-striving issue would not be a problem. It is possible then that there are better goals to manipulate, rather than the Administrative and Feedback-related conditions that were used. Future research should attempt to determine what goals are the most relevant to this particular population (or which ever population is being examined), and manipulate those. The key to effectively evaluating bias detection methods, such as DPF, is making sure goals are chosen that raters will accept and truly consider when giving ratings.

Another aspect that future research could explore is to examine each of the items in the Goal Importance Questionnaire (each represents a specific goal) and determine what types of ratings one would expect if a rater indicated they were pursuing that goal. For example, if someone indicated that they wanted to give "Fair Ratings," does that mean that he or she should be lenient, severe, or in line with standard ratings. Even if a rater were pursuing multiple goals, it would be interesting to see if all of those goals were congruent with the same type of ratings that particular rater should be providing. This would provide evidence that an individual had indeed adopted a particular goal and was making ratings accordingly.

Although the goal-based perspective postulates that goals direct an individual's behavior (i.e., performance ratings), and there have been empirical results that have supported these notions (e.g., Murphy, et. al, 2004; Wong & Kwong, 2007), the process by which goals actually influence behavior is not fully understood. Murphy and

Cleveland (1995) mentioned that in some cases people consciously consider their goals as well as the various strategies they may adopt to accomplish those goals, but others may not. In either case, it is unlikely that individuals will evaluate and weigh several attributes of a course of action and act to maximize utility (Edwards, 1990). Because it is likely that individuals are neither aware of, nor have the capability to process all possible alternatives (as well as courses of action for each), image theory (Beach, 1990; Beach & Mitchell, 1987; Mitchell & Beach, 1990) has been proposed as a process that may be better able to explain true decision making, especially where decisions are automatic, or intuitive.

Image theory suggests that the process of making decisions involves principles, goals, and plans, and focuses on when and under what circumstances people change plans and goals (Mitchell & Beach, 1990). According to Murphy and Cleveland, image theory implies that the process of fitting ratings to goals might involve a relatively simple (and often automatic) assessment of the extent to which performance ratings are consistent with the goal the rater is pursuing. To the extent that there is an inconsistency in ratings and goals, gradual and unconscious modifications create alignment. An image theory approach could be useful if organizations were interested in determining why individuals pursued certain goals, or even if organizations intended to develop interventions that were designed to influence the goals that raters pursue. Although an image theory approach may provide insights regarding goals and their effects on rating behavior, much research is needed before it can be considered a justifiable approach.

Summary

The primary purpose of this research was to determine the usefulness of a technique for detecting rater bias in performance ratings and to provide an empirical test of a goal-based perspective for performance ratings. The results supported the goal-based approach; however, they did not support the use of the DPF method for detecting lenient raters. The DPF method was also compared to traditional methods for detecting leniency, and the results indicated that they were all equally ineffective. Although there was some consistency in the classifications of biased raters, none of the techniques were sensitive to the goal manipulation. Even though participant reactions to the DPF procedure was positive and there are arguments against the traditional approaches, at this point it only seems fair to say that all of these methods should be used with caution, if at all, to detect leniency.

An examination of the Goal Importance Questionnaire provided evidence that the raters may have been pursuing multiple goals, and that individual differences may have played a role in what goals raters pursued. A deeper examination of the individual items and what types of ratings one would expect given an endorsement of a particular goal (i.e., item) is also warranted. Image theory (Beach, 1990) and other research may help to shed light on the underlying mechanisms regarding the influence of goals on rating behavior. Other personnel implications for both the DPF technique and the goal-based approach are presented below.

Personnel Implications

In regards to the differential person functioning (DPF) technique, the results of this research have several implications for the uses of bias detection techniques as part of personnel decisions. As discussed previously, rating inflation is one of the biggest concerns regarding performance evaluations. Among those who give inflated ratings, there is a great deal of variability in the extent to which those ratings are inflated (Saal, et. al., 1980). There has been research that has utilized tools to structure evaluations and possible courses of action when measuring the performance of individuals and groups (e.g., Edwards, 1980; Pritchard, 1990). The results of this research suggest that the DPF method is not appropriate for detecting leniency; however, the results also do not support the use of traditional methods either.

Saal and colleagues (1980) noted in their seminal work on rater errors, that the mean score method is the most popular approach to detecting leniency and although there have been more sophisticated methods recently developed based on item response theory (IRT), the mean scores continues to be the most widely used method for detecting leniency. Clearly, the current research provided evidence that the mean score method, as well as the skewness method, were also ineffective at detecting leniency during the instructor evaluation. This suggests that decision makers should be cautious about using these traditional approaches in similar situations. Perhaps research that uses DPF from an IRT perspective may be an avenue that provides further insights, but clearly more effective methods for detecting rater bias are needed.

Another possibility is to encourage organizations to reconsider forced-choice ratings scales. The DPF method did receive favorable participant reactions (a major issue with forced-choice scales); however, the method was ineffective at detecting bias. Even though high development costs and low user reactions have reduced the use of forced-choice formats, they have been shown to be resistant to leniency, and may be the best alternative that is available at this time.

Aside from attempting to detect bias, the DPF technique has also been used in other personnel contexts. Scherbaum (2003) demonstrated the usefulness of the DPF technique for detecting response distortion (i.e., faking) on personality inventories. Similarly, Scherbaum and colleagues (2005) used DPF to detect differential responding in biodata items. Their study was able to identify individuals who were responding to the biodata inventory differentially as a function of unique item attributes (i.e., verifiable vs. non-verifiable). Even though the DPF method was not successful at detecting leniency/severity in this study, several avenues for future research have been presented and continued research may yield more positive results for this particular method.

Conclusions

Although measuring performance is a critical component within any organization, the process has been plagued with issues since measurement began (Murphy & Cleveland, 1995). Indeed, because most performance ratings are based on subjective indices, rater bias is bound to occur. Although an abundant amount of research has been conducted to minimize the occurrence of rater errors in instruments as well as individuals, results have been mixed at best, and little progress has occurred. There are several traditional approaches (e.g., mean scores, skewness ratings), as well as more modern approaches (e.g., IRT) that have been used to detect rater bias, however; theses approaches have not been particularly successful in identifying biased individuals, and provide little information about why a rater may give biased ratings.

This research was an initial attempt to extend an alternative method (i.e., differential person functioning), to a performance evaluation context. The initial goal was to demonstrate that the alternative method (i.e., DPF) could provide decision makers (whether researchers or practitioners) with a tool that could be used to manage and understand rater bias within performance ratings. Although the findings from this research did not support the use of the DPF method for this purpose, it provided some evidence that questioned the usefulness, and even validity of the more common and traditional methods. The goal-based perspective was clearly supported; however, future research is needed to more fully understand the mechanisms by which goals influence rating behavior. This research was merely a first step for which future research can continue.

Figure 1 –Generic Item Characteristics Curve

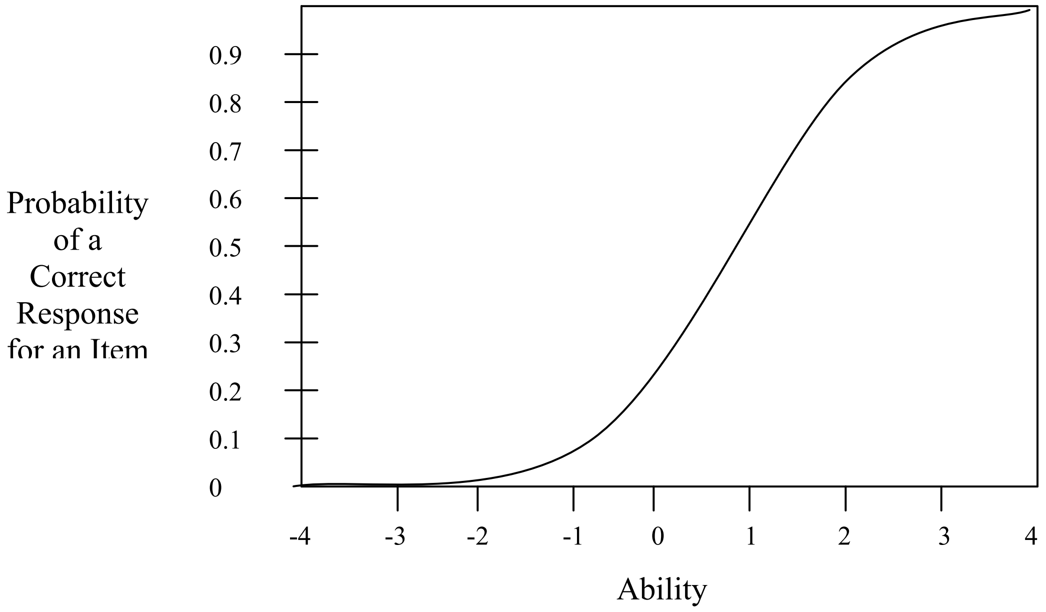

Figure 2 – DPF Analysis Showing Two PCCs for Different Types of Items: Demonstrating DPF

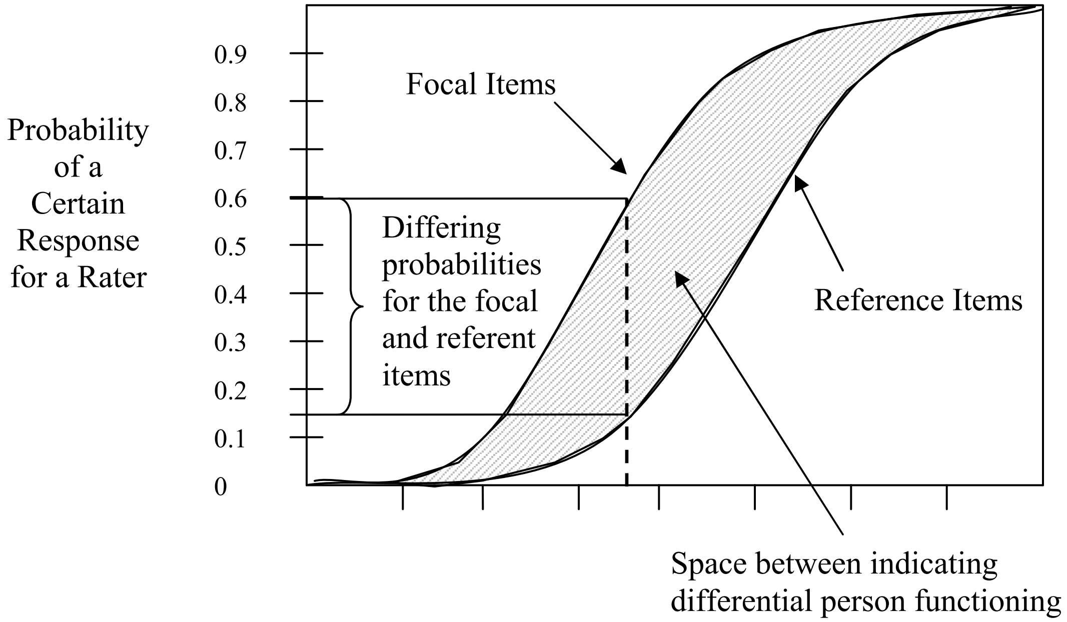

Figure 3 – Theoretical Decision Classification Table

		Technique 1		
		Bias	No Bias	
Technique 2	Bias	P_{11}	P_{10}	$P_{1.}$
	No Bias	P_{01}	P_{00}	$P_{0.}$
		$P_{.1}$	$P_{.0}$	

References

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings, *Annual Review of Psychology, 49,* 141-168.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72, 567-572.*

Austin, J. T. & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin,* 120, 338-375.

Bannister, B. D., Kinicki, A. J., DeNisi, A. S., & Hom, P. W. (1987). A new method for the statistical control of rating error in performance ratings. *Educational and Psychological Measurement, 47,* 583-596.

Barnes, J. L. & Landy, F. J. (1979). Scaling behavioral anchors. *Applied Psychological Measurement, 3,* 193-200.

Barnett, C. W., & Matthews, H. W. (1998). Current procedures used to evaluate teaching in schools of pharmacy. *American Journal of Pharmaceutical Education, 62,* 388-391.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6,* 15-43.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3,* 77-85.

Beach, L. R. (1990). *Image theory: Decision making in personal and organizational contexts.* Chichester, UK: John Wiley.

Beach, L. R., & Mitchell, T. R. (1987). Image theory: Principles, plans and goals in decision making. *Acta Psychologica,* 66, 201-220.

Bernardin , H. J. (1977). Behavioral expectation scales versus summated rating scales: A fairer comparison. *Journal of Applied Psychology, 62,* 422-427.

Bernardin, H. J., Alvarez, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology, 61,* 564-570.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

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

Bernardin, H. J., & Cascio, W. F. (1988). Performance appraisal and the law. In R. Schuler and S. Youngblood (Eds.), *Readings in Personnel/Human Resources,* 248-252. St. Paul: West Publishing.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvarez, K. M. (1976). Behavioral expectation scales: effects of developmental procedures and formats. *Journal of Applied Psychology, 61,* 75-79.

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

Bernardin, H. J., Orban, J. A., & Carlyle, J. (1981). Performance ratings as a function of trust in performance appraisal and rater individual differences. *Proceedings of the 41st annual meeting of the Academy of Management,* 311-315.

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

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. *Journal of Applied Psychology, 66,* 458-463.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. Locke (Ed.), *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Books.

Bernardin, H. J., & Villanova, P. (2005). Research streams in rater self-efficacy. *Group & Organization Management, 30,* 61-88.

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

Bernthal, P., Sumlin, R., Davis, P., & Rogers, B. (1997). *Performance management practices survey report.* Pittsburgh, PA: Developement Decisions International.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25,* 185-199.

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

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology (vol. 2, 2nd ed.).* Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology, 60,* 561-565.

Borman, W. C., & Vallon, W. R. (1974). A review of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology, 59,* 197-201.

Burnaska, R. F., & Hollmann, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology, 59,* 307-312.

Camilli, G., & Sheppard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications, Inc.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57,* 15-22.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice,* 31-44.

Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. In G. Ferris, & K. Rowland (Eds.), *Research in personnel and human resources management (vol. 10, pp. 121-185).* Greenwich, CT: JAI Press.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the linking methods under the graded response model. *Applied Psychological Measurement, 20,* 15-26.

Cotton, J., & Stoltz, R. E. (1960). The general applicability of a scale for rating research productivity. *Journal of Applied Psychology, 44,* 276-277.

Cozan, L. W. (1959). Forced choice: Better than other rating methods? *Personnel Psychology, 36,* 80-83.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth.

Damron, J. C. (1996). *Instructor personality and the politics of the classroom,* available at: www.mankato.msus.edu/dept/psych.htm.

Day, D. V., & Slusky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80,* 158-167.

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

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217-233.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*. Vol 1. Palo Alto, CA: Counsulting Psychologists Press, Inc.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Driver, R. S. (1942). Training as a means of improving employee performance ratings. *Personnel, 18,* 364-370.

Edwards, W. (1980). Multiattribute utility for evaluation: Structures, uses, and problems. In M. Klein & K. Teilmann (Eds.), *Handbook of criminal justice evaluation.* Beverly Hills: Sage.

Embertson, S., & Yang, X. (2006). Item response theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in educational research*, Mahwah, NJ: Lawrene Erlbaum.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different sources comparable? *Journal of Applied Psychology, 86,* 215-227.

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

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A p erceptual cognitive viewpoint. In K. Rowland & J. Ferris (Eds.), *Research in personnel and human resources management (vol. 4).* Greenwich, CT: JAI Press.

Feldman, K. A. (1978). Course characteristics and college students' rating of their teachers: What we know and what we don't. *Research in Higher Education, 9,* 199-242.

Feldman, K. A. (1996). Identifying exemplary teaching: Using data from course and teacher evaluations. *New Directions for Teaching and Learning, 65*, 41-50.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Utility of the Mantel-Haenszel procedure for deteting differential item functioning in small samples. *Educational and Psychological Measurement, 64,* 925-936.

Flowers, C. P., & Hancock, D. R. (2003). An interview protocol and scoring rubric for evaluating teacher performance. *Assessment in Education, 10*, 161-168.

Friedman, B. A., & Cornelius, E. T. (1976). Effects of rater participation in scale construction on the psychometric characteristics of two rating scale formats. *Journal of Applied Psychology, 61,* 210-216.

Funder, D. C. (1987). Errors and mistakes: evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75-90.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational & Psychological Measurement, 55*, 377-393.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: applying a mulit-level multi-domain perspective, *Leadership Quarterly,* 6, 219-247.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Guilford, J. P. (1954). *Psychometric methods*. (2nd ed.), New York: McGraw-Hill.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-162.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*, Newbury Park, CA: Sage Publications Inc.

Harvey, R. J. (1982). The future of partial correlation as a means to reduce halo in performance ratings. *Journal of Applied Psychology, 67,* 171-176.

Heneman, H. G., Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biological data, and interview order. *Journal of applied Psychology, 60,* 748-753.

Higgins, E. T., & King, G. (1981). Accessibility o social constructs: Information processing consequences of individual and contextual variability. In. N. Cantor & J. Kihlstrom (Eds.), *Personality, cognition, and social interaction.* Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity (pp. 129-145)*. Hillsdale, NJ: Lawrence Erlbaum.

Holzbach, R. L. (1978). Rater bias in performance ratings: superior, self-, and peer ratings. *Journal of Applied Psychology, 65,* 579-588.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5,* 64-86.

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

Hyde, A. C. (1982). Performance appraisal in the post-reform era. *Public Personnel Management,* 11, 294-305.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research inthe 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321-368.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior (vol. 5)*. Greenwich, CT: JAI Pres.

Johanson, G., & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62,* 435-443.

Johanson, G. A., & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education, 29,* 535-548.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kapel, D. E. (1974). Assessment of a conceptually based instructor evaluation form. *Research in Higher Education, 2*, 1-24.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A Windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments & Computers, 32*, 389-395.

Keaveny, T. J., & McGann, A. F. A. (1975). A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology, 60,* 695-703.

Kneeland, N. (1929). That lenient tendency in rating. *Personnel Journal, 7,* 356-366.

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

Landy, F. J. (1986). *Psychology of work behavior (3rd ed.),* Homewood, IL: Dorsey Press.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psycholgical 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., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for raring the performance of police officers. *Journal of Applied Psychology, 61,* 750-758.

Landy, F. J., & Gion, R. M. (1970). Development of scales for the measurement of work motivation. *Organization Behavior and Human Performance, 5,* 93-103.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize r ating errors in the observation of behavior. *Journal of Applied Psychology, 60,* 550-555.

Lombardo, M. M., & McCauley, C. D. (1990). *Benchmarks development reference points for managers and executives.* Greensboro, NC: Center for Creative Leadership.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monography, No. 7.* Iowa City, IA: Psychometric Society.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N. J.: Lawrence Erlbaum.

Lovell, G. D., & Haner, C. F. (1955). Forced choice applied to college faculty rating. *Educational and Psychological Measurement, 15,* 291-304.

Madden, J. M., & Bourdon, R. D. (1964). Effects of variations in rating scale format on judgment. *Journal of Applied Psychology, 48,* 147-151.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

March, J. C., & Simon, H. A. (1959). *Organizations*. New York: Wiley

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83,* 693-702

Mitchell, T. R., & Beach, L. R. (1990). Do I love thee? Let me count: Toward an understanding of intuitive and automatic decision making. *Organizational Behavior and Human Decision Processes,* 47, 1-20.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379-387.

Muchinsky, P. M. (2006). *Psychology Applied to Work (8th ed.).* Belmont, CA: Wadsworth.

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

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

Murphy, K. R., Balzer, W. K., Lockhart, M., & Eisenman, E. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*, 72-84.

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

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage Publications, Inc.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of applied Psychology, 89,* 158-164.

Murphy, K. R. & Constans, J. J. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology, 72*, 523-579.

Murphy, K. R., & Constans, J. J. (1988). Psychological issues in scale format research: Behavioral anchors as a source of bias in rating. In R. Cardy, S. Peiffer, & J. Newman (Eds.), *Advances in information processing in organizations (vol. 3).* Greenwich, CT: JAI Pres.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67,* 562-567.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet rasch measurement: Part II. *Journal of Applied Measurement, 5,* 189-227.

Newcomb, T. (1931). A design to test the validity of a rating technique. *Journal of Educational Psychology, 22,* 279-289.

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage Publications Inc.

Paterson, D. G. (1922). The Scott Company graphic rating scale. *Journal of Personnel Research,* 1, 351-376.

Phipps, S. D., Kidd, R. S., & Latif, D. A. (2006). Relationships among student evaluations, instructor effectiveness, and academic performance. *Pharmacy Education, 6,* 237-243.

Popovich, P. (2007). Personal correspondence, August 22, 2007.

Pritchard, R. D. (1990). *Measuring and improving organizational productivity: A practical guide*. New York: Praeger.

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

Raju, N., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19,* 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Riggio, R. E. (2003). *Introduction to Industrial/Organizational Psychology.* (4th Ed.). Upper Saddle River: NJ. Prentice Hall.

Ritti, R. R. (1964). Control of "halo" in factor analysis of a supervisory behavior inventory. *Personnel Psychology, 17,* 305-318.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1-10.

Ryan, F. J. (1958). Trait ratings of high school students by teachers. *Journal of Educational Psychology, 49,* 124-128.

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

Saal, F. E., & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance, 18,* 19-35.

Sackett, P. R., & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette, and L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology,* vol. 1 (pp 417-489), Palo Alto: CA, Consulting Psychologists Press.

Samejima, F. (1969). *Estimation of latent ability using response pattern of graded scores.* Psychometric Monograph No. 17, Iowa City, IA: Psychometric Society.

Scherbaum, C. A. (2003). *Detecting intentional response distortion on measures of the five-factor model of personality: An application of differential person functioning.* Unpublished Dissertation, Ohio University, Athens, Ohio.

Scherbaum, C. A., Yusko, K., Goldstein, H., & Kern, M. (2005). Differential person functioning related to biodata item attributes. Poster presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angles, CA.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousand Oaks, CA: Sage Publications, Inc.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84, 1-66.

Sharon, A., & Bartlett, C. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 22,* 252-263.

Silverman, S. B., & Wexley, K. N. (1984). Reactions of employees to performance appraisal interviews as a function of their participation in rating scale development. *Personnel Psychology, 37,* 703-710.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26,* 208-227.

Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training an rating accuracy: training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78,* 994-1003.

Staugas, L., & McQuitty, L. L. (1950). A new application of forced-choice ratings. *Personnel Psychology, 3,* 413-424.

Stockford, L., & Bissell, H. W. (1949). Factors involved in establishing a merit-rating scale. *Personnel, 26,* 94-116.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77,* 501-510.

Sulsky, L. M., & Keown, J. L. (1997). Performance appraisal in the changing world of work: Implications for the meaning and measurement of work performance. *Canadian Psychology, 39,* 52-59.

Swaminathan, H. Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation. *Journal of Educational Measurement,* 11, 263 – 268.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics (3rd ed).* New York: Harper Collins.

Taylor, E. K., Schneider, D. E., & Clay, H. C. (1954). Short forced-choice ratings work. *Personnel Psychology, 7,* 245-252.

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

Thissen, D. M., & Steinberg, L. (1986). A taxonomy of *item* response models. *Psychometrika, 51,* 567-577.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Tsui, A. S. & Barry, B. (1986). Interpersonal affect and rating errors. *Academy of Management Journal, 29,* 586-599.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group & Organization Management, 30,* 89-98.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15-25.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer.

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. *Organizational Behavior and Human Performance, 22,* 279-294.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 233-236.

Whisler, T. L., & Harper, S. F. (Eds.). (1962). *Performance appraisal: Research and practice*. New York: Hold, Rinehart, & Winston.

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

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46,* 35-51.

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study, *Journal of Applied Psychology, 92,* 577-585.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71-87.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Appendix A

Student Evaluation Form (Dichotomous Response Format)

Instructor Items

Intellect/knowledge

1. This instructor is knowledgeable in the field.

Agree Disagree

2. This instructor demonstrated command of the subject matter.

Agree Disagree

3. This instructor gives information and viewpoints not found in text.

Agree Disagree

4. This instructor answers course related questions effectively.

Agree Disagree

5. This instructor uses current information from the field in his/her lectures.

Agree Disagree

Motivation/Learning/Stimulation of Interest

6. This instructor creates a desire to learn and do well in this course.

Agree Disagree

7. This instructor motivated me to do my best.

Agree Disagree

8. This instructor wants to see all of his/her students do well.

Agree Disagree

9. This instructor got me interested in this subject.

Agree Disagree

10. This instructor helped me become interested in this material.

Agree Disagree

11. This instructor demonstrated a genuine interest in educating students.

Agree Disagree

12. This instructor was enthusiastic about the subject.

Agree Disagree

Preparation and Organization

13. This instructor is organized and well prepared for class.

Agree Disagree

14. This instructor makes good use of examples and illustrations.

Agree Disagree

15. This instructor is punctual (beginning/ending class on time).

Agree Disagree

16. This instructor returned exams and assignments in a timely manner.

Agree Disagree

17. This instructor was available outside of class time to give assistance.

Agree Disagree

18. This instructor utilized office hours effectively.

Agree Disagree

19. This instructor provided alternate resources for student assistance.

Agree Disagree

20. This instructor is able to cover the material in a timely manner without rushing.

Agree Disagree

21. This instructor utilized supplemental materials when needed.

Agree Disagree

Student Development/Learning Environment

22. This instructor encouraged student participation.

Agree Disagree

23. This instructor is insensitive to students' needs and problems.

Agree Disagree

24. This instructor sees students only as students and not individuals.

Agree Disagree

25. This instructor has no problems with students' questions.

Agree Disagree

26. This instructor respects the comments and suggestions of students.

Agree Disagree

27. This instructor helps students understand the course material.

Agree Disagree

Presentation

28. This instructor is clear and understandable when explaining class material.

Agree Disagree

29. This instructor speaks at a reasonable speech rate.

Agree Disagree

30. This instructor appears nervous and unable to effectively present the course material.

Agree Disagree

31. This instructor has difficulty expressing lecture material clearly.

Agree Disagree

32. This instructor takes different learning styles into account.

Agree Disagree

33. This instructor is effective at presenting material to others.

Agree Disagree

34. This instructor uses multiple instructional strategies (e.g., lecture, video, discussion, etc.).

Agree Disagree

35. This instructor explained difficult material clearly.

Agree Disagree

36. This instructor has nervous habits that interfere with the learning process.

Agree Disagree

Personality

37. This instructor has a good sense of humor.

Agree Disagree

38. This instructor has a personality that is well suited for teaching this course.

Agree Disagree

39. This instructor has a personality that is well suited for teaching in general.

Agree Disagree

40. This instructor has a poor attitude towards students.

Agree Disagree

8. Evaluation

41. For the amount of work done the instructor graded too harshly.

Agree Disagree

42. For the amount of work done the instructor graded fairly.

Agree Disagree

43. The instructor clearly explained the grading system.

Agree Disagree

Course Items

Organization

1. This course was well organized.

Agree Disagree

2. The use of instructional materials was effective.

Agree Disagree

3. The content of this course is current with the knowledge and issues in the field.

Agree Disagree

4. The format of this course is appropriate.

Agree Disagree

5. The material covered in this course was what I though it would be.

Agree Disagree

6. The use of technology was utilized to promote learning.

Agree Disagree

7. The number of students in this class is appropriate for this course.

Agree Disagree

8. The course content followed a logical progression.

Agree Disagree

Course Level/Difficulty

9. The difficulty in this course was appropriate.

Agree Disagree

10. The amount of material covered in this course is acceptable.

Agree Disagree

11. This course challenged me intellectually.

Agree Disagree

12. Attendance is necessary for understanding this material.

Agree Disagree

13. The course level designation (e.g., 100 level, 200 level, 300 level) assigned by the university is appropriate for this course.

Agree Disagree

14. Overall, this is a useful course.

Agree Disagree

Goals/Objectives/Electivity

15. This is an important course for students to take.

Agree Disagree

16. The format of the course is appropriate for the course objectives.

Agree Disagree

17. I would recommend this course to another student.

Agree Disagree

18. This course achieved its stated objectives.

Agree Disagree

19. Course requirements were clearly stated and followed.

Agree Disagree

20. This course improved my written communication skills.

Agree Disagree

21. This course improved my oral communication skills.

Agree Disagree

22. Overall this course is of great value.

Agree Disagree

23. The amount of material covered in this course is fair.

Agree Disagree

24. This course content is enjoyable.

Agree Disagree

<u>Subject matter of the course</u>

25. This course material is interesting to me.

Agree Disagree

26. The concepts from one topic flowed well into the concepts from other topics.

Agree Disagree

27. This course taught me to understand arguments on this topic.

Agree Disagree

28. There is always enough time to cover the needed material.

Agree Disagree

29. I learned much new information from taking this course.

Agree Disagree

30. This course will/has helped me understand information from my main area of study.

Agree Disagree

Evaluation

31. The evaluation procedures (exams) utilized in this course were fair.

Agree Disagree

32. Exams and quizzes helped me find my strengths and weaknesses.

Agree Disagree

33. Exams and/or quizzes cover material presented in class/textbook/activities.

Agree Disagree

34. The grading criteria were clearly communicated in this course.

Agree Disagree

35. The evaluation tools (exams/assignments) were appropriate for this course.

Agree Disagree

36. The material was covered in a meaningful/appropriate progression.

Agree Disagree

37. Assignments were returned in a reasonable period of time.

Agree Disagree

38. Readings are an important for understanding this material.

Agree Disagree

39. Homework assignments are a useful part of this course.

Agree Disagree

40. The supplemental material in this course is helpful.

Agree Disagree

1. What was the purpose of the ratings you were giving (what was your goal when rating the instructor)? ____

2. Do you like this instructor? Yes No

3. What grade do you expect to earn in this course? A B C D F

4. What is your sex? Male Female

Appendix B

Student Evaluation Form (Likert-Type Response Format)

Instructor Items

Intellect/knowledge

1. This instructor is knowledgeable in the field.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

2. This instructor demonstrated command of the subject matter.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

3. This instructor gives information and viewpoints not found in text.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

4. This instructor answers course related questions effectively.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

III. This instructor uses current information from the field in his/her lectures.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Motivation/Learning/Stimulation of Interest

6. This instructor creates a desire to learn and do well in this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

7. This instructor motivated me to do my best.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

8. This instructor wants to see all of his/her students do well.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

9. This instructor got me interested in this subject.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

10. This instructor helped me become interested in this material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

11. This instructor demonstrated a genuine interest in educating students.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

12. This instructor was enthusiastic about the subject.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Preparation and Organization

13. This instructor is organized and well prepared for class.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

14. This instructor makes good use of examples and illustrations.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

15. This instructor is punctual (beginning/ending class on time).

Strongly			Strongly
Agree	Agree	Disagree	Disagree

16. This instructor returned exams and assignments in a timely manner.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

17. This instructor was available outside of class time to give assistance.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

18. This instructor utilized office hours effectively.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

19. This instructor provided alternate resources for student assistance.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

20. This instructor is able to cover the material in a timely manner without rushing.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

21. This instructor utilized supplemental materials when needed.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

<u>Student Development/Learning Environment</u>

22. This instructor encouraged student participation.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

23. This instructor is insensitive to students’ needs and problems.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

24. This instructor sees students only as students and not individuals.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

25. This instructor has no problems with students' questions.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

26. This instructor respects the comments and suggestions of students.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

27. This instructor helps students understand the course material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Presentation

28. This instructor is clear and understandable when explaining class material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

29. This instructor speaks at a reasonable speech rate.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

30. This instructor appears nervous and unable to effectively present the course material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

31. This instructor has difficulty expressing lecture material clearly.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

32. This instructor takes different learning styles into account.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

33. This instructor is effective at presenting material to others.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

34. This instructor uses multiple instructional strategies (e.g., lecture, video, discussion, etc.).

Strongly			Strongly
Agree	Agree	Disagree	Disagree

35. This instructor explained difficult material clearly.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

36. This instructor has nervous habits that interfere with the learning process.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Personality

37. This instructor has a good sense of humor.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

38. This instructor has a personality that is well suited for teaching this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

39. This instructor has a personality that is well suited for teaching in general.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

40. This instructor has a poor attitude towards students.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

8. Evaluation

41. For the amount of work done the instructor graded too harshly.

Strongly Strongly
Agree Agree Disagree Disagree

42. For the amount of work done the instructor graded fairly.

Strongly Strongly
Agree Agree Disagree Disagree

43. The instructor clearly explained the grading system.

Strongly Strongly
Agree Agree Disagree Disagree

Course Items

Organization

1. This course was well organized.

Strongly Strongly
Agree Agree Disagree Disagree

2. The use of instructional materials was effective.

Strongly Strongly
Agree Agree Disagree Disagree

3. The content of this course is current with the knowledge and issues in the field.

Strongly Strongly
Agree Agree Disagree Disagree

4. The format of this course is appropriate.

Strongly Strongly
Agree Agree Disagree Disagree

5. The material covered in this course was what I though it would be.

Strongly Strongly
Agree Agree Disagree Disagree

6. The use of technology was utilized to promote learning.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

7. The number of students in this class is appropriate for this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

8. The course content followed a logical progression.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Course Level/Difficulty

9. The difficulty in this course was appropriate.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

10. The amount of material covered in this course is acceptable.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

11. This course challenged me intellectually.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

12. Attendance is necessary for understanding this material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

13. The course level designation (e.g., 100 level, 200 level, 300 level) assigned by the university is appropriate for this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

14. Overall, this is a useful course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

<u>Goals/Objectives/Electivity</u>

15. This is an important course for students to take.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

16. The format of the course is appropriate for the course objectives.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

17. I would recommend this course to another student.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

18. This course achieved its stated objectives.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

19. Course requirements were clearly stated and followed.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

20. This course improved my written communication skills.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

21. This course improved my oral communication skills.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

22. Overall this course is of great value.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

23. The amount of material covered in this course is fair.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

24. This course content is enjoyable.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

<u>Subject matter of the course</u>

25. This course material is interesting to me.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

26. The concepts from one topic flowed well into the concepts from other topics.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

27. This course taught me to understand arguments on this topic.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

28. There is always enough time to cover the needed material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

29. I learned much new information from taking this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

30. This course will/has helped me understand information from my main area of study.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

Evaluation

31. The evaluation procedures (exams) utilized in this course were fair.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

32. Exams and quizzes helped me find my strengths and weaknesses.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

33. Exams and/or quizzes cover material presented in class/textbook/activities.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

34. The grading criteria were clearly communicated in this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

35. The evaluation tools (exams/assignments) were appropriate for this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

36. The material was covered in a meaningful/appropriate progression.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

37. Assignments were returned in a reasonable period of time.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

38. Readings are an important for understanding this material.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

39. Homework assignments are a useful part of this course.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

40. The supplemental material in this course is helpful.

Strongly			Strongly
Agree	Agree	Disagree	Disagree

1. What was the purpose of the ratings you were giving (what was your goal when rating the instructor)? ____

2. Do you like this instructor? Yes No

5. What grade do you expect to earn in this course? A B C D F

3. What is your sex? Male Female

Appendix C

Evaluation Process and Participant Reactions

1. This evaluation is a fair way of evaluating an instructor's level of performance.

   Strongly Agree Agree Disagree Strongly Disagree

2. The psychology department should adopt this evaluation form for future instructor evaluations.

   Strongly Agree Agree Disagree Strongly Disagree

3. You can control the ratings you give (you can give high or low ratings when they are appropriate).

   Strongly Agree Agree Disagree Strongly Disagree

4. This type of evaluation gave you the ability to give objective ratings of your instructor.

   Strongly Agree Agree Disagree Strongly Disagree

5. Please rate the level of difficulty in using this evaluation method.

   Extremely Difficult Difficult Easy Extremely Easy

6. Please rate your level of satisfaction in using this evaluation method.

   Extremely Satisfied Satisfied Unsatisfied Extremely Unsatisfied

Appendix D

Goal Questionnaire (Murphy, Cleveland, Skattebo, Kinney, 2004)

1. Identify areas in which the instructor might need improvement.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

2. Rate my instructor fairly.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

3. Identify areas where the instructor needs more training.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

4. Convey my satisfaction with the instructor’s performance.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

5. Identify area that the instructor should focus on improving.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

6. Indicate where the instructor fell short in terms of performance.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

7. Give my instructor a rating that she or he will realize is based on performance, rather than my judgment of him or her as a person.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

8. Identify my instructor’s strengths and weaknesses.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

9. Highlight my instructor’s performance so that his or her success is visible to his or her department head.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

10. Improve my instructor’s confidence.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

11. Make it clear to my instructor that there is room for improvement.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

12. Identify my instructor’s performance deficiencies.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

13. Challenge my instructor to improve his or her performance.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

14. Clarify expected performance levels to the instructor.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

15. Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

16. Indicate where instructor has exceeded performance expectations.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

17. Encourage the instructor’s current level of performance.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

18. Encourage the instructor to improve performance.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

19. Motivate the instructor.

Strongly				Strongly
Agree	Agree	Neutral	Disagree	Disagree

Appendix E

Pilot Study Purpose

Hello,
The purpose of this study is to develop and refine a new evaluation instrument that may be used for evaluating instructors. The questionnaires that you will be presented with contain questions related to the course, the instructor, and the goals (purposes) for giving ratings.

Please consider your PSY101 (or PSY120) course and instructor when filling out these evaluation forms.

PLEASE DO NOT SKIP AHEAD...ANSWER THE QUESTIONS/SECTIONS IN THE ORDER THEY APPEAR.

Please read the Evaluation Instructions CAREFULLY so that you understand the PURPOSE for giving ratings (what they will be used for).

Thank you for your help and participation!

Appendix F

Online Consent Form

TITLE OF RESEARCH: Detecting Differential Rater Functioning in Performance
Ratings
PRINCIPLE INVESTIGATOR: Kevin B. Tamanini
DEPARTMENT: Psychology

I. Federal and university regulations require us to obtain consent for participation in research involving human subjects. After reading the statement below, please indicate your consent by clicking the button below.

II. STATEMENT OF PROCEDURE:

I understand that I will be asked to participate in a survey dealing instructor evaluations and goals. I understand that this study will take approximately 1 hour of my time should I complete the study. **I understand that for my participation in this study I will earn 1 credit toward mandatory or extra credit in certain general psychology courses.** I understand that my main task is to answer questions regarding the performance of my instructor and goals that I pursue. I understand that the results of my participation in the study and my responses to questions during the study will be kept in the strictest of confidence. Any identifying information, such as this signed consent form, will be kept separate from the data collected.

There are no known risks for participating in this research. The benefits include helping the investigators the nature of performance evaluations and participating in a research project. If you have any questions about this study, please contact Kevin Tamanini, Psychology Department, Ohio University at kt109402@ohio.edu.

III. I certify that I have read and understood the statement of procedure and agree to participate as a subject in the specific research described therein. I agree that all known risks to me have been explained to my satisfaction and I understand that no compensation is available from Ohio University and its employees for any injury resulting from my participation in this research. My participation in this research is given voluntarily. I understand that I may discontinue participation at any time without penalty or loss of any benefits to which I may otherwise be entitled. I certify that I am at least 18 year of age.

If you have any questions regarding your rights as a research participant, please contact Jo Ellen Sherow, Director of Research Compliance, Ohio Universtiy, (740)-593-0664.

By clicking the button below, you signify that you have read and understand this consent form and have given your consent to participate in this study.

Consent
Please click to continue

Appendix G

Psychology Department Faculty Evaluation Form

1. The instructor is knowledgeable in the field:

1	2	3	4	5
not at all		average		very much

2. The instructor is clear and understandable when explaining class material:

1	2	3	4	5
not at all		average		very much

3. The instructor makes appropriate use of examples and illustrations in explaining ideas:

1	2	3	4	5
not at all		occasionally		a lot

4. The instructor gives information and viewpoints not found in the text:

1	2	3	4	5
not at all		occasionally		a lot

5. The instructor is interested and enthusiastic about teaching:

1	2	3	4	5
not at all		average		very much

6. The instructor got me interested in the subject:

1	2	3	4	5
not at all		average		very much

7. From being in this course, I have learned:

1	2	3	4	5
nothing		average		much new information

8. This course helps me to understand and evaluate arguments and discussions on topics in this field:

1	2	3	4	5
not at all		occasionally		a lot

9. Readings and assignments are an important and useful part of this course:

1	2	3	4	5
not at all		occasionally		always

10. The difficulty of the work in this course was:

1	2	3	4	5
too easy		OK		too hard

11. Taking into account the amount of work I did, I feel that the instructor for this course graded:

1	2	3	4	5
too easy		just right		too hard

12. In an overall evaluation, I rate the examinations, quizzes, or other methods of evaluation:

1	2	3	4	5
very poor		average		very good

13. In an overall evaluation, I rate the instructor:

1	2	3	4	5
very poor		average		very good

14. I would recommend this course to another student:

1	2	3	4	5
not at all		maybe		definitely

15. I would recommend this instructor to another student:

1	2	3	4	5
not at all		maybe		definitely

16. In this class, I expect to receive a grade of:

1	2	3	4	5
A	B	C	D	F

17. My grade point average is approximately:

1	2	3	4	5
1.0	2.0	3.0	4.0	First Quarter

18. My interest in this subject before taking the course was:

1	2	3	4	5
very small		average		very great

Appendix H

Please express your opinions about the following aspects of the course. Your response will be a valuable feedback to the instructors of this class. Any kind of comments you have will be highly appreciated.

1. The text

2. The lectures and discussions

3. The instructor

4. The tests

5. The class format

6. I like best

7. I like least

Regarding this course, you might have some other constructive comments or suggestions that would improve this course. Please write down your opinions freely:

Appendix I

RESEARCH CREDIT REQUIREMENT - UPDATE

Hello everyone,

At this point in the quarter you should be close to completing your outside class research credits. At the beginning of the quarter you were told that the course evaluations this year were going to be conducted differently for this “mega section” of PSY101, and that because of this you would be able to use participation in the evaluation process as 1 credit (for use towards your 4 required credits or towards one of your 6 possible extra credit points).

This reminder is to inform you that this evaluation will be held on Thursday, November 1st during class time (1:00-2:00) in the regularly scheduled classroom (201 Morton Hall). There will be a brief lecture period at the beginning of class followed by the evaluation portion, at which time the instructor (Dr. Popovich) will leave the building.

NO SIGN-UP IS NECESSARY – SIMPLY SHOW UP TO CLASS!!! You will be asked to write your name on an index card and turn that in at the end so that the course TAs can give you credit.

Appendix J

Debriefing Form

Thank you for participating in this study. Yes, it was both an evaluation and a study! The purpose of the study was to utilize a new method for detecting rater bias within performance evaluations. Rating bias is a major issue that has influenced performance evaluations (whether instructor ratings or manager-supervisor ratings) for a long time. Because of this, there has been a great deal of research that has attempted to detect and eliminate rater bias. Unfortunately, much of the research on this topic has provided mixed results at best and the methods for detecting rater bias have not been effective. This study will utilize a new method for detecting rating bias. Additionally, there is a belief that the goals that one pursues (such as the instructions you were given) have an influence in the performance ratings that one gives. To test this question, we created different instructions to elicit different goals. For example, some were told that the ratings will be use for a promotion decision for Dr. Popovich. There was some deception here in that Dr. Popovich is not actually up for promotion at this time. However, it is true that student ratings do play a role in promotion decisions at this university. Also, some were told that the ratings will be used only for development purposes. This is also a role that the student evaluations play, but not exclusively. Finally, some were given the usual instructions for evaluations (which include the uses described above). In addition to different instructions, additional items were added to the evaluation form. Be assured, your voice will be heard, but only responses to the traditional items under the typical instruction set will be forwarded to the psychology department to keep the evaluation consistent with the typical approach. A comparison will be made between the other conditions and if no differences are detected with the traditional item set, they will also be forwarded to the department. As I mentioned previously, no names or personal information have been linked to the responses. You will receive 1 credit for your participation. If you have any questions about this study or the results please contact the principle investigator Kevin Tamanini at kt109402@ohio.edu. If you have any questions regarding your rights as a research participant, please contact Jo Ellen Sherow, Director of Research Compliance, Ohio Universtiy, (740)-593-0664.

Thank you for your participation!

Appendix K

Means and Standard Deviations in the Administrative, Feedback-Related, and Psychology Department Instruction Conditions for Each Item using the Likert-type Data (N=280)

	Admin.		Feedback		Dept.	
Item	M	SD	M	SD	M	SD
Instructor Items						
1. This instructor is knowledgeable in the field.	3.74	.44	3.73	.49	3.81	.42
2. This instructor demonstrated command of the subject matter.	3.60	.51	3.44	.58	3.60	.56
3. This instructor gives information and viewpoints not found in text.	3.62	.57	3.44	.61	3.67	.54
4. This instructor answers course related questions effectively.	3.62	.49	3.35	.50	3.49	.64
5. This instructor uses current information from the field in his/ her lectures.	3.38	.64	3.37	.555	3.45	.67
6. This instructor creates a desire to learn and do well in this course.	3.23	.71	2.97	.69	3.17	.66
7. This instructor motivated me to do my best.	2.97	.66	2.71	.65	2.93	.72
8. This instructor wants to see all of his/her students do well.	3.13	.72	3.05	.63	3.20	.64
9. This instructor got me interested in this subject.	3.00	.78	2.84	.78	3.06	.87
10. This instructor helped me become interested in this material.	3.06	.80	2.97	.74	3.18	.78
11. This instructor demonstrated a genuine interest in educating students.	3.39	.61	3.26	.67	3.48	.60

12. This instructor was enthusiastic about the subject.	3.55	.52	3.53	.50	3.67	.54
13. This instructor is organized and well prepared for class.	3.61	.53	3.48	.54	3.58	.52
14. This instructor makes good use of examples and illustrations.	3.49	.58	3.38	.53	3.49	.62
15. This instructor is punctual (beginning/ending class on time).	3.49	.54	3.38	.53	3.53	.60
16. This instructor returned exams and assignments in a timely manner.	3.38	.61	3.24	.61	3.43	.54
17. This instructor was available outside of class time to give assistance.	3.32	.53	3.23	.49	3.37	.51
18. This instructor utilized office hours effectively.	3.24	.54	3.21	.48	3.37	.53
19. This instructor provided alternate resources for student assistance.	3.16	.60	3.04	.65	3.09	.69
20. This instructor is able to cover the material in a timely manner without rushing.	3.36	.70	3.16	.72	3.30	.72
21. This instructor utilized supplemental materials when needed.	3.16	.61	3.09	.64	3.20	.62
22. This instructor encouraged student participation.	3.03	.68	2.955	.75	2.99	.71
23. This instructor is insensitive to students' needs and problems. (R)	2.95	.81	2.86	.75	2.89	.79
24. This instructor sees students only as students and not individuals. (R)	2.84	.77	2.81	.66	2.88	.75
25. This instructor has no problems with students' questions.	3.45	.56	3.31	.55	3.46	.58
26. This instructor respects the						

comments and suggestions of students.	3.28	.56	3.17	.60	3.29	.58
27. This instructor helps students understand the course material.	3.31	.57	3.13	.61	3.40	.61
28. This instructor is clear and understandable when explaining class material.	3.42	.65	3.31	.60	3.45	.60
29. This instructor speaks at a reasonable speech rate.	3.33	.77	3.33	.53	3.35	.66
30. This instructor appears nervous and unable to effectively present the course material. (R)	3.51	.60	3.56	.61	3.59	.56
31. This instructor has difficulty expressing lecture material clearly.	3.36	.69	3.39	.62	3.43	.70
32. This instructor takes different learning styles into account.	2.71	.71	2.69	.73	2.88	.66
33. This instructor is effective at presenting material to others.	3.29	.56	3.21	.50	3.35	.58
34. This instructor uses multiple instructional strategies (e.g., lecture, video, discussion, etc.).	3.04	.80	2.97	.69	3.14	.64
35. This instructor explained difficult material clearly.	3.13	.71	3.06	.67	3.19	.65
36. This instructor has nervous habits that interfere with the learning process.(R)	3.53	.52	3.46	.63	3.54	.58
37. This instructor has a good sense of humor.	3.36	.72	3.33	.55	3.56	.52
38. This instructor has a personality that is well suited for teaching this course.	3.45	.65	3.49	.52	3.55	.54
39. This instructor has a personality that is well suited for teaching in general.	3.41	.65	3.42	.54	3.58	.52

40. This instructor has a poor attitude towards students.(R)	3.48	.56	3.37	.74	3.58	.56
41. For the amount of work done the instructor graded too harshly.(R)	3.11	.70	2.91	.83	2.99	.64
42. For the amount of work done the instructor graded fairly.	3.22	.64	3.05	.69	3.22	.68
43. The instructor clearly explained the grading system.	3.48	.58	3.32	.61	3.48	.54
Course Items						
1. This course was well organized.	3.33	.63	3.33	.53	3.45	.62
2. The use of instructional materials was effective.	3.19	.63	3.08	.52	3.21	.64
3. The content of this course is current with the knowledge and issues in the field.	3.27	.57	3.15	.48	3.31	.53
4. The format of this course is appropriate.	3.21	.60	3.14	.58	3.30	.59
5. The material covered in this course was what I though it would be.	3.14	.63	3.04	.60	3.16	.64
6. The use of technology was utilized to promote learning.	2.87	.83	2.83	.68	2.95	.72
7. The number of students in this class is appropriate for this course.	2.74	.78	2.69	.69	2.80	.73
8. The course content followed a logical progression.	3.32	.57	3.17	.50	3.25	.57
9. The difficulty in this course was appropriate.	2.97	.71	2.84	.69	2.99	.69
10. The amount of material covered in this course is acceptable.	3.15	.69	2.99	.72	3.21	.57

11. This course challenged me intellectually.	3.36	.57	3.17	.61	3.24	.58
12. Attendance is necessary for understanding this material.	3.40	.71	3.20	.79	3.31	.79
13. The course level designation (e.g., 100 level, 200 level, 300 level) assigned by the university is appropriate for this course.	3.17	.68	2.93	.66	3.14	.68
14. Overall, this is a useful course.	3.31	.66	3.12	.62	3.31	.73
15. This is an important course for students to take.	3.18	.66	2.98	.67	3.22	.71
16. The format of the course is appropriate for the course objectives.	3.19	.63	3.15	.56	3.31	.53
17. I would recommend this course to another student.	3.19	.77	2.97	.75	3.14	.83
18. This course achieved its stated objectives.	3.31	.57	3.26	.51	3.34	.54
19. Course requirements were clearly stated and followed.	3.39	.55	3.29	.52	3.35	.58
20. This course improved my written communication skills.	2.17	.59	2.09	.70	2.04	.67
21. This course improved my oral communication skills.	2.15	.65	2.08	.68	2.07	.66
22. Overall this course is of great value.	3.11	.66	3.04	.63	3.14	.69
23. The amount of material covered in this course is fair.	3.18	.62	2.93	.64	3.05	.56
24. This course content is enjoyable.	3.06	.73	3.05	.72	3.16	.78
25. This course material is interesting to me.	3.18	.76	3.17	.70	3.24	.78

26. The concepts from one topic flowed well into the concepts from other topics.	3.30	.55	3.05	.59	3.21	.55
27. This course taught me to understand arguments on this topic.	3.15	.60	2.88	.56	3.04	.64
28. There is always enough time to cover the needed material.	3.13	.72	2.98	.67	3.15	.73
29. I learned much new information from taking this course.	3.36	.69	3.23	.61	3.34	.65
30. This course will/has helped me understand information from my main area of study.	2.69	.82	2.65	.81	2.84	.83
31. The evaluation procedures (exams) utilized in this course were fair.	3.04	.82	2.97	.72	3.08	.78
32. Exams and quizzes helped me find my strengths and weaknesses.	2.80	.77	2.73	.82	2.87	.75
33. Exams and/or quizzes cover material presented in class/textbook/activities.	3.45	.54	3.37	.60	3.40	.63
34. The grading criteria were clearly communicated in this course.	3.44	.52	3.29	.52	3.38	.59
35. The evaluation tools (exams/ assignments) were appropriate for this course.	3.17	.74	3.10	.66	3.15	.70
36. The material was covered in a meaningful/appropriate progression.	3.38	.64	3.23	.63	3.34	.58
37. Assignments were returned in a reasonable period of time.	3.13	.68	3.10	.70	3.26	.58
38. Readings are an important for understanding this material.	3.55	.56	3.52	.67	3.48	.64
39. Homework assignments are a useful part of this course.	2.33	.83	2.26	.78	2.37	.84

40. The supplemental material in this course is helpful.	3.02	.61	2.97	.52	3.05	.54

Note. Admin. = Administrative Response Condition; Feedback = Feedback-Related Response Condition; Dept. = Psychology Department Response Condition. All items have been recoded such that higher values are more desirable; (R) = items that were reverse scored.

Appendix L

Results of the Differential Person Functioning Analysis Using the Mantel-Haenszel Procedure for Dichotomous Scores

Rater	Condition	χ^2_{MH}	α	|*Delta*|	Directionality
1	Administrative	0.044	0.000	----	----
2	Administrative	----	----	----	----
3	Administrative	4.191***	0.000	----	----
4	Administrative	**4.327*** **	**0.261**	**3.156**	**L**
5	Administrative	0.213	0.000	----	----
6	Administrative	0.213	0.000	----	----
7	Administrative	**1.453***	**4.689**	**3.631**	**S**
8	Administrative	**1.209***	**0.207**	**3.701**	**L**
9	Administrative	0.535	0.570	1.321	L
10	Administrative	0.341	0.650	1.012	L
11	Administrative	0.409	0.292	2.893	L
12	Administrative	5.579*	0.000	----	L
13	Administrative	0.061	1.005	0.012	S
14	Administrative	0.064	0.972	0.067	L
15	Administrative	0.005	1.600	1.105	S
16	Administrative	0.467	0.280	2.991	L
17	Administrative	**2.694***	**0.277**	**3.017**	**L**
18	Administrative	0.063	1.069	0.157	S
19	Administrative	0.004	2.250	1.906	S
20	Administrative	**5.272*** **	**0.135**	**4.706**	**L**
21	Administrative	**1.515***	**2.158**	**1.808**	**S**
22	Administrative	0.001	1.123	0.287	S
23	Administrative	0.004	0.906	0.232	L
24	Administrative	0.314	0.553	1.392	L
25	Administrative	**2.750*** *	**0.211**	**3.656**	**L**
26	Administrative	0.433	0.000	----	
27	Administrative	0.004	1.362	0.726	S
28	Administrative	**11.276*** ***	**0.120**	**4.98**	**L**
29	Administrative	0.825	0.137	4.671	L
30	Administrative	0.473	1.916	1.528	S
31	Administrative	0.273	0.440	1.929	L
32	Administrative	----	----	----	----
33	Administrative	**1.926***	**0.418**	**2.050**	**L**
34	Administrative	0.046	0.429	1.989	L
35	Administrative	0.798	1.787	1.364	S
36	Administrative	----	----	----	----
37	Administrative	0.140	1.449	0.871	S
38	Administrative	**1.082***	**0.486**	**1.696**	**L**

39	Administrative	0.200	1.674	1.211	S
40	Administrative	0.320	1.167	0.363	S
41	Administrative	0.413	2.248	1.904	S
42	Administrative	**2.975****	**3.076**	**2.859**	**S**
43	Administrative	0.629	0.000	----	----
44	Administrative	0.023	0.368	2.349	L
45	Administrative	0.000	----	----	----
46	Administrative	0.085	0.486	1.696	L
47	Administrative	0.214	1.000	0.000	----
48	Administrative	0.477	1.781	1.356	S
49	Administrative	**7.450******	**5.318**	**3.927**	**S**
50	Administrative	**39.157******	**0.020**	**9.193**	**L**
51	Administrative	0.016	0.000	----	----
52	Administrative	0.211	1.530	0.999	S
53	Administrative	**1.331***	**0.464**	**1.804**	**L**
54	Administrative	0.321	0.479	1.730	L
55	Administrative	0.000	1.436	0.850	S
56	Administrative	**13.859******	**14.492**	**6.283**	**S**
57	Administrative	0.212	0.727	0.749	L
58	Administrative	0.010	1.410	0.807	S
59	Administrative	0.005	1.475	0.913	S
60	Administrative	0.418	0.507	1.596	L
61	Administrative	0.036	1.608	1.116	S
62	Administrative	0.289	1.544	1.021	S
63	Administrative	0.164	0.689	0.875	L
64	Administrative	**8.861******	**0.186**	**3.915**	**L**
65	Administrative	0.216	2.640	2.281	S
66	Administrative	0.225	0.335	2.570	L
67	Administrative	0.016	0.843	0.401	L
68	Administrative	0.002	1.245	0.515	S
69	Administrative	2.622*	0.000	----	----
70	Administrative	0.753	0.493	1.662	L
71	Administrative	0.128	1.602	1.107	S
72	Administrative	**5.602*****	**0.056**	**6.774**	**L**
73	Administrative	0.004	0.866	0.338	L
74	Administrative	**3.982*****	**0.331**	**2.598**	**L**
75	Administrative	**8.056******	**10.694**	**5.569**	**S**
76	Administrative	0.529	5.230	3.888	S
77	Administrative	0.002	0.850	0.382	L
78	Administrative	**3.237****	**8.926**	**5.144**	**S**
79	Administrative	0.200	0.703	0.828	L
80	Administrative	1.960*	0.000	----	----
81	Administrative	**2.620***	**0.346**	**2.494**	**L**
82	Administrative	1.212	4.434	3.500	S

83	Administrative	----	----	----	----
84	Administrative	**2.902****	**0.124**	**4.906**	**L**
85	Administrative	0.644	0.343	2.515	L
86	Administrative	0.130	1.740	1.302	S
87	Administrative	0.027	0.763	0.636	L
88	Administrative	0.024	1.439	0.855	S
89	Administrative	0.001	1.186	0.401	S
90	Administrative	0.485	----	----	----
91	Administrative	16.067****	----	----	----
92	Administrative	**1.857***	**0.271**	**3.068**	**L**
93	Administrative	0.000	----	----	----
94	Administrative	0.005	1.085	0.192	S
95	Feedback	0.485	0.000	----	----
96	Feedback	0.115	1.360	0.726	S
97	Feedback	1.028	8.867	5.128	S
98	Feedback	1.051	0.478	1.735	L
99	Feedback	**10.334******	**0.106**	**5.274**	**L**
100	Feedback	0.000	1.138	0.304	S
101	Feedback	**2.871****	**2.755**	**2.382**	**S**
102	Feedback	**3.075****	**0.222**	**2.640**	**S**
103	Feedback	0.372	0.000	----	----
104	Feedback	8.063****	0.000	----	----
105	Feedback	0.047	0.854	0.396	L
106	Feedback	1.311	0.440	1.929	L
107	Feedback	0.163	0.981	0.045	L
108	Feedback	0.308	0.456	1.845	L
109	Feedback	0.019	0.796	0.536	L
110	Feedback	0.000	1.667	1.201	S
111	Feedback	0.025	0.760	0.645	L
112	Feedback	0.015	1.313	0.640	S
113	Feedback	0.108	0.622	1.116	L
114	Feedback	**7.482******	**0.163**	**4.263**	**L**
115	Feedback	0.399	2.731	2.361	S
116	Feedback	**4.173*****	**0.177**	**4.069**	**L**
117	Feedback	0.030	1.534	1.006	S
118	Feedback	0.639	1.905	1.515	S
119	Feedback	**4.252*****	**3.543**	**2.973**	**S**
120	Feedback	**1.448***	**4.848**	**3.710**	**S**
121	Feedback	**18.549******	**0.062**	**6.534**	**L**
122	Feedback	**3.876*****	**0.209**	**3.679**	**L**
123	Feedback	0.014	0.552	1.396	L
124	Feedback	0.001	1.154	0.337	S
125	Feedback	1.065	0.348	2.481	L
126	Feedback	0.742	2.743	2.371	S

127	Feedback	0.035	1.456	0.883	S
128	Feedback	0.117	1.048	0.110	S
129	Feedback	0.213	0.000	----	----
130	Feedback	0.981	----	----	----
131	Feedback	0.010	1.126	0.279	S
132	Feedback	0.485	0.000	----	----
133	Feedback	**4.922*****	**0.191**	**3.890**	**L**
134	Feedback	**2.156***	**3.214**	**2.744**	**S**
135	Feedback	0.380	3.500	2.944	S
136	Feedback	0.013	1.129	0.285	S
137	Feedback	0.232	1.963	1.585	S
138	Feedback	0.521	0.423	2.022	L
139	Feedback	0.097	0.594	1.224	L
140	Feedback	1.476*	----	----	----
141	Feedback	2.250	0.368	2.349	L
142	Feedback	7.525	6.562	4.421	S
143	Feedback	0.204	0.392	2.201	L
144	Feedback	**1.215***	**0.375**	**2.305**	**L**
145	Feedback	**24.993******	**0.026**	**8.577**	**L**
146	Feedback	1.476*	----	----	----
147	Feedback	0.013	1.209	0.446	S
148	Feedback	**7.725******	**0.056**	**6.774**	**L**
149	Feedback	**11.125******	**0.115**	**5.083**	**L**
150	Feedback	0.070	3.018	2.596	S
151	Feedback	0.174	----	----	----
152	Feedback	0.178	3.400	2.876	S
153	Feedback	0.295	0.432	1.972	L
154	Feedback	**1.592***	**0.381**	**2.268**	**L**
155	Feedback	0.485	0.000	----	----
156	Feedback	**1.924***	**0.331**	**2.598**	**L**
157	Feedback	0.485	0.000	----	----
158	Feedback	0.305	0.650	1.012	L
159	Feedback	0.070	4.427	3.496	S
160	Feedback	0.020	0.809	0.498	L
161	Feedback	0.119	1.416	0.817	S
162	Feedback	0.589	2.090	1.732	S
163	Feedback	**2.450***	**0.452**	**1.866**	**L**
164	Feedback	**7.308******	**0.154**	**4.396**	**L**
165	Feedback	0.648	0.537	1.461	L
166	Feedback	0.051	1.221	0.469	S
167	Feedback	0.016	0.814	0.484	L
168	Feedback	**1.490***	**2.722**	**2.353**	**S**
169	Feedback	**14.490******	**11.638**	**5.768**	**S**
170	Feedback	6.040***	0.000	----	----

171	Feedback	2.385*	----	----	----
172	Feedback	**1.594***	**7.244**	**4.653**	**S**
173	Feedback	**3.798****	**0.113**	**5.124**	**L**
174	Feedback	0.125	1.472	0.909	S
175	Feedback	0.024	2.200	1.853	S
176	Feedback	0.487	1.714	1.266	S
177	Feedback	0.088	1.323	0.658	S
178	Feedback	0.032	0.560	1.363	S
179	Feedback	0.009	0.894	0.263	L
180	Feedback	0.061	0.891	0.255	L
181	Feedback	**7.150******	**7.301**	**4.672**	**S**
182	Feedback	0.865	0.388	2.225	L
183	Feedback	1.833	0.441	1.924	L
184	Feedback	----	----	----	---
185	Feedback	0.284	0.523	1.523	L
186	Feedback	0.419	0.591	1.236	L
187	Feedback	0.000	1.301	0.618	S
188	Feedback	0.007	0.738	0.714	L
189	Feedback	----	----	----	----
190	Department	0.000	0.861	0.351	L
191	Department	0.023	0.816	0.478	L
192	Department	0.527	1.886	1.491	S
193	Department	**5.191*****	**0.268**	**3.094**	**L**
194	Department	0.485	0.000	----	----
195	Department	**1.845***	**0.144**	**4.620**	**L**
196	Department	3.354**	0.000	**----**	**----**
197	Department	1.139	0.497	1.643	L
198	Department	**7.200******	**0.182**	**4.003**	**L**
199	Department	0.266	0.645	1.030	L
200	Department	0.068	1.427	0.836	S
201	Department	**40.960******	**0.005**	**12.451**	**L**
202	Department	0.174	1.642	1.165	S
203	Department	0.488	0.000	----	----
204	Department	0.016	1.750	1.315	S
205	Department	**5.286*****	**4.965**	**3.766**	**S**
206	Department	0.495	0.595	1.220	L
207	Department	**5.399*****	**0.085**	**5.792**	**L**
208	Department	0.000	1.370	0.740	S
209	Department	**5.494*****	**0.091**	**5.633**	**L**
210	Department	**11.177******	**0.067**	**6.352**	**L**
211	Department	1.282	0.509	1.587	L
212	Department	**2.471***	**0.298**	**2.845**	**L**
213	Department	0.214	0.364	2.375	L
214	Department	1.439	0.216	3.601	L

215	Department	1.211	0.294	2.877	L
216	Department	**28.328******	**0.013**	**10.206**	**L**
217	Department	0.422	1.793	1.372	S
218	Department	**4.075*****	**0.247**	**3.286**	**L**
219	Department	**4.279*****	**3.703**	**3.076**	**S**
220	Department	**1.677***	**0.441**	**1.924**	**L**
221	Department	**3.030****	**0.114**	**5.103**	**L**
222	Department	1.233	0.435	1.956	L
223	Department	0.032	1.229	0.485	S
224	Department	**4.896*****	**5.325**	**3.930**	**S**
225	Department	0.076	1.240	0.506	L
226	Department	0.135	0.653	1.002	L
227	Department	0.000	1.195	0.419	S
228	Department	1.393	1.924	1.538	S
229	Department	0.699	0.544	1.431	L
230	Department	0.031	1.374	0.747	S
231	Department	1.114	2.120	1.766	S
232	Department	0.856	2.094	1.737	S
233	Department	33.185****	0.000	----	----
234	Department	0.005	0.776	0.596	L
235	Department	0.102	1.296	0.609	S
236	Department	0.006	1.159	0.347	S
237	Department	0.035	0.930	0.171	L
238	Department	0.382	2.297	1.954	S
239	Department	0.757	0.395	2.183	L
240	Department	1.228	2.060	1.698	S
241	Department	**3.078****	**0.161**	**4.292**	**L**
242	Department	0.091	1.834	1.425	S
243	Department	**4.413*****	**0.274**	**3.042**	**L**
244	Department	0.213	1.267	0.556	S
245	Department	0.302	3.818	3.148	S
246	Department	1.344	2.451	2.107	S
247	Department	0.009	0.923	0.188	L
248	Department	0.009	0.000	----	----
249	Department	**20.015******	**0.054**	**6.859**	**L**
250	Department	**6.675*****	**0.166**	**4.220**	**L**
251	Department	0.050	0.806	0.507	L
252	Department	0.000	----	----	----
253	Department	0.002	0.719	0.775	L
254	Department	0.179	1.465	0.897	S
255	Department	0.070	0.962	0.091	L
256	Department	**2.427***	**3.252**	**2.771**	**S**
257	Department	**1.815***	**2.959**	**2.549**	**S**
258	Department	0.054	0.987	0.031	L

259	Department	0.013	1.125	0.277	S
260	Department	0.081	1.042	0.097	S
261	Department	0.051	1.153	0.335	S
262	Department	0.013	1.100	0.224	S
263	Department	0.130	1.396	0.784	S
264	Department	1.561	----	----	----
265	Department	0.006	1.141	0.310	S
266	Department	**2.295***	**2.586**	**2.233**	**S**
267	Department	0.501	0.618	1.131	L
268	Department	**4.893*****	**0.252**	**3.239**	**L**
269	Department	**5.743*****	**0.127**	**4.849**	**L**
270	Department	0.741	1.768	1.339	S
271	Department	----	----	----	----
272	Department	0.051	0.645	1.030	L
273	Department	**11.036*******	**0.153**	**4.412**	**L**
274	Department	0.033	0.950	0.121	L
275	Department	**7.472*******	**10.480**	**5.521**	**S**
276	Department	0.068	1.498	0.950	S
277	Department	0.000	1.467	0.901	S
278	Department	**8.251*******	**7.918**	**4.862**	**S**
279	Department	0.030	1.188	0.405	S
280	Department	0.040	1.286	0.591	S

Note: χ^2_{MH} is the Mantel-Haenszel chi-square statistic, α is the odds ratio from the MH chi-square, L=Leniency, S=Severity

**** *p* <. 01, *** *p* <. 05, ** *p* <. 10, * *p* < .20

Appendix M

Results of the Mean Score Method for Detecting Leniency/Severity

Rater	Condition	Instructor	t-test	Directionality
1	Administrative	3.02	**-3.092****	**S**
2	Administrative	2.98	**-9.170****	**S**
3	Administrative	3.02	-1.831	
4	Administrative	3.53	**3.823****	**L**
5	Administrative	3.05	**-2.509***	**S**
6	Administrative	3.07	**-2.334***	**S**
7	Administrative	2.53	**-5.025****	**S**
8	Administrative	3.09	-1.042	
9	Administrative	3.26	0.801	
10	Administrative	3.63	**4.156****	**L**
11	Administrative	3.14	-0.708	
12	Administrative	3.17	-0.243	
13	Administrative	3.51	**3.340****	**L**
14	Administrative	3.65	**4.936****	**L**
15	Administrative	2.60	**-5.270****	**S**
16	Administrative	3.16	-0.478	
17	Administrative	3.44	**3.015****	**L**
18	Administrative	3.35	1.281	
19	Administrative	2.88	**-4.424****	**S**
20	Administrative	3.35	1.601	
21	Administrative	3.40	**2.490***	**L**
22	Administrative	3.53	**3.823****	**L**
23	Administrative	3.37	1.932	
24	Administrative	3.19	-0.031	
25	Administrative	3.81	**9.089****	**L**
26	Administrative	2.77	**-5.256****	**S**
27	Administrative	3.07	**-2.334***	**S**
28	Administrative	3.58	**4.711****	**L**
29	Administrative	3.09	-1.492	
30	Administrative	3.42	**2.752****	**L**
31	Administrative	3.33	1.875	
32	Administrative	2.88	**-6.192****	**S**
33	Administrative	3.49	**3.552****	**L**
34	Administrative	2.98	**-2.342***	**S**
35	Administrative	3.60	**5.029****	**L**
36	Administrative	1.95	**-9.609****	**S**
37	Administrative	3.33	1.703	
38	Administrative	3.79	**9.569****	**L**
39	Administrative	3.84	**11.362****	**L**

40	Administrative	3.02	**-4.107****	**S**
41	Administrative	3.81	**7.496****	**L**
42	Administrative	3.33	1.380	
43	Administrative	4.00	-----**	**L**
44	Administrative	2.84	**-4.027****	**S**
45	Administrative	2.58	**-7.326****	**S**
46	Administrative	3.95	**23.496****	**L**
47	Administrative	2.88	**-3.033****	**S**
48	Administrative	3.65	**5.713****	**L**
49	Administrative	3.67	**6.700****	**L**
50	Administrative	3.84	**11.362****	**L**
51	Administrative	2.60	**-6.165****	**S**
52	Administrative	3.44	**2.801****	**L**
53	Administrative	3.40	**2.722****	**L**
54	Administrative	3.12	-0.825	
55	Administrative	2.79	**-3.529****	**S**
56	Administrative	3.35	1.968	
57	Administrative	3.56	**3.627****	**L**
58	Administrative	3.07	-1.429	
59	Administrative	3.05	-1.767	
60	Administrative	3.77	**7.182****	**L**
61	Administrative	3.14	-0.801	
62	Administrative	3.63	**5.362****	**L**
63	Administrative	3.47	**3.281****	**L**
64	Administrative	3.67	**5.239****	**L**
65	Administrative	2.77	**-4.856****	**S**
66	Administrative	3.09	-1.737	
67	Administrative	3.44	**2.801****	**L**
68	Administrative	3.07	-1.280	
69	Administrative	2.83	**-2.994****	**S**
70	Administrative	3.49	**3.868****	**L**
71	Administrative	3.77	**7.890****	**L**
72	Administrative	3.93	**10.610****	**L**
73	Administrative	3.84	**11.362****	**L**
74	Administrative	3.53	**3.587****	**L**
75	Administrative	2.84	**-3.364****	**S**
76	Administrative	3.79	**5.094****	**L**
77	Administrative	3.60	**5.496****	**L**
78	Administrative	2.70	**-5.064****	**S**
79	Administrative	3.70	**6.484****	**L**
80	Administrative	4.00	----**	**L**
81	Administrative	3.42	**3.003****	**L**
82	Administrative	2.95	**-3.574****	**S**
83	Administrative	2.91	**-6.315****	**S**

84	Administrative	3.97	**33.830****	**L**
85	Administrative	3.84	**9.810****	**L**
86	Administrative	3.81	**10.391****	**L**
87	Administrative	3.91	**12.844****	**L**
88	Administrative	3.09	-1.116	
89	Administrative	3.09	-0.757	
90	Administrative	2.79	**-6.361****	**S**
91	Administrative	2.91	**-6.315****	**S**
92	Administrative	3.95	**23.496****	**L**
93	Administrative	2.77	**-5.773****	**S**
94	Administrative	3.51	**3.340****	**L**
95	Feedback	2.88	**-5.138****	**S**
96	Feedback	3.33	1.703	
97	Feedback	2.81	**-5.478****	**S**
98	Feedback	3.35	1.700	
99	Feedback	3.51	**4.170****	**L**
100	Feedback	3.47	**2.182***	**L**
101	Feedback	3.44	**2.627***	**L**
102	Feedback	3.93	**18.831****	**L**
103	Feedback	2.81	**-4.929****	**S**
104	Feedback	3.30	1.435	
105	Feedback	3.09	-1.492	
106	Feedback	3.40	**2.490***	**L**
107	Feedback	3.12	-0.889	
108	Feedback	2.95	**-2.375***	**S**
109	Feedback	3.16	-0.274	
110	Feedback	3.92	**18.831****	**L**
111	Feedback	3.33	1.572	
112	Feedback	3.19	-0.066	
113	Feedback	2.63	**-3.377****	**S**
114	Feedback	3.42	**2.752****	**L**
115	Feedback	2.91	**-3.527****	**S**
116	Feedback	3.33	1.703	
117	Feedback	2.86	**-3.044****	**S**
118	Feedback	3.44	**2.358***	**L**
119	Feedback	3.35	1.968	
120	Feedback	2.63	**-4.681****	**S**
121	Feedback	3.56	**4.407****	**L**
122	Feedback	3.09	-0.886	
123	Feedback	2.93	**-3.723****	**S**
124	Feedback	2.77	**-2.782****	**S**
125	Feedback	3.33	1.703	
126	Feedback	3.14	-0.801	
127	Feedback	3.19	-0.044	

128	Feedback	3.09	-1.116	
129	Feedback	2.93	**-3.303****	**S**
130	Feedback	2.67	**-7.131****	**S**
131	Feedback	3.56	**4.407****	**L**
132	Feedback	3.02	**-7.170****	**S**
133	Feedback	3.36	1.754	
134	Feedback	3.61	**4.034****	**L**
135	Feedback	2.77	**-5.256****	**S**
136	Feedback	3.77	**7.182****	**L**
137	Feedback	3.05	-1.441	
138	Feedback	3.35	1.968	
139	Feedback	3.14	-0.503	
140	Feedback	2.79	**-6.361****	**S**
141	Feedback	3.44	**2.481***	**L**
142	Feedback	3.37	1.932	
143	Feedback	3.05	-1.635	
144	Feedback	3.67	**3.794****	**L**
145	Feedback	3.65	**5.713****	**L**
146	Feedback	2.74	**-6.622****	**S**
147	Feedback	3.05	-1.529	
148	Feedback	3.37	**2.230***	**L**
149	Feedback	3.70	**4.309****	**L**
150	Feedback	2.88	**-4.486****	**S**
151	Feedback	2.93	**-6.608****	**S**
152	Feedback	3.02	**-4.107****	**S**
153	Feedback	3.00	**-2.158***	**S**
154	Feedback	3.42	**2.752****	**L**
155	Feedback	2.88	**-5.138****	**S**
156	Feedback	3.91	**15.997****	**L**
157	Feedback	2.88	**-5.138****	**S**
158	Feedback	3.58	**5.142****	**L**
159	Feedback	2.44	**-5.945****	**S**
160	Feedback	3.56	**3.839****	**L**
161	Feedback	3.09	-0.814	
162	Feedback	3.37	1.367	
163	Feedback	3.79	**7.049****	**L**
164	Feedback	3.27	0.543	
165	Feedback	3.90	**15.591****	**L**
166	Feedback	2.84	**-3.207****	**S**
167	Feedback	3.37	**2.230***	**L**
168	Feedback	3.23	0.529	
169	Feedback	3.37	**2.230***	**L**
170	Feedback	4.00	-----**	**L**
171	Feedback	2.86	**-6.163****	**S**

172	Feedback	2.72	**-5.208****	**S**
173	Feedback	3.19	-0.066	
174	Feedback	3.35	1.700	
175	Feedback	2.16	**-6.914****	**S**
176	Feedback	3.12	**2.261***	**L**
177	Feedback	3.63	**5.871****	**L**
178	Feedback	3.16	-0.412	
179	Feedback	3.47	**3.050****	**L**
180	Feedback	3.00	**-2.158***	**S**
181	Feedback	3.14	-0.944	
182	Feedback	3.21	0.211	
183	Feedback	3.40	**2.041***	**L**
184	Feedback	2.93	**-6.608****	**S**
185	Feedback	3.16	-0.290	
186	Feedback	3.53	**4.481****	**L**
187	Feedback	3.09	**-2.164***	**S**
188	Feedback	3.07	-1.555	
189	Feedback	2.91	**-6.315****	**S**
190	Department	3.52	**3.560****	**L**
191	Department	3.33	1.703	
192	Department	3.65	**6.271****	**L**
193	Department	3.58	**5.142****	**L**
194	Department	2.79	**-5.620****	**S**
195	Department	2.93	**-3.087****	**S**
196	Department	2.70	**-5.064****	**S**
197	Department	2.81	**-2.545***	**S**
198	Department	3.12	-0.519	
199	Department	3.26	0.743	
200	Department	3.81	**10.391****	**L**
201	Department	3.84	**11.362****	**L**
202	Department	3.30	0.953	
203	Department	2.88	**-5.138****	**S**
204	Department	2.91	**-3.527****	**S**
205	Department	3.53	**3.390****	**L**
206	Department	3.60	**4.367****	**L**
207	Department	3.19	-0.041	
208	Department	3.14	-0.642	
209	Department	3.35	**2.160***	**L**
210	Department	3.37	1.516	
211	Department	3.49	**3.302****	**L**
212	Department	3.93	**18.831****	**L**
213	Department	2.91	**-2.865****	**S**
214	Department	3.63	**14.372****	**L**
215	Department	3.23	0.653	

216	Department	3.67	**6.700****	**L**
217	Department	3.65	**4.649****	**L**
218	Department	3.60	**4.122****	**L**
219	Department	3.53	**3.587****	**L**
220	Department	3.40	**2.490***	**L**
221	Department	2.95	**-2.147***	**S**
222	Department	3.86	**12.540****	**L**
223	Department	3.00	**-2.158***	**S**
224	Department	3.70	**7.164****	**L**
225	Department	2.95	**-2.695***	**S**
226	Department	3.21	0.226	
227	Department	3.37	**2.441***	**L**
228	Department	3.63	**5.871****	**L**
229	Department	3.72	**6.344****	**L**
230	Department	3.30	1.435	
231	Department	3.44	**2.801****	**L**
232	Department	3.44	**2.801****	**L**
233	Department	3.58	**3.363****	**L**
234	Department	2.84	**-4.027****	**S**
235	Department	2.93	**-3.723****	**S**
236	Department	3.51	**2.866****	**L**
237	Department	3.49	**3.868****	**L**
238	Department	3.15	-0.781	
239	Department	3.21	0.188	
240	Department	3.23	0.489	
241	Department	3.91	**15.997****	**L**
242	Department	2.93	**-2.821****	**S**
243	Department	3.40	1.938	
244	Department	2.84	**-4.353****	**S**
245	Department	3.86	**6.878****	**L**
246	Department	3.14	-0.708	
247	Department	3.51	**3.340****	**L**
248	Department	3.00	**-3.562****	**S**
249	Department	3.88	**14.025****	**L**
250	Department	3.91	**15.997****	**L**
251	Department	3.47	**3.575****	**L**
252	Department	2.86	**-6.163****	**S**
253	Department	2.26	**-4.955****	**S**
254	Department	3.58	**5.142****	**L**
255	Department	3.86	**12.540****	**L**
256	Department	3.58	**3.871****	**L**
257	Department	3.02	-1.831	
258	Department	3.49	**3.868****	**L**
259	Department	3.33	1.875	

260	Department	3.70	**4.700****	**L**
261	Department	3.19	-0.041	
262	Department	3.19	-0.035	
263	Department	3.47	**3.575****	**L**
264	Department	2.86	**-6.163****	**S**
265	Department	3.56	0.801	
266	Department	3.72	**6.913****	**L**
267	Department	3.63	**5.871****	**L**
268	Department	3.49	**3.302****	**L**
269	Department	3.37	**2.441***	**L**
270	Department	3.56	**4.094****	**L**
271	Department	2.40	**-8.370****	**S**
272	Department	3.28	1.064	
273	Department	3.74	**6.253****	**L**
274	Department	3.56	**4.804****	**L**
275	Department	3.02	**-2.365***	**S**
276	Department	3.72	**5.529****	**L**
277	Department	2.95	**-2.695***	**S**
278	Department	3.19	-0.066	
279	Department	3.16	-0.412	
280	Department	3.63	**5.871****	**L**

Note: L=Leniency, S=Severity; ---- refers to raters whose mean rating was a 4 on a 4 point scale. A t-statistic could not be calculated because there was no variance in the scores. These values are indicated as being statistically significant at *p* < .01.
** *p* < .01
* *p* < .05

Appendix N

Skewness Ratings

Rater	Condition	Skewness	Directionality
1	Administrative	-0.583	L
2	Administrative	**-9.110***	**L**
3	Administrative	**-1.026***	**L**
4	Administrative	-0.664	L
5	Administrative	-0.522	L
6	Administrative	-0.522	L
7	Administrative	-0.122	L
8	Administrative	**-1.004***	**L**
9	Administrative	-0.358	L
10	Administrative	**-1.615***	**L**
11	Administrative	0.000	
12	Administrative	-0.141	L
13	Administrative	**-0.714***	**L**
14	Administrative	**-1.385***	**L**
15	Administrative	0.032	S
16	Administrative	0.522	S
17	Administrative	0.251	S
18	Administrative	**-0.727***	**L**
19	Administrative	**-0.772***	**L**
20	Administrative	-0.309	L
21	Administrative	-0.326	L
22	Administrative	-0.679	L
23	Administrative	-0.100	L
24	Administrative	-0.595	L
25	Administrative	**-1.696***	**L**
26	Administrative	0.526	S
27	Administrative	**0.777***	**S**
28	Administrative	-0.060	L
29	Administrative	-0.073	L
30	Administrative	-0.210	L
31	Administrative	0.138	S
32	Administrative	**-2.059***	**L**
33	Administrative	-0.221	L
34	Administrative	-0.187	L
35	Administrative	**-1.110***	**L**
36	Administrative	-0.134	L
37	Administrative	**-0.853***	**L**
38	Administrative	**-1.830***	**L**
39	Administrative	**-2.905***	**L**

40	Administrative	**-0.735***	**L**
41	Administrative	**-4.150***	**L**
42	Administrative	**-0.811***	**L**
43	Administrative	**-6.068***	**L**
44	Administrative	-0.011	L
45	Administrative	-0.609	L
46	Administrative	**-3.047***	**L**
47	Administrative	-0.693	L
48	Administrative	**-1.717***	**L**
49	Administrative	**-1.157***	**L**
50	Administrative	-0.446	L
51	Administrative	-0.507	L
52	Administrative	-0.551	L
53	Administrative	0.501	S
54	Administrative	0.021	S
55	Administrative	0.193	S
56	Administrative	-0.394	L
57	Administrative	**-1.060***	**L**
58	Administrative	0.060	S
59	Administrative	-0.453	L
60	Administrative	**-1.442***	**L**
61	Administrative	**0.765***	**S**
62	Administrative	**-0.895***	**L**
63	Administrative	-0.287	L
64	Administrative	-0.692	L
65	Administrative	-0.138	L
66	Administrative	**0.711***	**S**
67	Administrative	**-0.872***	**L**
68	Administrative	-0.030	L
69	Administrative	-0.140	L
70	Administrative	0.040	S
71	Administrative	**-1.880***	**L**
72	Administrative	**-2.994***	**L**
73	Administrative	**-2.577***	**L**
74	Administrative	-0.429	L
75	Administrative	-0.318	L
76	Administrative	**-4.828***	**L**
77	Administrative	-0.123	L
78	Administrative	-0.175	L
79	Administrative	**-1.875***	**L**
80	Administrative	**-4.004***	**L**
81	Administrative	0.082	S
82	Administrative	-0.122	L
83	Administrative	**-2.685***	**L**

84	Administrative	**-2.565***	**L**
85	Administrative	**-2.308***	**L**
86	Administrative	**-1.981***	**L**
87	Administrative	**-3.401***	**L**
88	Administrative	-0.574	L
89	Administrative	-0.377	L
90	Administrative	**-0.897***	**L**
91	Administrative	0.076	S
92	Administrative	**-2.208***	**L**
93	Administrative	**-1.620***	**L**
94	Administrative	-0.823	L
95	Feedback	**-1.647***	**L**
96	Feedback	0.399	S
97	Feedback	-0.424	L
98	Feedback	**-0.780***	**L**
99	Feedback	0.209	S
100	Feedback	**-1.612***	**L**
101	Feedback	**-0.933***	**L**
102	Feedback	**-1.981***	**L**
103	Feedback	-0.452	L
104	Feedback	0.368	S
105	Feedback	**-1.272***	**L**
106	Feedback	0.047	S
107	Feedback	-0.619	L
108	Feedback	-0.290	L
109	Feedback	-0.420	L
110	Feedback	**-3.670***	**L**
111	Feedback	0.294	S
112	Feedback	0.267	S
113	Feedback	-0.239	L
114	Feedback	-0.025	L
115	Feedback	**-0.852***	**L**
116	Feedback	-0.016	L
117	Feedback	-0.227	L
118	Feedback	**-0.729***	**L**
119	Feedback	-0.488	L
120	Feedback	-0.087	L
121	Feedback	0.502	S
122	Feedback	-0.019	L
123	Feedback	**-1.168***	**L**
124	Feedback	-0.395	L
125	Feedback	0.185	S
126	Feedback	0.510	S
127	Feedback	-0.394	L

128	Feedback	-0.474	L
129	Feedback	-0.504	L
130	Feedback	**-1.155***	**L**
131	Feedback	-0.440	L
132	Feedback	**-2.089***	**L**
133	Feedback	-0.076	L
134	Feedback	**-1.792***	**L**
135	Feedback	-0.354	L
136	Feedback	**-2.308***	**L**
137	Feedback	-0.649	L
138	Feedback	-0.049	L
139	Feedback	**-0.742***	**L**
140	Feedback	**-0.827***	**L**
141	Feedback	-0.344	L
142	Feedback	-0.658	L
143	Feedback	-0.044	L
144	Feedback	**-1.318***	**L**
145	Feedback	-0.172	L
146	Feedback	**-1.025***	**L**
147	Feedback	-0.186	L
148	Feedback	0.067	S
149	Feedback	**-1.565***	**L**
150	Feedback	-0.484	L
151	Feedback	**-1.835***	**L**
152	Feedback	**2.115***	**S**
153	Feedback	-0.330	L
154	Feedback	0.135	S
155	Feedback	**-1.025***	**L**
156	Feedback	**-1.981***	**L**
157	Feedback	**-1.372***	**L**
158	Feedback	-0.488	L
159	Feedback	-0.482	L
160	Feedback	**-1.112***	**L**
161	Feedback	-0.672	L
162	Feedback	**-1.082***	**L**
163	Feedback	**-1.992***	**L**
164	Feedback	-0.366	L
165	Feedback	**-2.197***	**L**
166	Feedback	**-0.774***	**L**
167	Feedback	-0.051	L
168	Feedback	-0.437	L
169	Feedback	-0.628	L
170	Feedback	**-3.046***	**L**
171	Feedback	**-1.322***	**L**

172	Feedback	-0.179	L
173	Feedback	**2.565***	**S**
174	Feedback	-0.538	L
175	Feedback	0.573	S
176	Feedback	**-1.031***	**L**
177	Feedback	-0.533	L
178	Feedback	**0.809***	**S**
179	Feedback	-0.336	L
180	Feedback	-0.330	L
181	Feedback	0.545	S
182	Feedback	-0.002	L
183	Feedback	-0.198	L
184	Feedback	**-3.364***	**L**
185	Feedback	0.016	S
186	Feedback	**-0.702***	**L**
187	Feedback	0.522	S
188	Feedback	-0.058	L
189	Feedback	**-2.375***	**L**
190	Department	**-0.931***	**L**
191	Department	0.082	S
192	Department	**-1.271***	**L**
193	Department	-0.140	L
194	Department	**-1.162***	**L**
195	Department	0.087	S
196	Department	-0.315	L
197	Department	-0.135	L
198	Department	-0.393	L
199	Department	0.018	S
200	Department	**-1.947***	**L**
201	Department	**-1.105***	**L**
202	Department	-0.665	L
203	Department	**-1.025***	**L**
204	Department	-0.162	L
205	Department	**-1.882***	**L**
206	Department	**-1.310***	**L**
207	Department	0.063	S
208	Department	-0.435	L
209	Department	0.455	S
210	Department	-0.622	L
211	Department	-0.677	L
212	Department	**-2.516***	**L**
213	Department	-0.067	L
214	Department	**-2.760***	**L**
215	Department	**1.389***	**S**

216	Department	-0.003	L
217	Department	**-1.630***	**L**
218	Department	**-0.905***	**L**
219	Department	**-1.493***	**L**
220	Department	**-0.780***	**L**
221	Department	0.258	S
222	Department	**-1.593***	**L**
223	Department	-0.664	L
224	Department	**-1.439***	**L**
225	Department	**-1.456***	**L**
226	Department	0.025	S
227	Department	0.555	S
228	Department	-0.643	L
229	Department	**-1.821***	**L**
230	Department	-0.040	L
231	Department	-0.594	L
232	Department	**-0.737***	**L**
233	Department	**-0.890***	**L**
234	Department	-0.025	L
235	Department	-0.326	L
236	Department	**-1.372***	**L**
237	Department	-0.647	L
238	Department	0.509	S
239	Department	0.014	S
240	Department	-0.135	L
241	Department	**-2.471***	**L**
242	Department	-0.689	L
243	Department	-0.232	L
244	Department	**-0.733***	**L**
245	Department	**-5.908***	**L**
246	Department	0.071	S
247	Department	**-0.986***	**L**
248	Department	**-0.932***	**L**
249	Department	**-0.823***	**L**
250	Department	**-1.356***	**L**
251	Department	0.250	S
252	Department	**-1.162***	**L**
253	Department	0.489	S
254	Department	-0.380	L
255	Department	**-1.690***	**L**
256	Department	**-1.540***	**L**
257	Department	-0.031	L
258	Department	**-0.771***	**L**
259	Department	0.549	S

260	Department	**-2.008***	**L**
261	Department	**-0.855***	**L**
262	Department	-0.646	L
263	Department	0.323	S
264	Department	**-1.242***	**L**
265	Department	**-1.083***	**L**
266	Department	**-1.775***	**L**
267	Department	-0.680	L
268	Department	-0.261	L
269	Department	0.274	S
270	Department	**-1.091***	**L**
271	Department	-0.464	L
272	Department	-0.015	L
273	Department	**-1.366***	**L**
274	Department	-0.193	L
275	Department	0.155	S
276	Department	**-2.068***	**L**
277	Department	**-0.739***	**L**
278	Department	0.251	S
279	Department	-0.469	L
280	Department	-0.627	L

Note: L=Leniency, S=Severity
*p <. 01

Appendix O

Correlations for Total Score with Each Goal Questionnaire Item

Item	Admin(N=94)	Feedback(N=94)	Dept(N=91)
1. Identify areas in which the instructor might need improvement.	.16	.13	-.21*
2. Rate my instructor fairly.	.38**	.13	.33**
3. Identify areas where the instructor needs more training.	.03	-.01	-.05
4. Convey my satisfaction with the instructor's performance.	.36**	.09	.47**
5. Identify area that the instructor should focus on improving.	.01	-.03	.01
6. Indicate where the instructor fell short in terms of performance.	-.01	.02	-.01
7. Give my instructor a rating that she or he will realize is based on performance, rather than my judgment of him/her as a person.	.40**	.10	.24*
8. Identify my instructor’s strengths and weaknesses.	.36**	.07	.12
9. Highlight my instructor’s performance so that his or her success is visible to his or her department head.	.44**	.16	.33**
10. Improve my instructor’s confidence.	.21*	.08	.30**
11. Make it clear to my instructor that there is room for improvement.	-.23*	-.14	-.16
12. Identify my instructor’s performance deficiencies.	-.07	-.02	-.25*
13. Challenge my instructor to improve his or her performance.	-.03	.13	-.19

14. Clarify expected performance levels to the instructor.	.23*	.05	.03
15. Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.	.24*	.08	.33**
16. Indicate where instructor has exceeded performance expectations.	.45**	.29**	.37**
17. Encourage the instructor's current level of performance.	.41**	.30**	.53**
18. Encourage the instructor to improve performance.	-.03	.12	-.10
19. Motivate the instructor.	.12	.19	.18

*$p < .05$

**$p < .01$

Note: Admin = Administrative condition, Feedback=Feedback condition, Dept=Psychology Department condition

Appendix P

Analysis of Variance Results for the Goal Importance Questionnaire Items.(N=279)

Item	F	η^2
1. Identify areas in which the instructor might need improvement.	1.43	.01
2. Rate my instructor fairly.	1.19	.01
3. Identify areas where the instructor needs more training.	0.12	.00
4. Convey my satisfaction with the instructor's performance.	2.48	.09
5. Identify area that the instructor should focus on improving.	0.90	.01
6. Indicate where the instructor fell short in terms of performance.	0.04	.00
7. Give my instructor a rating that she or he will realize is based on performance, rather than my judgment of him/her as a person.	1.05	.01
8. Identify my instructor’s strengths and weaknesses.	2.72	.02
9. Highlight my instructor’s performance so that his or her success is visible to his or her department head.	5.03**	.04
10. Improve my instructor’s confidence.	2.42	.02
11. Make it clear to my instructor that there is room for improvement.	0.90	.01
12. Identify my instructor’s performance deficiencies.	2.29	.02
13. Challenge my instructor to improve his or her performance.	0.48	.00

14. Clarify expected performance levels to the instructor.	0.29	.00
15. Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.	6.21**	.04
16. Indicate where instructor has exceeded performance expectations.	1.81	.01
17. Encourage the instructor's current level of performance.	3.02	.02
18. Encourage the instructor to improve performance.	1.11	.01
19. Motivate the instructor.	1.73	.01

*$p < .05$

**$p < .01$

Note: Admin = Administrative condition, Feedback=Feedback condition, Dept=Psychology Department condition

Appendix Q

Point-Biserial Correlations for Leniency for Each Method with Each Goal Questionnaire Item

	Admin(N=94)			Feedback(N=94)			Dept(N=91)			Total(N=279)		
Item	D	M	S	D	M	S	D	M	S	D	M	S
1. Identify areas in which the instructor might need improvement.	.09	.06	.13	.24	-.04	.23*	.04	.10	-.10	.04	-.04	.07
2. Rate my instructor fairly.	.22**	.19	.32**	-.01	-.06	.05	.05	.11	.20	.10	.25**	.20**
3. Identify areas where the instructor needs more training.	.16	.12	.06	.05	-.07	.15	.06	.12	.03	.09	-.07	.08
4. Convey my satisfaction with the instructor's performance.	.17	.10	.22*	-.08	-.01	.02	-.04	.01	.18	.02	.29**	.15*
5. Identify area that the instructor should focus on improving.	19	.19	-.04	-.07	-.17	.01	.08	.04	-.03	.07	-.10	-.01
6. Indicate where the instructor fell short in terms of performance.	.19	.20	-.07	-.11	-.07	.08	.07	.13	.01	.06	.07	.00
7. Give my instructor a rating that she or he will realize is based on performance, rather than my judgment of him/her as a person.	.15	.11	.37**	.04	-.03	.13	.01	-.04	.18	.07	.21*	.24**
8. Identify my instructor’s strengths and weaknesses.	.20	.07	.29**	-.01	.06	.06	.02	-.02	.03	.08	.14*	.14*

9. Highlight my instructor’s performance so that his or her success is visible to his or her department head.	.25**	.17	.32**	.12	.01	.10	.09	.07	.21	.16**	.30**	.22**
10. Improve my instructor’s confidence.	.04	.02	.06	.18	-.13	.03	.13	.17	.08	.12	.11	.06
11. Make it clear to my instructor that there is room for improvement.	.16	.05	-.15	.00	.04	.10	.07	.21*	-.05	.08	-.18*	-.04
12. Identify my instructor’s performance deficiencies.	.24**	.18	-.09	.11	.06	.09	-.01	.10	-.07	.11	-.15*	-.02
13. Challenge my instructor to improve his or her performance.	.24*	.08	-.10	.08	-.07	.17	.01	.04	-.11	.11	-.05	-.02
14. Clarify expected performance levels to the instructor.	.24*	.02	.10	-.01	-.02	.07	.04	.18	.02	.09	.05	.07
15. Evaluate the instructor in a manner that clearly indicates what was done well and what was done poorly.	.20	.19	.20	-.10	.01	-.07	-.04	.13	.15	.04	.10	.10
16. Indicate where instructor has exceeded performance expectations.	.20	.03	.26*	.02	-.01	.00	.09	.20	.16	.11	.23**	.12*
17. Encourage the instructor’s current level of performance.	.19	.10	.28**	.05	-.04	-.01	.14	.27**	.28*	.13*	.30**	.19**
18. Encourage the instructor to improve performance.	.05	.06	-.25**	-.03	-.02	.15	.00	.08	.07	.01	-.10	-.02
19. Motivate the instructor.	.06	-.10	.05	.04	-.06	.14	-.01	.08	.04	.03	.09	.08

*$p < .05$

**$p < .01$; Note: Admin = Administrative condition, Feedback=Feedback condition, Dept=Psychology Department condition, D=DPF method, M=Mean score method, S=Skewness score method