REVISITING RATING FORMAT RESEARCH: COMPUTER-BASED RATING FORMATS AND COMPONENTS OF ACCURACY

by
Scott Parrill

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Psychology

Neil Hauenstein, chair
Roseanne Foti
John Donovan

12 May, 1999
Blacksburg, VA

Keywords: Rating, Format, Appraisals, Accuracy

REVISITING RATING FORMAT RESEARCH: COMPUTER-BASED RATING FORMATS AND COMPONENTS OF ACCURACY

Scott Parrill

Abstract

Prior to 1980, most research on performance appraisal focused on rating formats. Since then, most performance appraisal research has focused on the internal processes of raters. This study redirects the focus back onto rating format with a critical eye towards rating accuracy. Ninety subjects read several hypothetical descriptions of teacher behavior and then rated the teachers on different dimensions of teaching performance using computer-based rating formats. It was found that rating format does affect some measures of rating accuracy. In addition, support was found for the viability of a new rating format. Graphic rating scales with no anchors received higher accuracy scores on certain measures of accuracy, higher ratings for liking of the rating format, higher levels of comfort with the rating format, and higher levels of interrater reliability than either BARS or graphic rating scales with numerical anchors. This study supports the ideas that rating format research should be reexamined with a focus on rating accuracy and that computer-based graphic scales with no anchors should be considered as an alternative to more traditional rating methods.

CONTENTS

TABLES

CHAPTER 1: INTRODUCTION

The idea of appraising or evaluating the performance of a worker is not an idea originating in modern organizations. The general concept of performance appraisal has historical precedents dating back, at least, hundreds of years. As early as the third century A.D., a Chinese philosopher criticized the practices of the "Imperial Rater" because he rated workers according to how well he liked them, not based on how well they performed their duties (Murphy & Cleveland, 1995).

Formal, merit-based systems have existed in America and abroad since the early 1800's. As early as 1916, at least one major department store chain utilized a formal appraisal process not dissimilar from those in use today (Benjamin, 1952). Gradually, the idea of appraising worker performance has increased in popularity, sometimes spurred on by the social zeitgeist. The popularity of the civil rights movement of the 1960's and 70's and the ensuing legislation created a need for the increased usage of valid appraisal practices (Murphy & Cleveland, 1995). Since then, the attention devoted to performance appraisal has continued to increase to a point where job performance is now the most frequently studied criterion variable in both the areas of human resource management and organizational behavior (Heneman, 1986). With the growing number of successful, multinational companies, performance appraisal is rapidly becoming a global, not just an American, topic. For example, DeVries, Morrison, Schullman, and Gerlach (1986) found that approximately 82% of the organizations in Great Britain have instituted a formal appraisal process of some kind.

Throughout the years, there has been a tremendous amount of research dedicated to the appraisal process and appraisal instruments. Graphic rating scales date back to the first quarter of this century (Paterson, 1922). They were widely viewed as a method to accurately and fairly assess an employee's performance. Over time, researchers became more aware of the strengths and the limitations of graphic rating scales. The scales were simple to use and develop, but the evaluations were often contaminated by various rating errors. The desire to improve the quality of performance ratings spurred research into other rating formats. Most popular was Smith & Kendall's (1963) behaviorally anchored rating scale (BARS). Again, as researchers systematically explored the BARS method, the literature soon revealed their strengths and weaknesses as well. Numerous studies were conducted to compare the different formats and the conditions in which they were administered. Many other, less popular, styles of rating scales were developed and compared with the BARS and graphic rating scales. These include behavioral observation scales (Tziner, Kopelman, & Livneh, 1993), behavioral expectation scales (Keaveny & McGann, 1975; Schneier, 1977), mixed-standard scales (Kingstrom & Bass, 1981), and quantitative ranking scales (Chiu & Alliger, 1990). Through it all, however, the BARS and the graphic rating scales received, by far, the most attention.

The accumulated body of rating format research examined every part of the appraisal process, the rater, the ratee, intervening variables, gender, age, race, and the format itself. However, with all of this research, there was no clear consensus as to what was the best rating format. In 1980, Landy & Farr called for a moratorium on format

research. Soon, the focus of researchers began to move to a more cognitive perspective of the rating process, and they began to focus on improving the rater rather than the format. Researchers devised cognitive models of the appraisal process (DeNisi, Cafferty, & Meglino, 1984), developed methods of training raters to yield better appraisals (Bernardin & Buckley, 1981; Latham, Wexley, and Pursell, 1975; Stamoulis & Hauenstein, 1993), and experimented with motivational processes that affect the appraisal process (Neck, Stewart, & Manz, 1995).

With the focus on cognitive aspects of appraisal, little research is presently directed towards rating format. However, there may be a good reason to revisit a line of research that has been deemed fruitless. Previous expeditions into the realm of format research have been based largely upon the notions of freedom from halo and leniency errors, high levels of reliability, and increased variance in performance ratings (Borman & Dunnette, 1975; Friedman & Cornelius, 1976; Kingstrom & Bass, 1981). However, much of the research currently published in the area of performance appraisal is concerned with levels of rating accuracy (Day & Sulsky, 1995; Murphy & Balzer, 1989; Stamoulis & Hauenstein, 1993). This focus on rating accuracy is noticeably absent from the rating format research of the past. Previous conclusions that rating format has little or no effect on the quality of ratings may have been drawn due to the erroneous beliefs regarding the importance of halo and leniency error, reliability, and variability in ratings. Research has since revealed that these criteria are not as important as measures of rating accuracy for determining the merits of a method of rating performance. Therefore, based on the new focus on rating format, a second look at rating format may be necessary.

To date, there has been little applied research attempting to combine the domains of format research and the cognitive approach to performance appraisal. Not only do we, as a field need to re-examine rating format research based on the notion of maximizing accuracy, but there is also a necessity to combine format research with the new focus on the cognitions that surround and are involved with performance appraisal.

The Nature of Appraisals

Performance appraisal can be thought of as "…the systematic description of job-relevant strengths and weaknesses within and between employees or groups…(Cascio, 1998, p.58)." Simply put, performance appraisal is a measurement of how well someone performs a job-relevant task. Performance appraisals (PAs), however, are not limited to formal evaluations administered in an organizational setting. Other examples of PAs include praise from a boss or co-worker for a job well done, grades given to students, and statistics of proficiency for athletes. However, given the prevalence of formal appraisal instruments, one typically associates appraisal with organizational settings.

Formal appraisals are an important and integral part of any organization. One important purpose for appraisal is a basis for employers to take disciplinary action such as denying a pay increase or justification of employee termination (Jacobs, 1986). Organizations can also use performance appraisals for determining employee's strengths and weaknesses (Cleveland, Murphy, & Williams, 1989). Perhaps the most obvious use of performance appraisal is to assist in the decisions regarding promotions and/or pay

raises (Murphy & Cleveland, 1995). In addition to those purposes mentioned above, performance appraisals serve a host of other functions including, but not limited to: determination of transfers and assignments, personnel planning, assisting in goal identification, reinforcing the authority structure, and identifying widespread organizational developmental needs.

Since appraisals are so pervasive in modern organizations, it is only prudent that researchers investigate all factors that affect the rating process. Organizations spend millions of dollars per year to rate their employees for a variety of reasons. Because of this, as much attention as possible should be devoted to every facet of the rating process. Numerous studies have focused on characteristics of the rater and characteristics of the ratee. A third area of concern is the vehicle by which ratings are made: the appraisal instrument. Landy & Farr (1980) note that there is a general consensus that the appraisal instrument is a very important part of the appraisal process, and different instruments can affect the accuracy and the utility of the performance evaluation information. This is the focus of concern for the rest of this paper.

Appraisal Instruments

People administering performance appraisal instruments, or raters, have the option of two general categories of performance measures (Cascio, 1998). Objective performance measures have the benefit of being easily quantified, objective measures relative to job performance. They may include production data (how many units were produced, how many errors were committed, the total dollar value of sales) and employment data (tardiness, absences, accidents). Although these measures appear to be desirable, they do not focus on the behavior of the employee and are often impractical and unsuitable for appraisal purposes (Heneman, 1986).

Subjective measures, on the other hand, attempt to directly measure a worker's behavior. However, since they depend on human judgements, they are vulnerable to a whole host of biases. Subjective measures include relative and absolute ranking systems, behavioral checklists, forced-choice systems, critical incidents, graphic rating scales, and behaviorally anchored rating scales (Cascio, 1998). Behaviorally based ratings and graphic rating scales have received a great deal of the attention devoted to performance appraisal research (Landy & Farr, 1980). Therefore, these are the two formats upon which attention will be focused.

CHAPTER 2: LITERATURE REVIEW

Graphic Rating Scales

The first known method of graphically representing an employee's performance emerged from the disenchantment about the fairness of a seniority-based system for promotions and raises. In 1922, Paterson developed, and published, what he called the graphic rating scale. The scale was a straight line for each dimension of performance to be measured with adjectives placed underneath the line to indicate level of proficiency. However, these labels were not anchors of any kind, they were simply guides. The rater was free to place a check mark anywhere along the continuum he felt best evaluated the ratee on that dimension. To translate this check mark into a score, a stencil was placed over the line, indicating a corresponding numerical value for the rater's evaluation. The rater would repeat this procedure for all of the dimensions for a specific employee.

Paterson (1922) felt this method had several advantages over other methods of evaluation. First, the procedure is very simple. All the rater is required to do is place a check mark on a line indicating performance on a certain dimension. Secondly, the rater can make a precise judgment about a worker's performance. The rater is not restricted in his responses and is not forced to place the ratee in a category or class. Finally, the rater is freed from quantitative terms such as numbers to describe a worker's performance. Paterson felt that these quantitative terms influenced a rater's judgement. With this method, the rater can evaluate performance without numbers biasing his judgment.

The reaction to this method of ratings was overwhelming. Graphic rating scales rapidly grew in popularity. Within 30 years of Paterson's publication, the graphic rating scale was the most popular method for assigning merit-based ratings in organizations (Benjamin, 1952). Ryan (1958) observed that the graphic rating scale was used in almost any organizational activity where it was necessary to evaluate an individual's performance. Over the years, with the advent of new methods of ratings, popularity of the graphic rating scales has declined somewhat. However, it still continues to be one of the most widely used and distributed methods for evaluating performance (Bernardin & Orban, 1990; Borman, 1979; Cascio, 1998; Finn, 1972).

The reason why this method still retains its popularity more than 75 years after its inception is most likely due to its many advantages. To begin with, graphical rating scales are very simple. They are easily constructed (Friedman & Cornelius, 1976) and implemented (Chiu & Alliger, 1990), and they are a cost-effective method of evaluating employees. In comparison, other methods of evaluating performance are very expensive and require a more complex development process (Landy & Farr, 1980). Another advantage of graphical rating scales is that the results from this method are standardized (Cascio, 1998; Chiu & Alliger, 1990). This means that once the employees have been evaluated, comparison can be made to other ratees for the purposes of disciplinary action (Jacobs, 1986), feedback and development (Squires & Adler, 1998), promotions and advancement decisions (Cleveland, Murphy, & Williams, 1989), etc. Also, graphical rating scales have the advantage of being appealing to the actual evaluator, or rater.

Some research has demonstrated that raters actually prefer to rate using graphic rating scales due to their simplicity and ease of rating (Friedman & Cornelius, 1976). Raters are typically more reluctant to use a rating method that is rather complex and involved (Jacobs, 1986). Ease of development, simplicity of use, relatively little expense, and generalizablity across ratees all make for a method of evaluation that is attractive to organizations.

As originally proposed by Paterson, a graphic rating scale was a check mark, or evaluation, made on a continuous line. However, this began to change very rapidly. Soon, graphic rating scales were being scored on computers used by researchers to make their jobs easier. Instead of using continuous lines, however, researchers were designing scales with anchor points along a continuum. Each anchor was given a certain value to facilitate entry into the computer (Bendig, 1952a). Limiting answers to a set number of anchor points (e.g., five, seven or nine) on a line replaced answering on a continuum. Instead of a graphic rating scale, this format could have been more appropriately labeled a 'forced interval' format. For better or for worse, this new format was soon being referred to as the "traditional" graphical rating scale format (Taylor & Hastman, 1956).

Graphic rating scales are relatively simple to develop. The first step is to use job analysis to identify and define the most important and most relevant dimensions of job performance to be evaluated (Friedman & Cornelius, 1976; Jacobs, 1986). It is also recommended that after relevant dimensions have been identified, they should be carefully refined to echo exactly what facets of job performance the rater wants to measure (Friedman & Cornelius, 1976).

Following this, the rater should decide how many scale points, or anchors, are needed on the rating scale. [This begs the question of whether anchors are needed at all (Landy & Farr, 1980). However, Barrett, Taylor, Parker, & Martens (1958) conducted a study on clerical workers in the Navy that helps to resolve this issue. In reviewing different rating formats to measure performance, he found that anchored scales, on average, are more effective than scales without anchors.] Bendig (1952a, 1952b) conducted studies with students who were to rate the performance of their college teachers. He found that increasing the anchoring on the rating scales led to increased reliability of the scale. It was assumed, for a while at least, that more anchors lead to better ratings. However, other research disputes this claim.

Lissitz and Green (1975) conducted a Monte Carlo study that investigated this matter. They noted that previous studies concerned with the number of anchor points on graphic rating scales have advocated either one specific number or no specific number of anchor points. They felt that deciding the proper number of points on a scale is based on the objectives and purpose of the study. However, they did suggest that 7 points are optimal for a scale, but the increase in reliability begins to level off after 5 points. The idea that a smaller number of scale points (for example, seven as compared to twelve or fifteen) is preferable is a sentiment echoed by other researchers (Finn, 1972; Landy & Farr, 1980). McKelvie, (1978) investigated the effects of different anchoring formats by

having students rate personality characteristics of certain groups of people. His results were consistent with those of Lissitz and Greene (1975).

Once the number of scale points has been decided, the scale developer should decide the format of the anchors (Jacobs, 1986). Anchors can either be numerical, adjectival, or behavioral in nature. French-Lazovik and Gibson (1984) claimed that both verbal (behavioral and/or adjectival) and numerical anchors are preferable when anchoring a rating scale. Barrett et al. (1958), however, demonstrated that behavioral anchors tend to clearly be more effective than numerical or adjectival ones. Other research has also arrived at the same conclusion (Bendig, 1952a; Smith & Kendall, 1963). Jacobs (1986) notes that this is most likely because these types of anchors communicate more clearly, to the raters, what each point on the scale represents. (In fact, it was interest in these behavioral anchors that spawned research into a new type of rating format which will be discussed in more detail later in this paper.) However, it should be cautioned that anchors could become too complicated.

Barrett et al. (1958) found that scale effectiveness decreased when too much information was included in the anchors. The extra information seems to confuse the rater and interfere with the rating process. There is also evidence that reliability does not necessarily increase for scales with more specifically defined levels (Finn, 1972). However, there is a general consensus that behavioral anchors are preferable to adjectives or numbers (Landy & Farr, 1980). In general, it seems that when constructing graphic rating scales, one should make sure to have approximately seven anchor points that are behavioral in nature, taking care not to include too much information in any one anchor.

Graphic rating scales are not without their critics or criticisms. Although their use was very popular and widespread, graphic rating scales were not subjected to much empirical testing until the years following World War II. However, it became clear very quickly that problems existed with graphic rating scales. Questions were raised, and many researchers soon became concerned with how these problems could impact the effectiveness and appropriateness of graphic rating scales' widespread use in organizations.

One of the problems with graphic rating scales that quickly became apparent after their introduction is the so-called 'halo effect.' When examining graphic ratings of performance, Ford (1931) found that there was a tendency for raters to give similar scores to a ratee on all dimensions of performance. To rate a worker in this manner would be the equivalent of rating the worker on one single scale, as opposed to many different scales that measure different aspects of work performance. Other researchers also discovered this problem. Soon, there was a great deal of literature documenting the problem of halo when using graphic rating scales (Barrett et al., 1958; Ryan, 1945; Ryan, 1958; Taylor & Hastman, 1956). More current literature has also documented the problem of halo, indicating that it continues to be a pervasive problem with graphic rating scales (Cascio, 1998; Keaveny & McGann, 1975; Landy & Farr, 1980; Tziner, 1984).

For a while, it was thought that halo could be eliminated, or at least attenuated, by training. By warning raters of this pitfall associated with the graphic rating scales, scores would contain less halo, and the ratings would be more appropriate. However, research has shown this not to be the case (Ryan, 1958). Some have proposed the alternative of statistical correction to compensate for halo. However, this process, also, seems to lack promise (Feldman, 1986).

Halo has traditionally been considered a serious problem for the effectiveness of an appraisal system. Organizations typically use performance evaluations to make some sort of decision about a worker and his job (Cleveland, Murphy, & Williams, 1989). When evaluating a person, the organization attempts to measure the worker on several different criteria. In this way, the worker, with the help of the organization, is able to be aware of his strengths and can target areas for improvement. Halo eliminates the variance between measurements of different performance dimensions. The person scores similarly across all dimensions and, thus, is unable to know which areas are strengths and which areas should be targeted for development.

In addition to halo, a leniency bias also plagues the use of graphic rating scales. Leniency is characterized by the tendency of a rater to be generous in his evaluation of an employee's performance across all dimensions of performance and across all ratees (Cascio, 1998). Like halo, leniency has been well documented as a source of error when using graphic rating scales (Barrett et. al., 1958; Bendig, 1952; Bernardin & Orban, 1990; Borman, 1979; Borman & Dunnette, 1975; Keaveny & McGann, 1975; Landy & Farr, 1980; Taylor & Hastman, 1956).

Leniency presents a problem for organizations in the following way. Performance appraisals are used to establish variance between the performance level of employees. Typically, these evaluations are used so that some merit-based decision can be made about the employees for the purposes of raises, promotions, benefits, etc (Cleveland et. al., 1989). These evaluations could also be used for employment decisions, deciding which employees should be terminated due to poor performance or which employees should be kept in an era of downsizing and layoffs (Bernardin & Cascio; 1988). Leniency eliminates the variance between employees, making it very difficult, if not impossible, to make organizational decisions based on the measurement of employees' performance.

New research, however, challenges the traditional notion of associating these so-called rating errors with poor judgements of worker performance. One of the first scientists to challenge the traditional conception of leniency and halo error was Borman (1979). He noted that the literature of the time supported the idea that most performance ratings were probably contaminated by error (e.g., halo, leniency), thereby rendering inaccurate ratings of employees. However, the results from his study failed to support this notion, and he suggested that an increase in accuracy was not as strongly correlated with a decrease in rating errors as once believed.

Over time, greater numbers of researchers began to realize the danger of equating "rating error" with a lack of accuracy. Murphy & Balzer (1989) found that the average correlation between rating errors and measures of accuracy was near zero. Based on the data, they felt that rating errors were not very likely to contribute to the decrease in rating accuracy. Jackson (1996) also found evidence that the point of maximum accuracy for a task does not necessarily coincide with the lowest measures of rating errors. Some researchers (Balzer & Sulsky, 1992) went on to claim that any relation (high, low, or zero) could be empirically found between accuracy and rating errors. Nathan & Tippins (1990) were even so bold as to claim that ratings errors might actually contribute to an increase in accuracy.

Gradually, performance appraisal researchers were beginning to realize that rating errors are not reliable or consistent indicators as to the effectiveness of performance ratings despite what was thought in the past (Balzer & Sulsky, 1992). The traditional conception that leniency and halo were only measures of error was wrong. A more plausible conceptualization was that these "rating errors" actually contained some true-score variance, not just error (Hedge & Kavanagh, 1988). Regardless, the traditional criticism of the graphic rating scale's susceptibility to these "errors" no longer holds the same concern that it once did.

There are other problems associated with graphic rating scales besides the traditional problems of halo and leniency. Graphic rating scales have also been accused of having problems associated with validity (Tziner, 1984), poor inter-rater agreement (Barrett et. al., 1958; Borman & Dunnette, 1975; Lissitz & Green, 1975; Taylor & Hastman, 1956), and personal biases of a rater (Kane & Bernardin, 1982; Landy & Farr, 1980). Though important, these other problems associated with graphic rating scales are not as prevalent in the research literature and have not traditionally been attributed the same level of importance and influence as halo and leniency.

Behaviorally Anchored Rating Scales (BARS)

Due to the growing disenchantment with graphic rating scales (Ryan, 1958) and due to their own desire to develop a better method of rating employees' performance, Smith & Kendall (1963) devised a new method of appraising performance, Behaviorally Anchored Rating Scales (BARS). They felt that evaluations varied too widely from rater to rater using older methods. They wanted a method that relied on interpretations of behaviors in relation to specified traits. They felt that better ratings could be obtained from a rater "…by helping him to rate. We should ask him questions which he can honestly answer about behaviors which he can observe." (Smith & Kendall, 1963, p. 151)

The format they used was a series of graphic rating scales arranged in a vertical manner. Behavioral descriptions representing parts of desired performance dimensions were printed at various heights on the graphical scale, serving as anchors. A typical scale will usually contain seven or nine of these behaviorally anchored points (Landy & Barnes, 1979). These descriptions serve as anchors in aiding the rater's evaluation. Smith and Kendall (1963) note that "the examples we used, therefore, represent not

actual observed behaviors, but inferences or predictions from observations." (p. 150) The whole premise behind their method was to "…facilitate a common frame of reference so that they would look for the same kind of behaviors and interpret them in essentially the same way. (Bernardin & Smith, 1981, p. 460)"

This method, introduced by Smith and Kendall (1963), immediately became popular. It has several advantages over graphic rating scales. The behaviorally based method doesn't seem to suffer from the traditional problems of leniency and halo which plague graphic rating scales. Several researchers (Borman & Dunnette, 1975; Campbell, Dunnette, Arvey, & Hellervik, 1973; Friedman & Cornelius, 1976; Keaveny & McGann, 1975; Tziner, 1984) have noted a reduced susceptibility to halo and/or leniency error with BARS as compared to graphic rating scales. Keaveny and McGann (1975), however, did find some conflicting results. When students were asked to rate their professors with either a behaviorally based scale or a graphic rating scale, although the BARS method did have reduced halo error, the two methods did not differ in their amount of error due to leniency. However, two separate reviews of the performance appraisal literature (Kingstrom & Bass, 1981; Landy & Farr, 1980) both support the notion that the BARS method does, in fact, promote decreased levels of leniency and halo.

Another distinct advantage of the BARS method is that the scales are often developed by the same people who will eventually use them (Campbell et al., 1973; Smith & Kendall, 1963). This is helpful in the sense that when eventual raters participate in scale development, they may have a heightened understanding and awareness of the scale, and they might even gain more insight into the job they are to rate (Friedman & Cornelius, 1976; Smith & Kendall, 1963). Also, BARS are helpful to raters by their usage of technical terms. The BARS method defines the dimensions and behavioral labels in the language and terminology of the rater (Campbell et al., 1973; Smith & Kendall, 1963). This helps to ensure that the dimensions and the labels are interpreted the same by all raters (Campbell et al., 1973). Perhaps one of the most appealing features of BARS is that it uses specific behavioral incidents as anchors. Barrett et al. (1958) and Borman (1986) both noted that the increased specificity of the behavioral anchors gives the rater a more concrete guideline for making his evaluation. In addition, behaviorally based methods have been shown to lead to elevated levels of goal clarity, goal commitment, and goal acceptance compared to graphic rating scales (Tziner, Kopelman, & Livneh, 1993), and they have led to increased interrater agreement (Kingstrom & Bass, 1981). In summary, the BARS format not only corrects for errors in graphic rating scales, it provides many substantial advantages over other scale formats.

Admittedly, developing a BARS scale is an involved process. It requires careful planning and precision in order to develop a "successful" scale (Bernardin & Smith, 1981). Smith and Kendall's (1963) procedure for developing a BARS scale can be summarized in three steps. First, items must be selected that distinguish good from mediocre from poor incidents of performance. Borman (1986) notes that several examples of performance on all levels should be collected from individuals who are knowledgeable about the target position. These people are referred to as Subject Matter Experts (SMEs).

The second step is clustering. The items must be grouped into similar categories, or dimensions, of performance (e.g., communication with coworkers) that overlap as little as possible with other performance categories (Landy & Barnes, 1979). When the dimensions have been defined, a second group of SMEs place each behavioral example into what they believe are the appropriate categories (Borman, 1986). This process is known as retranslation (Cascio, 1998).

The third, and final, step involves scaling (Landy & Barnes, 1979). The investigator, or person developing the scale, then decides, based on the responses of the second group of SMEs, which behavioral incidents are to be included as anchors for each dimension of performance (Borman, 1986). Once this is finished, the behavioral anchors are placed on a vertical scale for each dimension, and the raters are able to record the behavior specified on each scale (Bernardin & Smith, 1981; Smith & Kendall, 1963).

A related issue, but one that is not given much attention, is the number of anchors on the scale. Previously, it was noted that approximately seven was the optimum number of scale points for graphic rating scales. Contrary to research that recommends otherwise (Landy & Farr, 1980; Lissitz & Greene, 1975), BARS scales typically have nine anchor points on their scales (Borman, 1986; Borman & Dunnette, 1975; Cascio, 1998; Landy & Barnes, 1979).

Though BARS scales became very popular after their introduction, this method, also, was not without its concerns. The first criticism of BARS scales came in the very publication which introduced them. Smith and Kendall (1963) noted that the raters would be judging behaviors that are complex in nature. This raises a potential problem if one rater attributes a behavior to one cause while a second rater attributes it to another. The idea that different raters all rate similarly, also called interrater agreement, is vital to a performance appraisal system. A lack of interrater agreement means that the results of the appraisals cannot be generalized across different raters. Employees rated by one rater may have received different scores on their appraisals had they been rated by a different rater. This issue is important when comparing employees with different supervisors, or raters, for the purposes of advancement or termination.

Another problem can potentially arise due to the nature of the behavioral incidents. Raters may have difficulty detecting similarities between the ratee's observed performance and the behavioral anchors (Borman, 1979). Because the anchors are very specific, finding congruence between the anchors and the performance can involve a high amount of inference. And, as Cascio (1998) notes, the more inferences made by a rater, the more likely that errors will occur. As well, it is possible that the ratee could have acted in direct accordance with two of the specific behavioral anchors (Bernardin & Smith, 1981). The problem for the rater is then to decide which example is more correct. This, again, involves inferences that could lead to rating errors. There is even some evidence that the nature of the behavioral anchors seems to increase rater error (Murphy & Constans, 1987). The specific nature of the behavioral incidents may trigger memories of individual incidents of behavior that match the anchors rather than serve to facilitate a more general recall of behavior. Also, given that many individuals will be rated after the

rater has already seen the appraisal instrument, the specific nature of the behavioral anchors may serve to prime the rater to look more carefully for behaviors that match those on the rating scale (Murphy & Constans, 1987).

Perhaps the largest problem with BARS scales is their development. Several authors (Borman & Dunnette, 1975; Campbell et al., 1973; Landy & Farr, 1980) have noted that BARS scales are extremely expensive and time-consuming to develop. Campbell et al. (1973) noted that managers did learn a great deal from the process. However, they invested a tremendous amount of time and energy. The time spent could be occupied with any number of tasks from administrative duties to rating employees. The company not only has to pay the manager for time spent helping to develop the scale, but they also have to pay many other managers for the same activity as well as fund the staff that is overseeing the construction process. It becomes a question of whether the gains of the BARS method outweigh the costs of development and administration. Some authors (Borman, 1979) cast serious doubt on the idea that the advantages outweigh the disadvantages of BARS scales, claiming that the time and effort spent is unwarranted.

The research literature in the 1970's focused extensively on the differences between different rating formats, most often the differences between graphic rating scales and BARS. However, the actual differences between the scales were very simple. Different numbers and different styles of anchors are the real difference between graphic rating scales and BARS. Graphic rating scales rely on relatively simple anchors, numbers, to guide performance evaluations. On the other hand, behavioral anchors are much more complex in nature, and they are much more specific. So, at the most basic level, the only real difference between graphic rating scales and BARS is the specificity of the anchors.

Traditionally, the specific nature of the anchors in BARS has been considered one of its advantages. Some claim that the specificity of behavioral anchors communicates what each point on the scale represents better than less specific anchors (Jacobs, 1986). As mentioned previously, these more specific anchors were considered to be more preferable than less specific anchors (Barrett et. al, 1958; Bendig, 1952a; Smith & Kendall, 1963). The general belief was that scales with more specific anchors were more effective. However, the "effectiveness" criteria included increased reliability and variance of the ratings (Finn, 1972). Researchers typically did not focus on measures of rating accuracy when evaluating scales.

Rating accuracy has long been a neglected criterion for determining the effectiveness of performance appraisals. Literature has recently begun to focus on this measure of rating quality, however. (The importance of accuracy as a criterion of rating effectiveness will be addressed at length in a subsequent section of this paper.) Based on this new focus, research should re-examine old conclusions about anchor specificity with a new criterion, accuracy. Behavioral incidents are not necessarily the "ideal" anchor that some researchers have claimed them to be. As early as 1959, Kay cautioned that critical incidents of behavior may be too specific for use as scale anchors. Barrett et al.

(1958) also noted that to much information contained in an anchor can have adverse effects on evaluating performance.

Although research on the topic is limited, there is some evidence to indicate that more specific anchors can actually serve to bias the subject to respond in a certain manner. A verbal label near the middle of a rating scale can actually serve to increase or depress the value of ratings (French-Lazovik & Gibson, 1984). This, of course, would not occur with a less specific type of anchoring such as a numerical anchor. Murphy & Constans (1987) have also focused on the biasing nature of anchors. They note that increased specificity of anchors does not always translate into increased rating effectiveness. They claim that behavioral anchors, specifically, affect the rater by biasing memory for behavior.

Behavioral anchors are not as 'ideal' as once suspected. As previously noted, research needs to be conducted that challenges the notion that behaviorally-based anchors are superior to other forms of anchoring with a focus on rating accuracy as the determinant of 'effectiveness.' Current literature cautions that increased specificity of anchors may bias the raters to rate in a certain manner. Taking this into consideration, one might arrive at the conclusion that less specific anchors are actually better.

A Quandary with Appraisal Instruments

It is quite apparent from the voluminous research that although they both suffer from the same types of subjective biases, Graphic rating scales (GRS) and BARS scales also have their own advantages and their disadvantages. The question to both researchers and consultants in the field of Industrial Psychology then becomes: Which format is the more desirable? Borman (1979) noted that despite the numerous studies devoted to rating format research, the research provides no clear picture of which type of scale is the "best." Two years later, Kingstrom & Bass (1981) published a study that echoed Borman's sentiments. Although they felt it was inappropriate to claim that BARS are not superior to traditional rating methods, there is relatively little empirical evidence to support such a position. As early as 1958, Barrett et al. revealed that variability in ratings across ratees, a desirable characteristic of performance appraisals, did not systematically vary as a function of the rating format. Around the same time, Taylor & Hastman (1956) arrived at a similar conclusion. They found that varying the rating format did not result in increased interrater reliability, nor did it result in increased variability of ratings or, as they called it, dispersion.

It appears that rating format research has come full circle. There are currently several different methods available to organizations who wish to appraise their employees' performance, including GRS and BARS. However, they are faced with a difficult task of deciding which method to use. Unfortunately, all of the aforementioned research indicates that if they picked a method at random, they would get very similar results. Organizations are left to decide for themselves what constitutes the best method of performance appraisal. Doverspike, Cellar, & Hajek (1987) note that the problem in the research literature stems from failure to achieve consensus regarding what criteria constitute the "best" method of rating performance. They point out that some researchers

use freedom from leniency or halo error as the prime criterion for determining the superiority of a rating method. Other methods of determining a rating scale's worth include variability in ratings and interrater agreement. Based on Doverspike et al.'s argument, it is difficult, if not impossible to determine which rating format is superior. However, new research is focusing on rating accuracy as the best criterion for assessing the quality of performance ratings.

The concern for accuracy in the ratings of workers' performance levels has always been present. However, historically, the level of accuracy in performance evaluations has always been implied by examining other statistical measures. Most commonly, rating errors, or the lack thereof, have implied the level of accuracy in performance evaluations (Murphy & Cleveland, 1995) . The vast majority of studies that compared different rating formats examined criteria other than explicit measures of rating accuracy. Several studies compared rating formats on the basis of reliability (Barrett et al., 1958; Borman & Dunnette, 1975; Kingstrom & Bass, 1981) and variance among performance ratings of workers (Barrett et al., 1958; Borman & Dunnette, 1975). However, many studies compared rating formats based on the notion of reducing leniency and halo, so-called "rating errors" (Borman & Dunnette, 1975; Chiu & Alliger, 1990; Friedman & Cornelius, 1976; Kingstrom & Bass, 1981). The conclusion of these studies, both independently and collectively, was that performance ratings are not affected by the rating format. Landy & Farr (1980) examined the research conducted to date and concluded that format does not affect the outcome of the performance ratings. The foundation for these arguments, however, was faulty.

As previously stated, it was traditionally assumed that accuracy was linked to rating errors. More specifically, the fewer rating errors contained in a rating instrument (lower levels of leniency and halo), the higher the level of accuracy. However, research soon began to accumulate that suggested this was not entirely true. Bernardin & Pence (1980) actually found a situation where lowered levels of leniency and halo corresponded with a decrease in rating accuracy. They trained subjects to evaluate performance in a manner such that there was greater variation of ratings both within subjects and across subjects. However, the dilemma they encountered was that the "true" level of performance exhibited by ratees may such that there is little variation in performance from one ratee to the next or across dimensions within the same ratee. So, by lowering these measures of leniency and halo, Bernardin & Pence actually lowered the level of rating accuracy as well.

In subsequent research, the ideas of Bernardin & Pence were expanded upon. For example, Balzer & Sulsky (1992) commented that it is even possible to achieve a zero correlation between measures of accuracy and halo error. In a meta-analysis, Murphy & Balzer (1989) even made the bold claim that rating errors could actually contribute to, rather than take away from, measures of rating accuracy. So, based on current research, it would be erroneous to conclude that the level of accuracy of performance evaluations can be inferred from the level of rating errors. In light of these recent findings, it is apparent that rating format research should be revisited. This time, instead of focusing on reliability, variance in ratings, or freedom from rating errors such as halo or leniency,

research should use rating accuracy as the primary criterion for determining the effectiveness of ratings.

CHAPTER 3: HYPOTHESIS

Theory

When Landy & Farr (1980) made their claim that rating format research was useless and should be abandoned, researchers turned their attention away from research on rating formats. Instead, research turned inward, focusing upon the cognitive processes of the rater. When studying these cognitive processes, researchers focused on the purpose of appraisal (Murphy, Philbin, & Adams, 1989), the cognitive makeup of the rater (Schneier, 1977), internal drive and self-affirmation for the rating task (Neck, Stewart, & Manz, 1995), and the effect of memory and judgement on ratings (Woehr & Feldman, 1993). Instead of focusing on devising better rating instruments, the focus of this new approach is to make better raters.

In light of the new focus on the cognitive aspects of ratings, perhaps one goal of the research on performance appraisals should be to devise a method of rating performance that helps the rater rate. All major rating formats explored by research thus far, BARS, Behavioral Observation Scales, and even GRS, do not allow a rater to document workers' performance levels at the same level at which the rater cognitively interprets and rates the behaviors. All of these methods require a rater to judge performance on a scale, usually from a value to 1 to a value of 5, 7, or 9. All major rating format research to date has not allowed the rater to finely, and accurately, discriminate between stimuli. Some research has suggested that as a rater rates more and more ratees, the rater's ability to discriminate between the different levels of performance can actually increase (Coren, Porac, & Ward, 1979). To allow the raters to rate at this high level of discrimination, continuous rating scales should be used.

Rating performance on a continuous scale would allow raters to discriminate between ratees at as precise of a level as they desire. No longer would raters be tied to anchors as the only possible responses about performance levels. Simply increasing the number of anchors on a rating scale would not have the same effect. Increasing the number of anchors to eleven, fifteen, or even twenty points does not allow for the rater to rate in the areas between the scale anchors. Increased anchoring still does not allow for the maximum level of discrimination between ratees that is possible. A continuous scale, however, does allow for this level of discrimination.

A continuous rating scale can be thought of, conceptually, as having an infinite number of anchor points. Some research has focused on the effects of increasing the number of anchors on a rating scale. Most of the studies that focus on the topic of scale anchors (Finn, 1972; French-Lazovik & Gibson, 1984; Lissitz & Greene, 1975) have not examined the effect of the number of scale points on rating accuracy. (The importance of rating accuracy as the primary criterion for judging the effectiveness of a rating scale was discussed earlier in the paper.) Their conclusions that five or seven scale points is the "optimum" number are, therefore, based on criteria such as reliability, variance in ratings, and freedom from rating errors in the performance ratings. Therefore, no assumptions can be made about the detrimental effect of increased anchoring on rating accuracy.

The focus of this study comes from previous theoretical work and empirical support. Clearly, the rating format research of the past has garnered few, if any, real advances in the area of performance appraisal. It is the supposition of this paper that: (1) the lack of substantial advances in the area of rating format research in the past is due to the fact that previous research did not seek to modify performance appraisals that fit with the cognitive structure of the raters (i.e., previous research did not adopt some method of rating on a continuous scale), and (2) there was an inappropriate focus on minimizing "rating errors" rather than striving to increase accuracy. Most of the traditional rating format research has focused on inappropriate criteria: leniency, halo, and variability of ratings. Instead, rating format research should be re-examined and focus on maximizing accuracy rather than minimizing error or increasing variability in ratings.

It should be noted that accuracy can actually be conceptualized in different ways (Cronbach, 1955). Elevation can be thought of as the overall level of rating, combined across raters and different performance dimensions. Differential Elevation collapses accuracy judgements across rating dimensions, but it examines each rater separately. The opposite of differential elevation, stereotype accuracy, examines each performance dimensions separately, but it combines the judgements of all raters. The final measure of accuracy, differential accuracy, examines each performance dimensions separately for each rater.

Through the history of rating format research, there has been an implicit assumption that scale anchors are necessary for rating scales to be effective. Perhaps this is the reason that, to date, no one has examined a "true" graphical rating scale sans anchors. As previously noted, there is some support for the idea that anchors can actually serve to bias a rater (Murphy & Constans, 1987). The presence of these anchors misdirects the observations or recall of the ratee's behavior. The presence of any type of anchor (numerical, adjectival, or behavioral) can have the effect of biasing a rater's judgement. However, behavioral anchors generate the largest amount of bias on a rater's performance judgements. Murphy & Constans concluded that a rating scale might not always benefit from scale anchors, particularly those of a behavioral nature.

Also, traditional rating format research has focused on the rating instrument but has largely ignored the rater. As demonstrated above, the natural tendencies of human perception, to notice very slight differences in stimuli (worker performance), have been incompatible with the rating formats used to date.

Hypothesis 1: Performance evaluations conducted using a graphic rating scale (GRS) without any type of anchors will demonstrate higher levels rating accuracy when compared to a standard than will performance evaluations conducted using a graphic rating scale (GRS) containing numerical anchors.

Hypothesis 2: Performance evaluations conducted using a graphic rating scale (GRS) without any type of anchors will demonstrate higher levels of rating

accuracy when compared to a standard than will performance evaluations conducted using a behaviorally anchored rating scale (BARS).

Hypothesis 3: Performance evaluations conducted using a graphic rating scale with numerical anchors will demonstrate higher levels of rating accuracy when compared to a standard than will performance evaluations conducted using a behaviorally anchored rating scale (BARS).

Interrater reliability is also a topic of concern for performance appraisals. This measure determines the extent to which the various raters agree on the level of observed performance of the ratees. One could argue that continuous rating scales with no anchors should lead to higher reliability than scales with anchors. When a rater observes performance from several ratees, the rater mentally rank-orders the ratees and then rates each ratee. A ratee's rating depends on their rank order compared to other ratees. A continuous scale with no anchors allows the rater to precisely preserve the ratees' rank orders in the rating process with no bias. Since there is an almost infinite number of places along the rating continuum where a ratee will fit, each ratee can be rated accurately, and the rank order of the ratees can stay consistent.

In continuous scales with anchors, however, performance ratings can be biased by the increasing specificity of anchors (Murphy & Constans, 1987). Due to this bias, a rater's evaluations are more likely to fall closer to the scale anchors. When the performance ratings fall closer to the anchors, there is a greater chance that the integrity of the rank ordering of the ratees is not preserved. For example, if a ratee's performance has a true score of 4.63, the anchors may bias the rater to actually record the level of performance closer to the "5.00" anchor. The score may actually be recorded as a 4.8. However, ratees' true scores of 4.58 and 4.75 may also be recorded as 4.8 due to the biasing effect of the anchors. When this happens, the rank order of the ratees is not preserved. All three ratees appear to have the same level of performance when, in fact, they do not. A different rater may not be as biased by the anchors and may record a performance value closer to the true score for each ratee. Thus, there is a lack of agreement between the two raters regarding the level of performance exhibited by these three ratees. Due to the biasing effect of the anchors, there would be a low level of interrater reliability. In a scale without the bias from anchors, this problem would not occur.

Hypothesis 4: Graphic rating scales with no anchors will demonstrate higher levels of interrater reliability for performance evaluations than will BARS or graphic rating scales with numerical anchors.

CHAPTER 4: METHODS

102 subjects (51 female) participated in one of three rating format conditions. Students participating in this experiment were drawn from an Introductory Psychology course at Virginia Tech University. The study is a 3 (format) x 10 (ratee) x 6 (dimension) design with accuracy measures collapsing across both ratees and dimensions. For analysis, the study is a one-way design. The three between subject conditions of the study are based on format used to evaluate performance: a traditional BARS scale, a graphic rating scale with numerical anchors, and a graphic rating scale without any kind of anchors. All subjects were randomly assigned to one of the three between-subjects conditions. In each condition, subjects were asked to rate the performance of ten different people on each of six different performance dimensions. The students received extra credit for their participation.

Vignettes and Rating Scales

The stimuli for the subjects in this experiment were a series of vignettes that describe the actions of teachers in the classroom. The vignettes were composed of a set of critical incidents depicting teachers' behaviors in the context of a class setting. The critical incidents used for the vignettes came from a larger pool of critical incidents. The same pool of critical incidents was also used to construct behavioral anchors in the BARS format condition of the experiment. Although both the vignettes and the BARS scales draw on the same pool of critical incidents, the two groups do not share any common incidents of behavior.

The critical incidents of teaching behavior were developed in accordance with specific guidelines concerning the development of BARS scales (Cascio, 1998; Smith & Kendall, 1963). The first step in the scale development was to collect critical incidents of teaching behavior. Smith & Kendall (1963) suggested that the people used to generate critical incidents of the ratee's behavior should be knowledgeable in the specific field. College students were judged to have the appropriate familiarity and knowledge of the ratee's job behaviors, so they were used to generate critical incidents of teaching behaviors. Thirty-six undergraduate students from Virginia Tech were used to gather the critical incidents. Once the lists of behaviors that the students generated were screened to eliminate repetitions, vague items, and non-specific behaviors, there were 234 specific teaching behaviors remaining.

The next step in the development of the scale was to assign each behavior to the appropriate dimension of teaching performance. The experimenter screened the list of behaviors for common trends or groups of teaching performance. In all, ten distinct dimensions of teaching performance were identified in the list: teacher dedication, class preparation, classroom organization, technological savvy, teacher expertise, courtesy and respect for students, adequately preparing students for exams, appropriate class content, appropriate grading procedures, and classroom delivery and presentation. Once the dimensions were identified by the experimenter, another group of twenty undergraduate students from Virginia Tech University were used to assign each behavior to a specific

dimension. Of the original 234 incidents of behavior, seven items were deleted due to lack of agreement among subjects when assigning those items to a specific dimension.

The final step in scale development involved retranslation. This is a process of assigning a numerical value to each of the 227 specific incidents of behavior. Traditional BARS scales are based on a scale of one to nine (Cascio, 1998; Campbell et al., 1973; Smith & Kendall, 1963). Hence, when assigning numerical values during the retranslation process, subjects were asked to rate each item as to the extent it represents effective teaching within the context of its appropriate dimension. Again, the subjects used for the retranslation process were 28 undergraduate students from Virginia Tech University.

This whole process yielded a list of 227 specific incidents of teaching behaviors grouped into ten distinct dimensions of performance. A combination of these incidents was then used to construct ten different teaching scenarios with each scenario representing the actions of a different teacher within the context of class. Each scenario contained specific incidents related to all six performance dimensions. Rating scales were constructed to measure the teacher's level of performance across each of the six dimensions. A continuous graphic scale with numerical anchors and a continuous graphic scale without anchors were constructed along with a BARS scale. For the BARS scale, the behavioral anchors for each dimension were placed along the scale according to their values as assigned in the retranslation process.

BARS scales traditionally range from one to nine. In order to make fair comparison across rating formats, it was decided that both types of graphic scales should also yield rating values between one and nine. This allows for direct comparison of responses on a certain performance dimension for a specific vignette across rating formats. In addition, all rating formats are continuous scales. In other words, a subject can also rate performance at any point between the numerical anchors along the rating scale and not just be limited to rating on the numerical anchors. The computer program allowed precise measurement of the ratings to two decimal places. This is to allow for a fair comparison between the "observed score" and the "true score" since the "true score" for each of the behavioral incidents used in the teaching vignettes is also measured to two decimals.

All six performance dimensions measure behaviors that are familiar to the subject and can be easily identified or recognized. Teacher dedication refers to behaviors that show a teacher's attitude toward the class and their commitment to both the class material and the students. Classroom preparation and organization is concerned with the "nuts and bolts" of the teacher in the classroom: how organized they are, whether classroom behavior is random or it has a planned purpose, does the teacher always have the necessary materials for class to run effectively, etc. For the dimension of teacher expertise, the teacher's overall level of knowledge about the class material is being measured. Incidents included in the courtesy and respect for students dimension are concerned with topics such as the how respectful the teacher is toward the students, the level of consideration displayed toward the students, and the teacher's general demeanor

in interacting with students. The adequately prepares students for exams dimension measures how well effective a teacher is in giving students the tools and skills necessary to perform well on the class tests. The last dimension, classroom delivery and presentation, refers to the style of a teacher's lectures, the effectiveness of a teacher in class, and how successful they are in presenting the class material to the students.

Procedures

Subjects were randomly divided into three groups, or experimental conditions. In the first condition, the subjects rated incidents of teaching performance on a BARS scale. In the second condition, subjects rated teaching performance using a continuous graphic rating scale with numerical anchors. In the third condition, subjects used a continuous graphic rating scale with no anchors to rate incidents of teaching performance. After the subject signed the informed consent form, the entire experiment was conducted on the computer. The subject was first asked a set of demographic questions (gender, ethnic background, classification in school, and age). Next, the computer provided a detailed list of instructions for the subject as well as a sample vignette and questions. The BARS format (see Appendix A), the graphic scale with numerical anchors format (see Appendix B), and the graphic scale with no anchors format (see Appendix C) all have very slight differences in the instructions. Any instructional variation was only meant to familiarize the subject with the proper rating scale and not to vary the experimental procedures across the different rating formats. Once the instructions were administered, the subject was presented with a vignette depicting a teacher's behavior in the context of a class (see Appendix A, B, and C). The same behavioral vignettes will be used for all rating formats. The subject read the short vignette and then answered a series of questions about the performance of the teacher in each of the six dimensions previously noted (see Appendix A, B, and C). The six dimensions are the same for all rating formats. The scale used by the subject to evaluate teacher performance was dependent upon the experimental condition to which the subject was assigned.

In the BARS condition, a nine-point BARS scale overlays the scale continuum (see Appendix A). The only difference from a traditional BARS scale is that the subject was free to respond to the dimension at any point along the scale line between one and nine. In the continuous graphic scale condition with numerical anchors overlaid on the scale continuum (see Appendix B), the subject was allowed to respond to the question anywhere along the scale. However, to guide a subject's responses, the scale contained the numerical anchors one through nine equally spaced along the scale continuum. In the continuous graphic scale condition with no anchors, the same graphic scale will be used, but the anchors will be removed (see Appendix C). All that will remain will be "Low" at the extreme bottom end of the scale and "High" at the extreme top end of the scale. As in the previous conditions, the subject was allowed to respond anywhere along the scale continuum. In all rating conditions, the subject used the computer mouse to click at the point along the scale continuum that represented the subject's rating of the teacher's level of performance. Based on the point along the continuum indicated by the subject, the computer generated a value indicating the level of performance (on a scale of one to nine) precise to two decimal places.

The subject rated the teacher presented in the vignette on each of the six dimensions listed above. Once finished with a vignette, the subject then proceeded to the next vignette and rated that teacher on the six performance dimensions. While rating a specific teacher, any of the ratings within that specific vignette could be altered as many times as necessary. However, once a subject proceeded on to the following vignette, the ratings of the teacher were not be able to be changed. The subject followed the same procedure until all ten teachers (vignettes) had been evaluated on all six performance dimensions. At that time, the subject then answered a series of six likert-type questions concerning personal preference and liking for the appraisal instrument (see Appendix D). Following these questions, the subject's participation was concluded.

Follow-up Questions

At the end of the experiment, the subject were asked a series of likert-type questions that dealt with their feelings about the vignettes and the scales used to judge teaching performance (see Appendix D). The subjects were asked (1) how comfortable they were in using the scales, (2) how clear were the behavioral scenarios, (3) the quality of the rating scales, (4) how well they liked the scale, (5) how realistic the scenarios were, and (6) how well the behavioral scenarios were constructed. These questions allow the researchers to determine the quality of the appraisal instrument. These responses could serve to guide construction of future appraisal instruments. In addition, these responses allow us to examine the level of subjects' rating accuracy as compared to their liking for the appraisal instrument.

Dependent Measures

Subjects in this study evaluated a teacher's level of performance on class-related duties. Six performance dimensions were evaluated for each of ten different teachers for a total of sixty total ratings. Two of the scales used by the subjects were labeled with anchors ranging from one to nine, and the third scale contained no anchors. Based on the number of pixels on the computer screen between the two ends of the rating scale, the computer program calculated the value for each of the ratings based on where the subject responded along the scale continuum. Based on the position of the subject's response on the answer continuum, the computer calculated a value of "level of performance" precise to two decimal places for all three rating formats. The performance dimensions that were evaluated included teacher dedication, class preparation and organization, teacher expertise, courtesy and respect for students, adequate preparation for exams, and classroom delivery and presentation.

True scores of performance were developed through the BARS procedure for developing behavioral incidents of performance (see Table 1). The true scores were established in the same manner as the value for the behavioral anchors used in the BARS. Since these true scores were established using the proper BARS procedure, they are more diagnostic of their respective performance dimensions than a list of behavioral incidents that did not go through the retranslation procedure. Because they did go through the proper procedure, there is less overlap across performance dimensions for specific

behavioral incidents and they retain more diagnosticity of their respective dimensions of teaching performance.

Table 1: True Scores of Behavioral Incidents

	Dim.1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Ratee Average
Ratee 1	7.61	2.61	8.56	1.46	8.36	2.11	5.12
Ratee 2	2.04	2.11	7.5	7.86	7.5	2.93	4.99
Ratee 3	7.39	7.71	1.54	1.18	8.14	7.54	5.583
Ratee 4	7.36	2.29	8.00	1.71	7.75	7.71	5.81
Ratee 5	1.68	7.43	7.50	2.00	1.46	2.82	3.82
Ratee 6	8.11	2.11	2.11	7.82	8.21	1.82	5.03
Ratee 7	8.18	7.43	7.29	7.86	1.96	2.29	5.84
Ratee 8	3.04	8.00	2.14	1.96	8.36	2.50	4.33
Ratee 9	8.11	7.39	1.71	8.32	8.14	2.11	5.96
Ratee 10	7.79	7.86	8.56	8.11	8.25	7.79	8.06
Dim. Average	6.13	5.49	5.49	4.83	6.81	3.962	*5.45

Note: The value marked with "*" represents the overall mean of all true scores

Dependent variables for the study include the four separate measures of accuracy (Cronbach, 1955). The first of these measures, elevation, reflects the judgements of all ratees combined across all performance dimensions being evaluated compared to the "true score" or target ratings. Elevation is the measure associated with the ideas of rating errors such as leniency or halo (Hauenstein, Facteau, & Schmidt, 1999).

$$E = \text{Square Root}[(X.. - T..)^2]$$

In this equation, X.. represents the grand mean of all observed scores of performance, and T.. represents the grand mean of all true scores. Differential elevation, is the second type of accuracy judgement. However, this type of accuracy collapses across dimensions and only focuses on the evaluation of each ratee. It can be represented by the following:

$$DE = \text{Square Root}[\{\sum[(X_{i.}-X..)-(T_{i.}-T..)]^2\}/6]$$

In the figures X_{ij} or T_{ij}, "i" refers to the ratee being evaluated and "j" represents the dimension on which ratee "i" is being rated. A "." represents the mean of the appropriate level of the statistic, and X represents observed scores while T represents true scores. For example, $X_{1.}$ would represent the mean of a rater's observed scores on ratee 1 averaged across all performance dimensions.

Stereotype accuracy is the opposite of differential elevation. It focuses on how a rater judges performance on the different dimensions, but it collapses across ratees. It is computed using the following equation:

$$SA = \text{Square Root}[\{\sum[(X_{.j}-X..)-(T_{.j}-T..)]^2\}/10]$$

Differential accuracy, the final accuracy judgement, reflects how a rater rates each ratee on each dimension. This measure examines accuracy for all ratees on each dimension and can be computed with the following equation:

$$DA = \text{Square Root}[\{\sum[(X_{ij}-X_{i.}-X_{.j}+X..)-(T_{ij}-T_{i.}-T_{.j}+T..)]^2\}/60]$$

It should be noted that on all accuracy measures, values closer to zero are better. That is, the lower the value of any measure of accuracy, the more accurate the performance ratings are as compared to the true score. Without specifying purpose of appraisal, there would be no a priori reason to state which measure of accuracy is the "best" for this study. Elevation might be the most desirable if the purpose was to assess a global pattern of a rater's ratings. Differential elevation would be the preferred variable if one was interested in a ratee's overall level of performance across performance dimensions. If the area of concern was improving areas of performance that were substandard across all employees, stereotype accuracy would be the preferred measure of accuracy on which to focus. Also, if one was interested in looking at each ratee's level of performance on each performance dimension, differential accuracy would be the most important measure. However, without any of these concerns, there is no reason to state, a priori, which is the most important measure of accuracy for this study.

Measures of interrater reliability were also used as dependent measures. The reliability estimates are calculated within each of the three rating formats. Interrater reliability is a measure that illustrates the degree to which the responses of one rater are similar to other raters within that same format. To assess reliability within a format, the responses of each subject were correlated with the responses of every other subject within that same rating format. A mean of these correlations was used for the measure of interrater reliability.

In addition, the answers to the follow-up questions asked of the subjects are dependent variables. The six questions are measured on a 7-point likert scale. These measures were collected in order to assess preferences and comfort levels with the rating scale used by each subject.

CHAPTER 5: RESULTS

Before assessing the hypotheses, potential moderators of the rating format/accuracy relationship were tested. An analysis of variance revealed that there were no mean differences in the four types of accuracy based on age, classification, sex, or ethnicity significant at the $p < .05$ level (see Table 2). Descriptive statistics for the subjects' responses can be found in Tables 3, 4, and 5. These tables contain the means and the standard deviations for each of the sixty performance evaluations made by each subject. In the tables, the results are separated according to rating format used. The overall mean of the observed scores for each of the three formats was higher than the overall mean of the true scores. This indicates a slight tendency to rate in a favorable manner. This bias toward more positive ratings can also be seen by an examination of the statistics. Also, Table 6 provides the intercorrelations between the performance dimensions. The levels of intercorrelation between the performance dimensions range from .214 to .592 with an average correlation of .388. These are, generally speaking, modest level correlations. Typically, performance appraisals have higher levels of intercorrelation between performance dimensions.

Table 2: Demographic Variables' Influences on Rating Accuracy

	Variable	DF	F-value	Prob > F
Elevation	Age	2, 99	.124	.884
	Class	4, 97	.241	.914
	Ethnic	4, 97	.191	.943
	Sex	1, 100	.393	.532
Differential Elevation	Age	2, 99	.196	.822
	Class	4, 97	.716	.583
	Ethnic	4, 97	2.014	.109
	Sex	1, 100	.048	.827
Stereotype Accuracy	Age	2, 99	2.329	.103
	Class	4, 97	1.706	.155
	Ethnic	4, 97	1.641	.170
	Sex	1, 100	.048	.827
Differential Accuracy	Age	2, 99	2.431	.092
	Class	4, 97	1.761	.143
	Ethnic	4, 97	2.291	.072
	Sex	1, 100	.479	.491

Table 3: Mean Observed Scores for BARS

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Ratee Average
Ratee 1	5.62	4.67	7.19	4.53	6.67	4.67	5.56
Ratee 2	4.99	4.76	7.03	5.47	4.87	4.20	5.22
Ratee 3	6.07	7.03	4.49	2.39	6.14	5.94	5.34
Ratee 4	6.86	4.88	7.14	6.82	4.74	7.26	6.28
Ratee 5	4.47	5.95	6.11	2.76	2.78	4.39	4.41
Ratee 6	6.97	4.44	4.56	8.06	7.31	3.31	5.77
Ratee 7	7.06	7.18	7.09	5.65	3.60	6.06	6.11
Ratee 8	4.56	4.26	4.08	3.36	6.49	3.29	4.34
Ratee 9	6.72	6.33	5.14	7.35	4.91	5.70	6.02
Ratee 10	8.13	8.25	8.41	8.06	8.22	8.20	8.21
Dim. Average	6.15	5.77	6.12	5.44	5.57	5.30	*5.73

Note: The value marked with "*" represents the average of all observed scores

Table 4: Mean Observed Scores for Graphic Scales with Numerical Anchors

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Ratee Average
Ratee 1	6.54	5.82	7.34	6.13	7.49	4.93	6.73
Ratee 2	4.91	4.11	6.98	5.60	5.15	4.24	5.17
Ratee 3	6.18	6.71	4.29	3.05	6.16	5.73	5.34
Ratee 4	7.46	6.45	7.01	6.89	6.45	8.05	7.05
Ratee 5	4.57	6.12	6.68	3.83	3.18	4.23	4.77
Ratee 6	7.72	4.47	5.11	8.30	7.49	3.61	6.12
Ratee 7	7.50	7.43	7.63	7.42	7.35	6.36	7.28
Ratee 8	4.64	4.48	5.17	3.53	5.77	3.39	4.50
Ratee 9	6.91	6.23	4.90	7.22	5.50	5.23	6.00
Ratee 10	8.09	8.25	8.20	8.38	8.23	8.43	8.27
Dim. Average	6.45	6.01	6.33	6.04	6.27	5.42	*6.09

Note: The value marked with "*" represents the average of all observed scores

Table 5: Mean Observed Scores for Graphic Scales with No Anchors

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Ratee Average
Ratee 1	6.33	4.94	6.86	5.95	7.11	4.34	5.92
Ratee 2	4.03	3.65	6.73	5.43	3.97	3.80	4.60
Ratee 3	6.42	6.93	5.15	2.45	6.96	5.63	5.59

Ratee 4	7.37	5.65	6.47	6.54	5.30	7.69	6.50
Ratee 5	4.30	5.19	6.03	2.89	2.90	4.53	4.31
Ratee 6	7.07	3.97	4.64	7.99	7.17	2.81	5.61
Ratee 7	7.26	6.65	6.87	6.91	6.58	5.56	6.64
Ratee 8	3.63	4.04	4.21	3.12	6.11	3.18	4.05
Ratee 9	6.58	5.06	4.32	7.14	4.60	4.71	5.40
Ratee 10	8.22	8.04	8.24	8.43	8.16	8.29	8.23
Dim. Average	6.12	5.41	5.95	5.69	5.89	5.05	*5.69

Note: The value marked with "*" represents the average of all observed scores

Table 6: Correlations Among Performance Dimensions

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6
Dim. 1	1.00					
Dim. 2	.514	1.00				
Dim. 3	.312	.378	1.00			
Dim. 4	.573	.238	.313	1.00		
Dim. 5	.478	.302	.214	.395	1.00	
Dim. 6	.504	.592	.416	.319	.276	1.00

Note: All correlations are significant at the 0.05 level

Regression equations were used to test the effects of rating format on the various measures of rating accuracy. The "model summary" for each of these regression equations was examined for all four measures of accuracy. They revealed the overall effect that rating format has on the various measures of rating accuracy (see Tables 7 and 8). For the effects of rating format on the various measures of rating accuracy, elevation had an R-square of .137, differential elevation had an R-square of .011, stereotype accuracy had an R-square of .099, and differential accuracy had an R-square of .007 (see Table 7). Table 8 presents descriptive statistics for the various measures of accuracy for each of the three rating formats.

Table 7: Analysis of Rating Format's Effect on Measures of Accuracy

Type of Accuracy	R-Square	MS Between Groups	MS Within Groups	F-value	Prob > F
Elevation	.137	3.892	.494	7.881	.001
Differential Elevation	.011	110.54	71.73	1.541	2.19
Stereotype Accuracy	.099	13.43	2.47	5.445	.006
Differential Accuracy	.007	1373.52	4027.61	.341	.712

Table 8: Analysis of Rating Format's Effects on Measures of Accuracy

Type of Accuracy	Format	N	Mean	Standard Deviation
Elevation	1	35	.401	.341
	2	33	.756	.553
	3	34	.429	.320
Differential Elevation	1	35	.875	.351
	2	33	1.023	.368
	3	34	.869	.299
Stereotype Accuracy	1	35	.862	.164
	2	33	.743	.177
	3	34	.758	.152
Differential Accuracy	1	35	.288	.288
	2	33	.206	.206
	3	34	.226	.226

Note: Format1 = BARS, Format 2 = GRS with Numerical Anchors, Format 3 = GRS with No Anchors

The hypotheses predicted that certain rating formats would yield higher measures of accuracy than others would. The first hypothesis was that graphic scales with no anchors would be more accurate than graphic scales with numerical anchors. A series of dummy-coded regression equations were used to test if there were significant mean differences in the four measures of rating accuracy between the two rating formats. Table 9 provides a comparison of the two formats. For the elevation measure of accuracy, graphic rating scales with no anchors were significantly more accurate than graphic rating scales with numerical anchors, $t(65) = 3.228$, $p < .05$. I followed Cohen's (1992) guidelines for computing effect sizes for the comparisons of independent means. The effect size for the elevation measure of this comparison was .742. For differential elevation, graphic scales with no anchors were also significantly more accurate than graphic scales with numerical anchors, $t(65) = 1.843$, $p < .05$. The comparison yielded a small to medium effect of .445. In addition, there were no significant mean differences present either for measures of stereotype accuracy, $t(65) = .378$, $p > .05$, or for measures of differential accuracy, $t(65) = .231$, $p > .05$. The effect for stereotype accuracy was .089 , and the effect for differential accuracy was very small, .057. Since graphic scales with no anchors were more accurate than graphic scales with numerical anchors only for elevation and for differential elevation, hypothesis 1 was only partially supported.

Table 9: A Comparison of GRS with No Anchors to GRS with Numerical Anchors

Accuracy Measure	[a]t-value	Effect Size
Elevation	*3.228	.742
Differential Elevation	*1.843	.445
Stereotype Accuracy	.378	.089
Differential Accuracy	.231	.057

Note. N=67

[a]All hypotheses tested using one-tailed tests of significance

*$p < .05$

Hypothesis 2 stated that graphic scales with no anchors would yield higher levels of accuracy than BARS. Dummy-coded regression equations were employed to test this hypothesis also. A comparison of the two formats can be found in Table 10. No significant mean differences in elevation were found between the two different rating formats, $t(67) = .278$, $p > .05$. The effect size for the comparison was very small, .063. Also, for measures of differential elevation, there were no significant mean differences, $t(67) = .070$, $p > .05$. Only a small effect of .017 was present for differential elevation. Analysis revealed that graphic scales with no anchors did result in higher levels of stereotype accuracy than BARS, $t(67) = 2.628$, $p < .05$. The effect of .607 was of medium to large size. However, there were no differences between the formats for measures of differential accuracy, $t(67) = .68$, $p > .05$, and the effect size was a small .165. Since graphic scales with no anchors were more accurate than BARS only for stereotype accuracy, hypothesis 2 was only partially supported.

Table 10: A Comparison of GRS with No Anchors to BARS

Accuracy Measure	[a]t-value	Effect Size
Elevation	.278	.063
Differential Elevation	.070	.017
Stereotype Accuracy	*2.628	.607
Differential Accuracy	.680	.165

Note. N=69
[a]All hypotheses tested using one-tailed tests of significance
*$p < .05$

Hypothesis 3 predicted that graphic scales with numerical anchors would be more accurate than BARS. As seen in Table 11, subjects using graphic scales with numerical anchors were more accurate, as measured by elevation, than ratings obtained using BARS, $t(66) = 3.526$, $p < .05$. There was a large effect size of .805 for elevation. For measures of stereotype accuracy, graphic scales with numerical anchors were more accurate than BARS, $t(66) = 2.989$, $p < .05$. Also, there was a medium to large effect of .701. However, BARS were more accurate than graphic scales with numerical anchors when measuring differential elevation, $t(66) = 1.787$, $p < .05$. For differential elevation, there was a medium-sized effect of .429. Although graphic scales with numerical anchors were more accurate than BARS according to measures of differential accuracy, the differences were not significant, $t = .442$, $p > .05$. Differential accuracy had only a small effect size of .108. Although graphic scales were slightly more accurate than BARS for differential accuracy, graphic scales were significantly more accurate only for measures of stereotype accuracy. Therefore, hypothesis 3 was only partially supported.

Table 11: A Comparison of GRS with Numerical Anchors to BARS

Accuracy Measure	[a]t-value	Effect Size
Elevation	*3.526	.805
Differential Elevation	*1.787	.429

Stereotype Accuracy	*2.989	.701
Differential Accuracy	.442	.108

Note. N=68
[a]All hypotheses tested using one-tailed tests of significance
*p < .05

In order to test hypothesis 4, interrater reliability estimates were calculated separately for each rating format. Within each format, pairwise comparisons were made between each subject's responses and the responses of all of the other subjects. Each of these pairwise comparisons yielded a correlation representing the reliability between the two subjects for the rendered performance judgements. The average of these correlations yielded a measure of interrater reliability for each rating format. Descriptive statistics for the reliability estimates can be found in Table 12. Subjects using the BARS had an average reliability of .4934 with a standard deviation of .1498. Graphic scales with numerical anchors yielded a reliability of .5378 and a standard deviation of .1369. Subjects using graphic scales with no anchors had the highest level of interrater reliability with .6071 and a standard deviation of .1018. As predicted in hypothesis 4, graphic scales with no anchors were more reliable than both of the other rating formats. In addition to being more reliable, graphic scales with no anchors also had the lowest amount of variability in the distribution of correlations.

Table 12: Comparisons of Interrater Reliability Estimates

	Mean	SD	Range	Skewness
GRS w/ No Anchors	.6071	.1018	.76	-.258
GRS w/ Numerical Anchors	.5378	.1369	.68	-.300
BARS	.4934	.1497	.84	-.432

I also conducted analyses on the follow-up questions (see Appendix D). An analysis of variance was performed on all of the follow-up measures to determine if rating format affected the way in which subjects responded. The results for the tests of significance can be found in Table 13, and the means and standard deviations are located in Table 14.

Table 13: Analysis of Rating Format's Effect on Follow-up Questions

Question	MS Between Groups	MS Within Groups	F-value	Prob > F
1	10.416	2.343	4.446	.014
2	6.060	2.147	2.823	.064
3	5.591	2.178	2.576	.082
4	16.964	2.770	6.123	.003
5	2.411	1.593	1.513	.225
6	4.263	2.978	1.431	.244

Table 14: Statistics for Follow-up Questions Based on Rating Format

Question	Format	N	Mean	SD
1	1	35	4.43	1.63
	2	33	5.30	1.38
	3	34	5.44	1.56
2	1	35	4.86	1.59
	2	33	5.61	1.39
	3	34	5.56	1.40
3	1	35	4.40	1.74
	2	33	5.15	1.35
	3	34	5.03	1.29
4	1	35	4.11	1.83
	2	33	5.30	1.57
	3	34	5.35	1.57
5	1	35	5.77	1.59
	2	33	6.18	1.07
	3	34	6.26	1.02
6	1	35	3.86	1.73
	2	33	4.42	1.79
	3	34	4.50	1.66

Note: Format 1 = BARS, Format 2 = GRS w/ no anchors, Format 3 = GRS w/ numerical anchors

According to the analysis, format had a significant main effect on how comfortable the subjects felt using the rating scale, $F(2, 99) = 4.446$, $p < .05$. A Bonferroni's test was used post hoc to determine if there were significant mean differences between the different formats. Table 15 illustrates the comparisons of mean scores on question one for each of the rating formats. There were no significant differences between the means for BARS vs. GRS with numerical anchors or for GRS with no anchors vs. GRS with numerical anchors. However, subjects reported significantly greater liking for the GRS with no anchors compared to the BARS.

Table 15: Comparisons of Mean Responses for Follow-up Questions w/ Sig. Effects

	Format (I)	Format (J)	Mean Difference (I-J)	Std. Error	Prob > F
Question1	1	2	-.87	.371	.062
	1	3	-1.01	.369	.021
	2	3	-.14	.374	1.00
Question 4	1	2	-1.19	.404	.012
	1	3	-1.24	.401	.008
	2	3	-.0499	.407	.407

Note: Format 1 = BARS, Format 2 = GRS w/ numerical anchors, Format 3 = GRS w/ no anchors

Analysis of variance was also conducted to reveal that format does not have a significant main effect on question two, $F(2, 99) = 2.823$, $p > .05$, the subject's assessment of how clearly the scenarios describe teaching behavior. Since there was no main effect, no post-hoc tests were performed on the data.

Question three measured the extent to which the subjects felt the scale they used allowed and accurate assessment of the teachers' performance. Analysis of variance revealed that there is no significant main effect of rating format on the results of question three, $F(2, 99) = 2.567$, $p > .05$.

In question four, subjects were asked how well they liked the scale they used compared with others they used in the past. Analysis of variance demonstrated that there was a significant main effect in question four due to rating format, $F(2, 99) = 6.123$, $p < .05$. A Bonferroni's test was performed post hoc to test for significant mean differences in response based on rating format. Table 15 illustrates the comparisons of mean scores for each of the rating formats. Subjects significantly preferred both the GRS with numerical anchors and the GRS with no anchors to the BARS. There were no significant mean differences in preference between the GRS with numerical anchors and the GRS with no anchors.

The next question, question five, measured how true to life the subject felt the teaching scenarios were. The ANOVA showed that there was no main effect of rating format on responses to question five, $F(2, 99) = 1.513$, $p > .05$.

The final question measured the extent to which the subjects felt they had enough information to make an accurate judgement about the teachers' levels of performance. Analysis revealed no significant main effects of format on responses to the question, $F(2, 99) = 1.431$.

CHAPTER 6: DISCUSSION

The results of this study provide partial support for the hypotheses presented. Three of the four measures of accuracy had a significant portion of their variance attributable to format effects. For elevation and differential elevation, graphic rating scales with no anchors were superior to graphic rating scales with numerical anchors, providing support for hypothesis one. For stereotype accuracy, graphic scales of both types were superior to BARS. This provides partial support for hypotheses 2 and 3. Hypothesis 4 concerning interrater reliability was supported in that those using graphic rating scales without anchors provided more reliable evaluations.

In general, the results indicate that format potentially affects rating accuracy. For both differential elevation and differential accuracy, format does not predict a significant amount of variance in ratings. However, rating format accounts for 14% of the variance in elevation and 10% of the variance in stereotype accuracy. Although there is a large portion of variance for both of these measures of accuracy that is not accounted for by rating format, these results should not be overlooked. There are numerous factors that can affect the assignment of performance ratings to a ratee. For one variable to account for 10 or 14 percent of the variance is an important finding. Taken as a whole, the results show that rating format can affect the reliability and accuracy of performance ratings.

Earlier in the paper, it was noted that more specific anchors might bias the rater and prevent accurate ratings. Murphy and Constans (1987) noted that anchors can bias ratings if a ratee exhibited a specific behavior that was anchored in the scale, but this one behavior was not altogether indicative of the ratees overall performance. Their argument was on a very specific level. However, this paper extended their work and tested the possibility of anchor bias on a more general level. Specifically, previous research (Murphy & Constans, 1987) indicated that behavioral anchors could negatively affect the accuracy of performance ratings. Our study provided mixed results for this idea. For most accuracy measures, BARS did not produce a lowered level of rating accuracy. Yet, for stereotype accuracy, subjects using the BARS generated less accurate ratings.

Stereotype accuracy collapses across ratees to assess how a rater rates on each dimension. It is possible that the specific nature of the behavioral anchors provided more information about the performance dimension than was needed by the rater. This may cause the confusion to which Murphy & Constans (1987) refer. The specific nature of the anchor provides the rater with more information than is necessary to make an accurate judge about ratees' levels of performance on each dimension. This idea is not new. Barrett et al. (1958) noted that too much information could be contained in scale anchors.

The idea that BARS results in lower stereotype accuracy provides even more support for Murphy and Constans (1987) because stereotype accuracy collapses across ratees, it indicates how accurately a raters judge each dimension. The results indicate that raters, in general, have a tougher time of rating accurately on a dimension when using BARS than when using the other types of scales. The more specific anchors appear to lead to more difficulty in correctly interpreting the nature of the performance dimensions.

This is the basic argument that Murphy & Constans (1987) made. Behavioral anchors can have a biasing effect on performance ratings and, possibly, lead to more inaccurate ratings.

The results for elevation measures of accuracy are more difficult to explain. Graphic scales with no anchors yielded more accurate ratings than graphic scales with numerical anchors. At first blush, this appears to fall directly in line with our hypotheses and support Murphy & Constans' (1987) idea that more specific anchors can have a biasing effect and, possibly, lead to decreased measures of accuracy. However, analysis also showed that BARS had higher levels of accuracy, as measured by elevation, than did graphic scales with numerical anchors. Murphy & Constans' explanation clearly does not work here. Perhaps a suitable explanation can come from previous rating format research. Elevation is simply a comparison of the mean of observed scores to the mean of the true scores. It essentially gives a measure of how lenient/severe a rater is compared to the true score. Previous research (Borman & Dunnette, 1975; Campbell et al., 1973; Kingstrom & Bass, 1981) has noted that BARS can result in lowered levels of leniency error compared to graphic rating scales. This could explain why the BARS in this study resulted in higher accuracy measures of elevation compared to the graphic scales with numerical anchors. The superiority of BARS over graphical scales with no anchors could be due to a different reason, however. Previous research has clearly documented the reluctance of raters assigning low numbers for performance ratings. BARS may contain enough specific behavioral information within the scales that raters match observed behavior with the anchors in the scales. In this way, the BARS may prevent lenient ratings. However, graphics scales with no anchors provide so little information and feedback to raters, they might not feel as if they are assigning low ratings. These combined factors could have led to the superiority of BARS over graphic scales with no anchors.

The results for measures of differential elevation are quite interesting as well. BARS were significantly more accurate than graphic scales with numerical anchors. This could be due to the nature of the behavioral anchors. They could provide enough information so that raters avoid their tendency to be lenient with their ratings and to not distinguish among the ratees' performance levels. However, the raters using graphic scales with numerical anchors fall prey to the well-documented tendency to be lenient with their performance evaluations and to not distinguish between ratees and their levels of performance. In contrast, graphic scales with no anchors were also superior to graphic scales with numerical anchors but for a slightly different reason. It's possible that the graphic scales with no anchors removes the rating tendency to use high numbers and to rate leniently. The removal of this bias would lead to greater distinction among the performance of the ratees compared to the graphic scales with numerical anchors. This, in turn, could have led to the increased level of differential elevation accuracy in graphic scales with no anchors compared to those with numerical anchors.

The interrater reliability hypothesis proposed was supported in that graphic scales with no anchors were more reliable than BARS and also more reliable than graphic scales with numerical anchors. In addition to increased reliability for subjects using graphic

scales with no anchors, there was also a lower standard deviation. Figures 1, 2, 3, and 4 provide a graphic representation for the reliability results. As indicated by the frequency distributions, not only does the mean of the reliability scores increase, but the "tightness" of the plot increases as well. This is indicative of more scores residing closer to the mean (i.e., a lower standard deviation). A reliability estimate of .6071 for graphic scales with no anchors is quite an interesting finding. Viswesvaran, Ones, & Schmidt (1996) conducted a meta-analysis on reliability estimates in performance appraisals. They found that reliabilities range from .50-.54. Based on these numbers, the findings of this study lend support to the idea that the graphic scale with no anchors can be a more reliable way to measure performance than traditionally seen.

Figure 1

Reliability for BARS

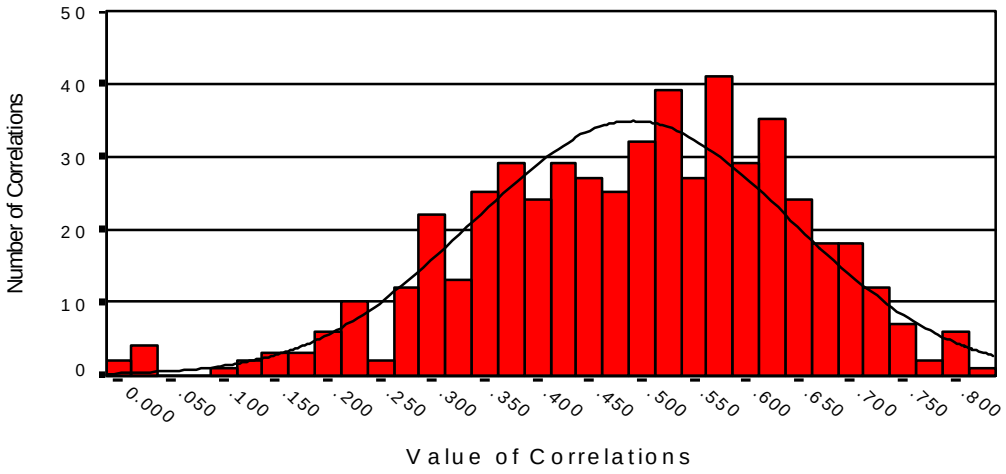


Figure 2

Reliability for GRS w/ Numerical Anchors

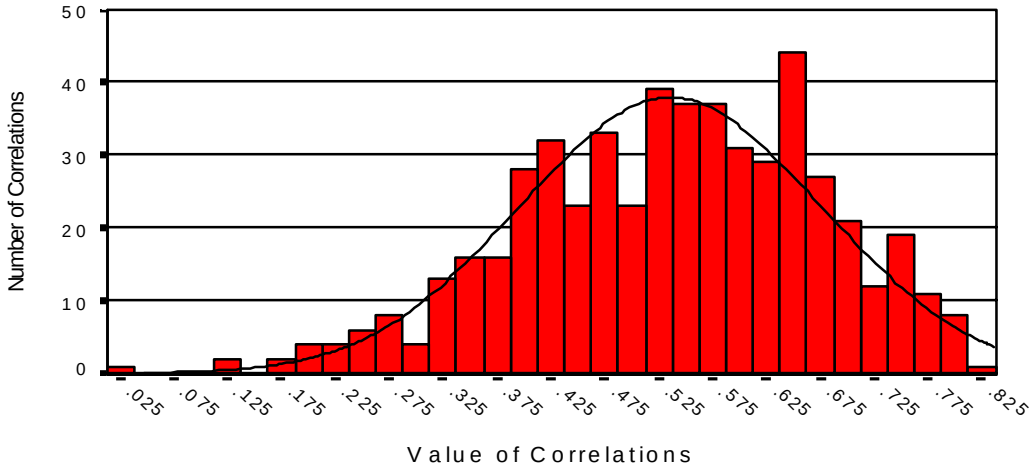
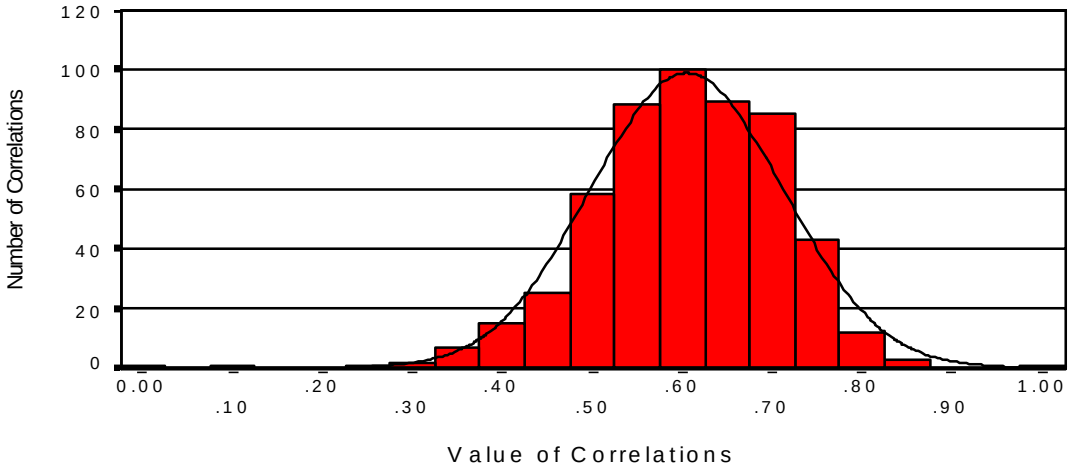


Figure 3

Reliability for GRS w/ No Anchors



At this point, we must be careful with any conclusions we make. It was mentioned earlier in this paper that this study was not a comparison of rating formats. It would be unfair to make conclusions about the effectiveness of BARS or graphic rating scales since these rating formats are not, traditionally, continuous rating scales. Rather, since all scales were continuous, this study was more a comparison of *anchor specificity*. Based on the results of this study, one could conclude that anchor specificity can influence measures of rating accuracy, but more specific anchors do not necessarily lead to more accurate performance ratings. It could be possible to then extend a generalization to the traditional forms of the various rating scales. One could draw the conclusion that, based on the data presented here, rating format affect measures of rating accuracy with the caveat that this study modified the "traditional" versions of these rating scales.

The data collected in this study do offer some support for the idea that rating format affects measures of rating accuracy. There was a general pattern to the data that suggested that graphic scales with no anchors might be valuable for evaluating performance. For differential elevation and differential accuracy, graphic scales with no anchors were the most accurate method of rating. For elevation and stereotype accuracy, they were the second most accurate. Also, graphic scales with no anchors proved to have the highest measures of interrater reliability of the three rating formats. The reliability estimates were also higher than the "average" range established by previous research (Viswesvaran et al., 1996). In addition, the graphic scales with no anchors format received good ratings on the follow-up questions. Subjects liked the graphic scales without anchors just as much as the graphic scales with numerical anchors and significantly more than the BARS. Also, subjects felt significantly more comfortable using the graphic scales with no anchors than both of the other formats. This is quite an interesting finding given the lack of exposure subjects have had to this particular rating format. Taking all of these factors into consideration, a strong case can be made for the use of graphic scales with no anchors for computer-based performance appraisal systems.

Conclusions

Clearly, the data showed that graphic scales with no anchors could be a potentially valuable tool for accurately and appropriately evaluating employee performance. However, it would be a mistake to draw a definite conclusion about rating format's effect on accuracy on the basis of one study. As previously mentioned, this study is delving into an old line of research but with a different perspective. There are few, if any, studies in the literature that examine effects of rating format on rating accuracy. Subsequent research should also investigate this problem but try to correct for some of the limitations of this study.

The first limitation is the general problem of using college students, with little vested interest, as raters. Of the different rating formats, the BARS is the most complex and hardest to use. As such, an interaction between the BARS complexity and the low motivation level of subjects may account for the findings that the BARS were less accurate than expected and that reaction to the BARS scale was the least favorable.

A second limitation with this study is the manner in which true scores for behavioral incidents were operationalized. The behavioral incidents were developed in accordance to the procedures outlined by Smith & Kendall (1963) and summarized by Landy & Barnes (1979). The values for the behavioral incidents were used as the "true scores" of performance. When assigning values to these incidents in the BARS development procedure, these incidents are treated as independent. They are evaluated and assigned values by themselves. However, these single behavioral incidents were combined with other incidents in order to construct a comprehensive scenario that gives an overall picture of a fictional ratee's performance-related behaviors. These observed values of these single behavioral incidents could possibly be affected or influenced by the surrounding behavioral incidents. Although this may have some affect on the findings, the dimensions were not correlated as highly as typically seen in performance appraisal research (See Table 6).

Another limitation drawback to this particular study is its simple design. Because this topic was one that has received little attention, this study was designed, purposely, to be a simple analysis of the influence of rating format on accuracy. Subsequent research should continue to investigate the nature of the format/accuracy relationship (or the lack thereof), but taking into account more factors. This means investigating potential moderators and/or mediators of the relationship. It also means that, if format does influence rating accuracy, we should strive to discover why and how the process occurs.

Previously in the paper, it was mentioned that we need to help raters to do a better job of rating. It was assumed that one step in this direction would be to provide raters with continuous scales so they could discriminate between levels of performance as finely as they desired. However, there was not a test of continuous vs. forced-choice formats. This is an area that should demand attention of researchers. Primary difficulty with such a task is to allow a methodologically fair comparison of the two different types

of scales. True scores that lie between anchors, or integers, "stack the deck" in favor of continuous scales because a rater using forced-choice scales could never match the true score with the proper observed score. Similarly, true scores that fall on an anchor, or integer, favor forced-choice formats. Because their choices are very limited, a rater using a forced-choice scale has a much better chance of recording an observed score that is congruent with the true score. However, this line of research is necessary and should demand greater attention.

In conclusion, this study has revisited an old line of research with a new perspective. Examining accuracy based on rating format is not simply "old wine in a new bottle." The effects of rating format have been largely unexplored. Current research (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993) is focusing on rating accuracy as the most important facet of performance ratings. This study follows in the footsteps of previous work conducted on rating format. However, the focus on accuracy is novel approach and should be continued in the literature. This study could serve as a springboard for future research in the area of rating accuracy.

This study has moved research forward by fully integrating current technology. By using modern computers and programs, we can explore new and different methods of rating, rating procedures, and the entire rating process. This study showed that it is possible to integrate traditional rating methods, such as BARS and graphic rating scales, into a form that can be accessed with current technology. By using computer, we were also able to measure performance ratings to a level of precision that has not been reached before. This study serves as a benchmark for future research. Not only can computers move research ahead and allow scientists to study performance appraisals as never before, but they also allow huge gains in our ability to precisely, and accurately, measure variables of interest.

It appears that performance appraisal research has come full circle. This study follows the footsteps of the multitude of research prior to 1980 that examined the effect of rating format on the quality of performance appraisals. However, unlike past research, this particular study seems to have generated some clear results. In line with Murphy and Constans' (1987) argument, it appears that decreased anchor specificity can affect the accuracy of performance evaluations. Also, since BARS had the lowest estimates of reliability and graphic scales with no anchors had the highest, one could conclude that decreased anchor specificity can also affect interrater reliability of performance evaluations. From the data gathered in this study, a strong case can be made for the future importance of graphic scales with no anchors in computer-based performance evaluations.

One implication from conclusions drawn along these lines is that organizations can benefit from simpler, less involved rating procedures. Follow-up questions one and four assessed how well the subjects liked the rating scale compared to other scales they have used in the past and also how comfortable the subjects felt using their particular rating scale. Subjects using the BARS liked their format significantly less than did subjects using either form of the graphic rating scales, and they felt significantly less

comfortable in using the BARS than they did using either of the graphic rating scales. If people feel more comfort and preference for simpler rating scales, organizations can save millions of dollars per year in development and administration of performance management systems.

Another implication of the data is that some other variables significantly affect the manner in which we rate performance. Previous rating format research has been conducted investigating the influence of variables such as education and job experience (Cascio & Valenzi, 1977), purpose for appraisal (Bernardin & Orban, 1990), sex, and race (Hamner, Kim, Lloyd, & Bigoness, 1974; Schmitt & Hill, 1977). However, these studies did not focus on measures of rating accuracy as the primary criterion to determine the "quality" of the rating formats. These variables should all be re-examined with a critical eye on measures of rating accuracy.

In conclusion, Landy & Farr (1980) were wrong to call for a moratorium on rating format research. Instead, they should have requested a shift in the variable of interest. No longer are we, as a field, interested in halo, or leniency. Instead, we are concerned with rating accuracy as measured by elevation, differential elevation, stereotype accuracy, and differential accuracy. The goal is to increase rating reliability and accuracy. This study supports the idea that rating format can potentially affect the level of rating accuracy. At this point, graphic rating scales without anchors appear to hold promise for computer-based performance appraisals. They can potentially promote increased levels of rating accuracy in raters. They can also yield more reliable results. In addition, raters appear to like these simple scales and appear to be comfortable in using them. Until enough data is accumulated to draw these conclusions, however, the effects of rating format on the various measures of rating accuracy should occupy a large and important place in the literature.

Appendix A

BARS Scales

Instructions

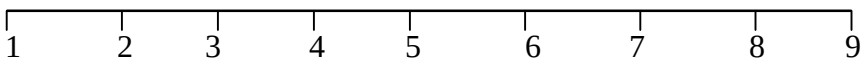
Given below are the instructions received by the subjects placed into the experimental group that rated performance using BARS scales. Please note that the scale was presented horizontally instead of vertically. Also, the behavioral anchors are not placed along side the scale as they were in the experimental condition. Rather, the anchors and their corresponding values are, instead, listed below the scale. Both of these differences are intended to conserve space in the Appendix.

Thank you for your participation in this experiment. Your involvement in this project is critical to its success. This entire project will be completed on this computer. However, if at any time you have questions, please feel free to ask the experimenter. Your task is as follows: you will be presented with a series of 10 vignettes, or scenarios, that depict a list of behaviors that a specific teacher exhibits in the classroom. Following each behavioral list is a series of 6 questions. The questions will measure how effective you thought that particular teacher was in a specific area, or dimension, of teaching performance. For example, read the sample behavioral list below:

> Teacher X is a math teacher here at Virginia Tech.
> Regardless of X's personal conflicts and obligations, X is always present for class.
> Often times, X comes to class without the proper transparencies for the overhead projector.
> On occasion, X has forgot to bring the lecture notes for the day.
> X has a tendency to speak in mean or overly harsh tones toward his students.
> X often has difficulty explaining concepts to the class in terms that the students can easily understand and relate to.
> Even though X teaches a math class, X rarely reviews how to achieve the correct answers on homework problems.

If this was a real vignette instead of one for practice, you would then answer a set of questions that ask you rate X's effectiveness on a particular dimension of teaching performance. Take, for example, the following question:

> On a scale of 1-9, how would you rate X's teaching performance in the area of classroom organization and preparation?



anchors: 2.11—teacher repeatedly forgets to bring necessary materials to class; 3.89—teacher deviates from preciously planned activity; 7.68—teacher arrives on-time for class

To answer this question, you would consider the information presented in the vignette about X's teaching behaviors that are involved with classroom organization and participation. You would then move the mouse and click anywhere on the response line that best represents your evaluation of X's performance in the area of classroom organization and performance. A set of sample behaviors and their values are placed next to the answer line. The values associated with each behavior represent the correct, or "true" score. These behavioral anchors serve as a comparison, or guide, when making your evaluation about the behaviors illustrated in the scenario. Remember, you can answer anywhere along the answer line you see fit, and your answers are not restricted to only the anchors. A valid response can be given at any point between any of the anchors. Please note that "1" represents the lowest value of teaching performance and "9" represents the highest teaching rating.

You would then proceed to the next question that asks for your judgement about a different dimension of teaching performance. Answer all six questions about a particular teacher, and then please proceed to the next teaching vignette that contains a behavioral list of a different teacher. Please note that while evaluating a specific teacher, you can change your ratings for each teaching dimension as many times as you wish. However, once you proceed to the next behavioral list, you cannot return to a previous behavioral list to change your answers.

You will read each behavioral list and answer each of the 6 questions about that particular teacher. Then, proceed to the following behavioral list and do the same. Continue in this manner until all six questions have been answered for each of the 10 vignettes. At the end, there will be some questions for you to answer about this experiment. Once all of those questions are answered, the computer will inform you that the experiment is over. At that time, please notify the experimenter. If you have any questions about any part of this experiment, please ask the experimenter before proceeding. Should any questions arise throughout the course of the experiment, please feel free to ask the experimenter for assistance. At this time, please proceed to the first scenario.

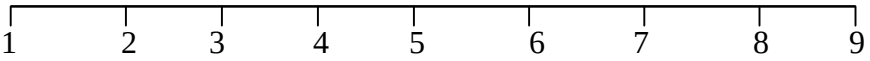
Sample Vignette

Below, a sample vignette is given. Space prevents inclusion of all ten vignettes from the appraisal instrument, so only one is included. In this manner, the reader can still get a sense of the instrument used. The entire appraisal instrument can be obtained upon request. In this sample, the behavioral list, or vignette, is followed by the questions regarding performance on the separate performance dimensions and by the scale to respond to each question. To conserve space in the appendix, the scales are presented horizontally rather than vertically. Also, the behavioral anchors and their respective values are listed below the scale rather than placed in their proper place next to the scale continuum. Again, this was to conserve space for the appendix.

Behavior list for teacher “A”

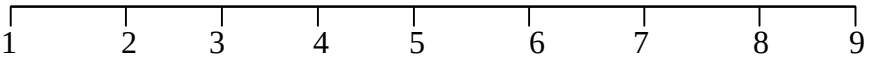
- “A’s” class is rather large. However, “A” always has the students’ tests graded and returned to the students within a week of taking the exam.
- Before every exam, “A” prepares a study guide that outlines the relevant material for the test.
- “A” is very knowledgeable about the current literature that pertains to the class and is able to answer questions about the material.
- “A” sometimes forgets lecture notes and has to “wing it” in class lectures.
- “A” speaks very quickly, making it difficult for students to keep pace while taking notes.
- “A” often arrives late for class but expects the students to stay late after class.

On a scale of 1 to 9, how would you rate A’s performance in the area of teacher dedication?



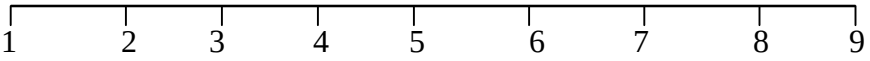
anchors:1.64—teacher is frequently late to class and occasionally misses class with no previous warning; 6.39—teacher informs students about their personal research interests; 7.68—even though the class is large, teacher tries to learn all students’ names so as to make the class more personal

On a scale of 1 to 9, how would you rate A’s performance in the area of classroom organization and preparation?



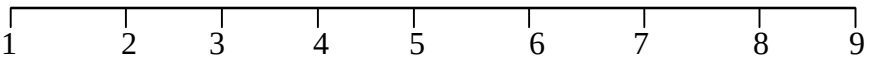
anchors: 2.11—teacher repeatedly forgets to bring necessary materials to class; 3.89—teacher deviates from preciously planned class activity; 7.68—teacher arrives on-time for class

On a scale of 1 to 9, how would you rate A’s performance in the area of teacher expertise?



anchors: 2.25—teacher has difficulty understanding the students’ questions and has difficulty answering them in class; 7.86—teacher performs experiments with the students in class to illustrate points; 8.39—teacher is well-versed and up to date on current research and literature that relates to the class

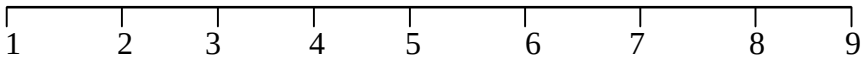
On a scale of 1 to 9, how would you rate A’s performance in the area of courtesy and respect for students?



anchors: 1.21—teacher makes fun of a student’s appearance in front of the class; 3.57—regardless of the excuse, teacher will not accept late papers/ assignments;

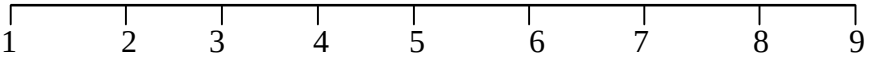
7.00—teacher praises students for class participation regardless of the quality of the comments

On a scale of 1 to 9, how would you rate A's performance in the area of adequately preparing students for exams?



anchors:1.43—teacher designs the test to be so tough that teacher cannot adequately explain the rationale behind the correct test answers; 2.71—teacher has several trick questions on tests: there seem to be multiple correct answers on one question; 8.25—teacher conducts a review session a few days before the test

On a scale of 1 to 9, how would you rate A's performance in the area of classroom delivery and presentation?



anchors: 1.70—while lecturing, the teacher tolerates extraneous conversations among the students in class which, in turn, contributes to the overall noise level in class, making it difficult to hear the teacher; 6.82—teacher always dresses "professionally" for class; 8.07—teacher uses real-life examples to clarify a point

Appendix B

Graphic Scales with Numerical Anchors

Instructions

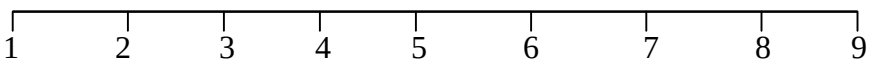
Below are the instructions given to the subjects who were assigned to the experimental condition where ratings were made using a graphic scale with numerical anchors.

Thank you for your participation in this experiment. Your involvement in this project is critical to its success. This entire project will be completed on this computer. However, if at any time you have questions, please feel free to ask the experimenter. Your task is as follows: you will be presented with a series of 10 vignettes, or scenarios, that depict a list of behaviors that a specific teacher exhibits in the classroom. Following each behavioral list is a series of 6 questions. The questions will measure how effective you thought that particular teacher was in a specific area, or dimension, of teaching performance. For example, read the sample behavioral list below:

> Teacher X is a math teacher here at Virginia Tech.
> Regardless of X's personal conflicts and obligations, X is always present for class.
> Often times, X comes to class without the proper transparencies for the overhead projector.
> On occasion, X has forgot to bring the lecture notes for the day.
> X has a tendency to speak in mean or overly harsh tones toward his students.
> X often has difficulty explaining concepts to the class in terms that the students can easily understand and relate to.
> Even though X teaches a math class, X rarely reviews how to achieve the correct answers on homework problems.

If this was a real vignette instead of one for practice, you would then answer a set of questions that ask you rate X's effectiveness on a particular dimension of teaching performance. Take, for example, the following question:

How would you rate X's teaching performance in the area of classroom organization and preparation?



To answer this question, you would consider the information presented in the vignette about X's teaching behaviors that are involved with classroom organization and participation. You would then move the mouse and click anywhere on the response line that best represents your evaluation of X's performance in the area of classroom organization and performance. You can

refer to the numbers as guides, or anchors, but your answers are not restricted to only the anchors. A valid response can be given at any point between any of the anchors. Please note that "1" represents the lowest value of teaching performance and "9" represents the highest teaching rating.

You would then proceed to the next question that asks for your judgement about a different dimension of teaching performance. Answer all six questions about a particular teacher, and then please proceed to the next teaching vignette that contains a behavioral list of a different teacher. Please note that while evaluating a specific teacher, you can change your ratings for each teaching dimension as many times as you wish. However, once you proceed to the next behavioral list, you cannot return to a previous behavioral list to change your answers.

You will read each behavioral list and answer each of the 6 questions about that particular teacher. Then, proceed to the following behavioral list and do the same. Continue in this manner until all six questions have been answered for each of the 10 vignettes. At the end, there will be some questions for you to answer about this experiment. Once all of those questions are answered, the computer will inform you that the experiment is over. At that time, please notify the experimenter. If you have any questions about any part of this experiment, please ask the experimenter before proceeding. Should any questions arise throughout the course of the experiment, please feel free to ask the experimenter for assistance. At this time, please proceed to the first scenario.

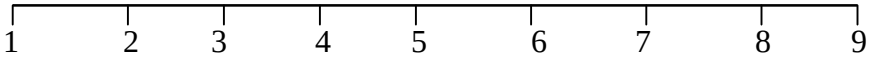
Sample Vignette

Below, a sample vignette is given. Space prevents inclusion of all ten vignettes from the appraisal instrument, so only one is included. In this manner, the reader can still get a sense of the instrument used. The entire appraisal instrument can be obtained upon request. In this sample, the behavioral list, or vignette, is followed by the questions regarding performance on the separate performance dimensions and by the scale to respond to each question.

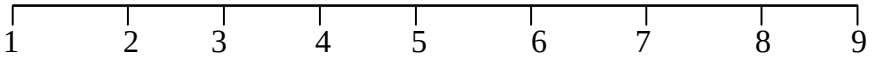
Behavior list for teacher "A"

- "A's" class is rather large. However, "A" always has the students' tests graded and returned to the students within a week of taking the exam.
- Before every exam, "A" prepares a study guide that outlines the relevant material for the test.
- "A" is very knowledgeable about the current literature that pertains to the class and is able to answer questions about the material.
- "A" sometimes forgets lecture notes and has to "wing it" in class lectures.
- "A" speaks very quickly, making it difficult for students to keep pace while taking notes.
- "A" often arrives late for class but expects the students to stay late after class.

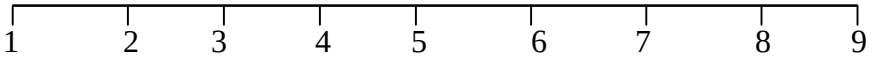
On a scale of 1 to 9, how would you rate A's performance in the area of teacher dedication?



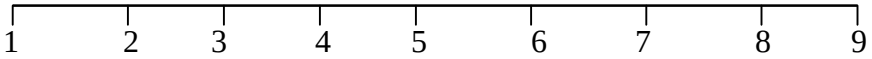
On a scale of 1 to 9, how would you rate A's performance in the area of classroom organization and preparation?



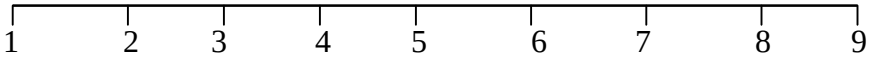
On a scale of 1 to 9, how would you rate A's performance in the area of teacher expertise?



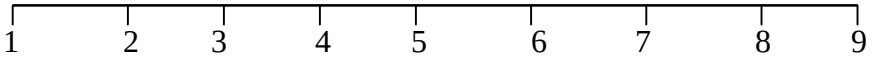
On a scale of 1 to 9, how would you rate A's performance in the area of courtesy and respect for students?



On a scale of 1 to 9, how would you rate A's performance in the area of adequately preparing students for exams?



On a scale of 1 to 9, how would you rate A's performance in the area of classroom delivery and presentation?



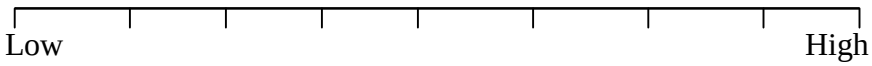
Appendix C

Graphic Scales Without Anchors

Instructions

Below are the instructions given to the subjects who were assigned to the experimental condition where ratings were made using a graphic scale without anchors.

> Thank you for your participation in this experiment. Your involvement in this project is critical to its success. This entire project will be completed on this computer. However, if at any time you have questions, please feel free to ask the experimenter. Your task is as follows: you will be presented with a series of 10 vignettes, or scenarios, that depict a list of behaviors that a specific teacher exhibits in the classroom. Following each behavioral list is a series of 6 questions. The questions will measure how effective you thought that particular teacher was in a specific area, or dimension, of teaching performance. For example, read the sample behavioral list below:
>
> Teacher X is a math teacher here at Virginia Tech.
> Regardless of X's personal conflicts and obligations, X is always present for class.
> Often times, X comes to class without the proper transparencies for the overhead projector.
> On occasion, X has forgot to bring the lecture notes for the day.
> X has a tendency to speak in mean or overly harsh tones toward his students.
> X often has difficulty explaining concepts to the class in terms that the students can easily understand and relate to.
> Even though X teaches a math class, X rarely reviews how to achieve the correct answers on homework problems.
>
> If this was a real vignette instead of one for practice, you would then answer a set of questions that ask you rate X's effectiveness on a particular dimension of teaching performance. Take, for example, the following question:
>
> How would you rate X's teaching performance in the area of classroom organization and preparation?



> To answer this question, you would consider the information presented in the vignette about X's teaching behaviors that are involved with classroom organization and participation. You would then move the mouse and click anywhere on the response line that best represents your evaluation of X's

performance in the area of classroom organization and performance. A valid response can be given at any point along the line. Please note that the extreme left end of the scale represents the lowest value of teaching performance and the extreme right end of the scale represents the highest teaching rating. There are no anchors or guides to assist you in your judgement. Simply realize that the better the level of performance is, the farther right you should mark on the scale. Likewise, the poorer the level of performance, the farther left you should rate performance on the scale.

You would then proceed to the next question that asks for your judgement about a different dimension of teaching performance. Answer all six questions about a particular teacher, and then please proceed to the next teaching vignette that contains a behavioral list of a different teacher. Please note that while evaluating a specific teacher, you can change your ratings for each teaching dimension as many times as you wish. However, once you proceed to the next behavioral list, you cannot return to a previous behavioral list to change your answers.

You will read each behavioral list and answer each of the 6 questions about that particular teacher. Then, proceed to the following behavioral list and do the same. Continue in this manner until all six questions have been answered for each of the 10 vignettes. At the end, there will be some questions for you to answer about this experiment. Once all of those questions are answered, the computer will inform you that the experiment is over. At that time, please notify the experimenter. If you have any questions about any part of this experiment, please ask the experimenter before proceeding. Should any questions arise throughout the course of the experiment, please feel free to ask the experimenter for assistance. At this time, please proceed to the first scenario.

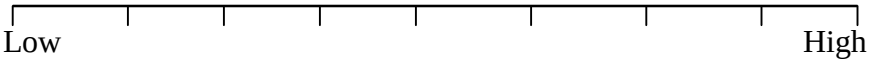
Sample Vignette

Below, a sample vignette is given. Space prevents inclusion of all ten vignettes from the appraisal instrument, so only one is included. In this manner, the reader can still get a sense of the instrument used. The entire appraisal instrument can be obtained upon request. In this sample, the behavioral list, or vignette, is followed by the questions regarding performance on the separate performance dimensions and by the scale to respond to each question.

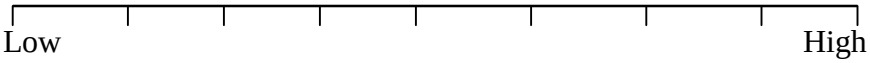
Behavior list for teacher "A"

- "A's" class is rather large. However, "A" always has the students' tests graded and returned to the students within a week of taking the exam.
- Before every exam, "A" prepares a study guide that outlines the relevant material for the test.
- "A" is very knowledgeable about the current literature that pertains to the class and is able to answer questions about the material.
- "A" sometimes forgets lecture notes and has to "wing it" in class lectures.
- "A" speaks very quickly, making it difficult for students to keep pace while taking notes.
- "A" often arrives late for class but expects the students to stay late after class.

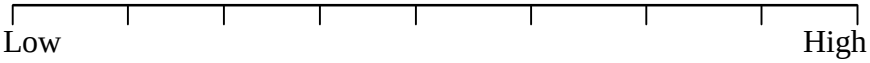
How would you rate A's performance in the area of teacher dedication?



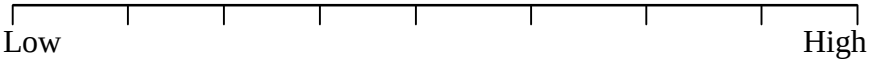
How would you rate A's performance in the area of classroom organization and preparation?



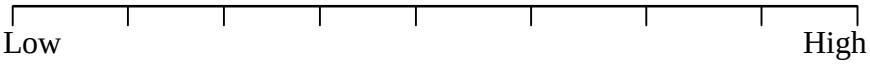
How would you rate A's performance in the area of teacher expertise?



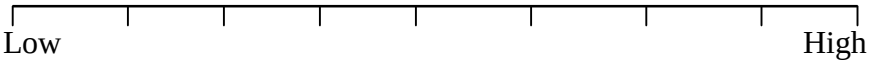
How would you rate A's performance in the area of courtesy and respect for students?



How would you rate A's performance in the area of adequately preparing students for exams?



How would you rate A's performance in the area of classroom delivery and presentation?

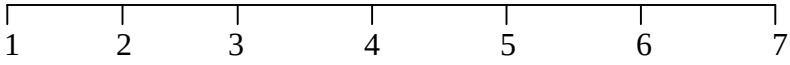


Appendix D

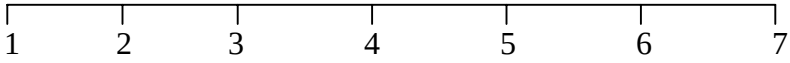
Follow-up Questions

Listed below are the six questions that are presented to the subjects following the last vignette. These questions simply assess the subjects' reactions to the vignettes and the different rating formats. All questions are likert-type answers ranging from one to seven.

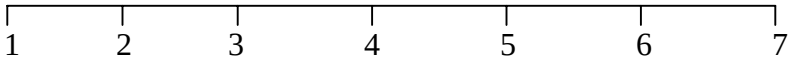
1. To what extent did you feel comfortable with the scales used to evaluate the performance of the teachers?



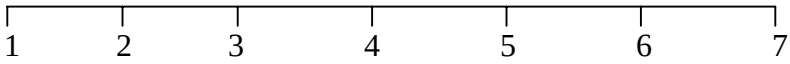
2. To what extent do you think the scenarios were clear in their description of the teachers' behaviors?



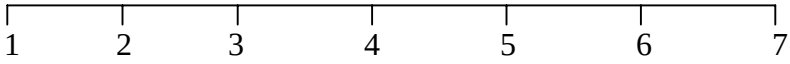
3. To what extend did you feel that the scales allowed an accurate assessment of the teachers' different dimensions of teaching performance?



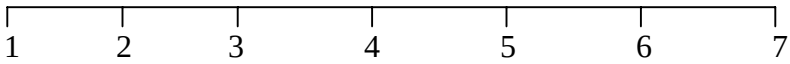
4. To what extent did you like the scales compared with other scales you have used to judge a person's performance in the past (i.e., in-class teacher evaluations)?



5. How realistic, or "true to life," do you feel the teaching scenarios were? In other words, to what extent do you think that these scenarios could have been modeled after real teachers here on campus?



6. To what extent do you feel you had enough information to make an accurate judgement about the teachers on each of the six performance dimensions?



Bibliography

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. Journal of Applied Psychology, 77 (6), 975-985

Barrett, R. S., Taylor, E. K., Parker, J. W., & Martens, L. (1958). Rating scale content: I. Scales information and supervisory ratings. Personnel Psychology, 11, 333-346.

Bendig, A. W. (1952). A statistical report on a revision of the Miami instructor rating sheet. The Journal of Educational Psychology, 43, 423-429. (a)

Bendig, A. W. (1952). The use of student rating scales in the evaluation of instructors in introductory psychology. The Journal of Educational Psychology, 43, 167-175. (b)

Benjamin, R. (1952). A survey of 130 merit rating plans. Personnel, 29, 289-294.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6 (2), 205-212.

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

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

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65 (1), 60-66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification on some issues regarding the development and use of behaviorally anchored rating scales (BARS). Journal of Applied Psychology, 66 (4), 458-463.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64 (4), 410-421.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (ed.), Performance Assessment: Methods and Applications (pp. 100-120). Baltimore, MD: Johns Hopkins University Press.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based traits versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60 (5), 561-565.

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 (1), 15-22.

Cascio, W. F. (1998). Applied Psychology in Human Resource Management. Upper Saddle River, NJ: Prentice Hall.

Cascio, W. F. & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters. Journal of Applied Psychology, 62 (3), 278-282.

Chiu, C. K., & Alliger, G. M. (1990). A proposed method to combine ranking and graphic rating in performance appraisal: The quantitative ranking scale. Educational and Psychological Measurement, 50, 493-503.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal and correlates. Journal of Applied Psychology, 74 (1), 130-135.

Coren, S., Porac, C., & Ward, L. M. (1979). Sensation and Perception. New York, NY: Academic Press.

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

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of reference training and information configuration on memory organization and rating accuracy. Journal of Applied Psychology, 80 (1), 158-167.

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

DeVries, D. L., Morrison, A. M., Schullman, S. L., & Gerlach, M. L. (1986). Performance Appraisal on the Line. Greensboro, NC: Center for Creative Leadership.

Doverspike, D., Cellar, D. F., & Hajek, M. (1987). Relative sensitivity to performance cue effectiveness as a criterion for comparing rating scale formats. Educational and Psychological Measurement, 47, 1135-1139.

Feldman, J. M. (1986). A note on the statistical correction of halo error. Journal of Applied Psychology, 71 (1), 173-176.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educational and Psychological Measurement, 32, 255-265.

Ford, A. (1931). A Scientific Approach to Labor Problems. New York, NY: McGraw Hill.

French-Lazovik, G., & Gibson, C. L. (1984) effects of verbally labeled anchor points on the distributional parameters of rating measures. Applied Psychological Measurement, 8 (1), 49-57.

Friedman, B. A., & Cornelius, E. T. (1976). Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 61 (2), 210-216.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. Journal of Applied Psychology, 59, (6), 705-711.

Hauenstein, N. M. A., Facteau, J. & Schmidt, J. A. (1999, April). Rater Variability Training: An Alternative to Rater Error Training and Frame-of-Reference Training. Poster session presented at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. Journal of Applied Psychology, 73 (1), 68-73.

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

Jackson, C. (1996), An individual differences approach to the halo-accuracy paradox. Personal Individual Differences, 21 (6), 947-957.

Jacobs, R. R. (1986). Numerical rating scales. In R. A. Berk (ed.), Performance Assessment: Methods and Applications (pp. 82-99). Baltimore, MD: Johns Hopkins University Press

Kane, J. S., & Bernardin, H. J. (1982). Behavioral observation scales and the evaluation of performance appraisal effectiveness. Personnel Psychology, 35, 635-641.

Kay, B. R. (1959) The use of critical incidents in a forced-choice scale. Journal of Applied Psychology, 60, 695-703.

Keaveny, T. J., & McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 60 (6), 695-703.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34 (2), 263-289.

Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3 (2), 193-200.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87 (1), 72-107.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60 (5), 550-555.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A monte carlo approach. Journal of Applied Psychology, 60 (1), 10-13.

McKelvie, S. J. (1978). Graphic rating scales: How many categories? British Journal of Psychology, 69, 185-202.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74 (4), 619-624.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives. Thousand Oaks, CA: SAGE Publications.

Murphy, K. R., & Constans, J. I. (1987). Behavioral anchors as a source of bias in rating. Journal of Applied Psychology, 72 (4), 573-577.

Murphy, K. R., Philbin, T. A., & Adams, S. R. (1989). Effect of purpose of observation on accuracy of immediate and delayed performance ratings. Organizational Behavior and Human Decision Processes, 43, 336-354.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo error in performance ratings: A field study of the moderating effect of halo test validation results. Journal of Applied Psychology, 75 (3), 290-296.

Neck, C. P., Stewart, G. L., & Manz, C. C. (1995). Thought self-leadership as a framework for enhancing the performance of performance appraisers. Journal of Applied Behavioral Science, 31 (3), 278-302.

Paterson, D. G. (1922). The Scott Company graphic rating scale. The Journal of Personnel Research, 1, 361-376.

Ryan, F. J. (1958). Trait ratings of high school students by teachers. Journal of Educational Psychology, 49 (3), 124-128.

Ryan, T. A. (1945). Merit rating Criticized. Personnel Journal, 24, 6-15.

Schmidt, N. & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. Journal of Applied Psychology, 62 (3), 261-264.

Schneier, C. E. (1977). Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 62 (5), 541-548.

Squires, P., & Adler, S. (1998). Linking appraisals to individual development and training. In J. W. Smith (ed.), Performance Appraisal: State of the Art in Practice (pp. 132-162). San Francisco, CA: Jossey-Bass Publishers.

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 (2), 149-155.

Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78 (6), 994-1003.

Taylor, E. K., & Hastman, R. (1956). Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology, 9, 181-206.

Tziner, A. (1984). A fairer examination of rating scales when used for performance appraisal in a real organizational setting. Journal of Occupational Behaviour, 5, 103-112.

Tziner, A., Kopelman, R. E., & Livneh, N. (1993). Effects of performance appraisal format on perceived goal characteristics, appraisal process satisfaction, and changes in rated job performance: A field experiment. Journal of Psychology, 127 (3), 281-291.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81 (5), 557-574.

Woehr, D. J., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. Journal of Applied Psychology, 78 (2), 232-241.

VITA

EDUCATION

M.S. Industrial/Organizational Psychology, May 1999
Virginia Polytechnic Institute and State University, Blacksburg, VA

B.A. Psychology, Magna cum Laude in Psychology, 1996
Texas Tech University, Lubbock, TX

PROFESSIONAL EXPERIENCE

Consultant, Virginia Polytechnic and State University Office of Admissions, Blacksburg, VA (September 98 to December 98)

Worked on a problem analysis team to investigate the effects of implementing a new computer technology on department production and information flow. Conducted job analyses for multiple positions, developed job descriptions, and diagramed organizational work flow paths. Also conducted job observation sessions, employee brainstorming meetings, and mediated communication between multiple organizational levels. Collaborated with team to review and evaluate current work procedures, communication lines, and training systems in order to make recommendations concerning future job structuring, training, and staff level technical support.

RESEARCH EXPERIENCE

Primary Investigator, Virginia Polytechnic Institute and State University, Blacksburg, VA (May 98 to May 99)

Conducted research to examine the effects of different rating formats on measures of performance appraisal rating accuracy. Designed a new computer-based rating scale to evaluate performance. Constructed a behaviorally-based rating scale to evaluate performance. Analyzed data using both analysis of variance and multiple regression.

Primary Investigator, Virginia Polytechnic Institute and State University, Blacksburg, VA (Sept 98 to Dec 98)

Examined job analysis results of the common-metric questionnaire in predicting pay. Also, explored gender differences in job analysis results and pay prediction. Finally, using factor analytic methods, tested a path analysis model of the relationship between derived CMQ subscales and "pay construct" indicators.

Research Assistant, Virginia Polytechnic Institute and State University, Blacksburg, VA (Sept 97 to Dec 97)

Worked on team to develop and administer a survey instrument designed to assess civilian voting attitudes and behavior. Used factor analytic methods to identify various attitudinal constructs and convergent/divergent validation methods to assess scale validities.

Research Assistant, Texas Tech University (Sept 95 to Dec 96)

Participated in study designed to test information processing as a function of subgroup membership, information quality, and computer-generated feedback. Designed situational scenarios containing information of varying quality. Designed and implemented a novel coding scheme used to assess quality of subjects' verbal responses.

TEACHING EXPERIENCE

Teaching Assistant: Personality Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA (Jan 99 to May 99)

Advised and tutored students on course related topics. Assisted in test construction, administration and evaluation.

Teaching Assistant: Introductory Psychology Recitation, Virginia Polytechnic Institute and State University, Blacksburg, VA (Sept 97 to Dec 98)

Full responsibility of teaching course. Topics covered included research methodology, sensation/perception, classical/operant conditioning, physiological psychology, developmental psychology, abnormal psychology, and industrial/organizational psychology.