


1-1-2017

# Efficacy Of A Structured Free Recall Intervention To Improve Rating Quality In Performance Evaluations

Maximum Mgrdich-Ararat Sirabian  
*Wayne State University,*

Follow this and additional works at: [https://digitalcommons.wayne.edu/oa\\_theses](https://digitalcommons.wayne.edu/oa_theses)

 Part of the [Organizational Behavior and Theory Commons](#), and the [Psychology Commons](#)

---

## Recommended Citation

Sirabian, Maximum Mgrdich-Ararat, "Efficacy Of A Structured Free Recall Intervention To Improve Rating Quality In Performance Evaluations" (2017). *Wayne State University Theses*. 586.  
[https://digitalcommons.wayne.edu/oa\\_theses/586](https://digitalcommons.wayne.edu/oa_theses/586)

This Open Access Thesis is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Theses by an authorized administrator of DigitalCommons@WayneState.

**EFFICACY OF A STRUCTURED FREE RECALL INTERVENTION TO IMPROVE  
RATING QUALITY IN PERFORMANCE EVALUATIONS**

by

**MAXIMUM MGRDICH-ARARAT SIRABIAN**

**THESIS**

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

**MASTER OF ARTS**

2017

MAJOR: PSYCHOLOGY (Industrial and  
Organizational)

Approved by:

---

Advisor

Date

## **DEDICATION**

**To myself...**  
for being AWESOME!

**To my parents...**  
for all the love and encouragement  
“hi mom and dad” \*waves\*

**To my advisor...**  
for all the support and patience

**To creativity, novelty, and imagination...**  
without whom this thesis would have been completed much, much sooner.

## TABLE OF CONTENTS

<b><u>Chapter</u></b>	<b><u>Page</u></b>
DEDICATION .....	ii
LIST OF TABLES .....	iv
LIST OF FIGURES .....	v
CHAPTERS	
CHAPTER 1 – Introduction.....	1
CHAPTER 2 – Method.....	26
CHAPTER 3 – Data Analysis.....	31
CHAPTER 4 – Results .....	32
CHAPTER 5 – Discussion.....	36
REFERENCES .....	47
APPENDICES	
Appendix A – Information Sheet .....	67
Appendix B – Script.....	68
Appendix C – Manipulation Check .....	69
Appendix D – Rating Scales.....	70
ABSTRACT.....	73
AUTOBIOGRAPHICAL STATEMENT.....	74

## LIST OF TABLES

<b><u>Table</u></b>	<b><u>Page</u></b>
Table 1: Means and Standard Deviations for Structured Free Recall, Frame of Reference Training and Control Groups for Halo.....	59
Table 2: Accuracy: Means and Standard Deviations for Structured Free Recall, Frame of Reference Training and Control Groups for Accuracy.....	60

## LIST OF FIGURES

<b><u>Figure</u></b>	<b><u>Page</u></b>
Figure 1: Method.....	61
Figure 2: Data Cleaning.....	62
Figure 3: Additional Analysis Overall Halo.....	63
Figure 4: Additional Analysis Overall Accuracy.....	64
Figure 5: Additional Analysis Pre and Post FoRT Accuracy.....	65
Figure 6: Additional Analysis Pre and Post FoRT Halo.....	66

## **Chapter 1: Introduction**

Performance evaluations are a very important process in the workplace and have been widely discussed and researched in industrial and organizational psychology (DeNisi, 1997; Murphy & Cleveland, 1995). According to researchers, effective performance evaluations are a cornerstone for an organization and critical to their success (Banks & May, 1999; MacLean & Chelladurai, 1995). Many personnel decisions within an organization rely heavily on the successful administration of these performance evaluations. Typically, the performance evaluation provides the manager (who will be referred to as the rater) an opportunity to rate an employee (who will be referred to as the ratee) and assess their job performance, goals, and organizational priorities. More often than not, the data obtained from these performance evaluations can be used to determine promotions and validate selection choices, as well as any other decisions that are made in an organization. Unfortunately, many performance evaluation procedures can be spoiled by numerous psychometric errors that can have negative effects on the reliability, validity, and accuracy of the obtained measurements (Bernardin & Pence 1980).

A common psychometric error found in the workplace during performance evaluations is a halo error. A halo error has been defined as “the influence of global evaluation on individual attributes of a person” (Nisbett & Wilson, 1977, p. 250). Meaning, during the performance evaluation a rater measures the performance of a ratee according to an overall impression instead of specific traits that are relevant exclusively to each performance dimension. Studies on the halo error have found evidence that a rater’s overall impression can strongly influence ratings of specific attributes across multiple performance dimensions during the performance evaluation process (Cooper, 1981). Therefore, it is commonly accepted that the halo error has a negative impact on the effectiveness of the decisions made based on the performance evaluation, as well

as the quality and accuracy of the obtained measurements (Kinicki, Bannister, Horn, & DeNisi, 1985; Saal & Knight, 1988). Thus, researchers are in agreement that removing the halo error from the performance evaluation process is an acceptable and meaningful endeavor (Bartlett, 1983; Bernardin & Beatty, 1984; Myers, 1965).

Over the years, researchers have focused on removing psychometric rater errors by developing training programs to improve the quality and accuracy of the performance evaluation process (DeCotiis & Petit, 1978; Dunnette & Borman, 1979). The significance of training programs to reduce rater errors during performance evaluations has been acknowledged since the mid 1900's (Bitner, 1948). Since then, numerous studies have been conducted which have concluded that, in general, rater training programs have a strong positive effect on reducing psychometric error during performance evaluations (Bernardin & Pence, 1980; Bernardin, 1978; Borman, 1979). For example, error training familiarizes raters with the halo error and encourages them to avoid it, whereas Frame of Reference Training (FoRT) provides raters with appropriate standards of the dimensions that will be rated and emphasizes the multidimensionality of each measurement (Woehr & Huffcutt, 1994). Ultimately, researchers agree that raters who participate in rater training programs demonstrate a superior level of rating quality and accuracy, as well as fewer psychometric errors than individuals who do not (Bernardin & Buckley, 1981; Smith, 1986; Spool, 1978).

Based on the findings of previous research, the primary goal of this current study is to examine a new training program that modifies the cognitive retrieval process of raters, which can maximize the psychometric quality and accuracy of performance evaluation measurements. Although past research has provided various effective rater training programs such as error training and FoRT to reduce halo errors and improve the quality of ratings, neither forms employ



a cognitive based approach that modifies the retrieval threshold of a rater's observed behaviors (Baltes & Parker, 2000; Roch, Woehr, Mishra, & Kieszczyńska, 2012; Segrest, 2010). Specifically, error training is mainly concerned with the reduction of rating errors, and as such, researchers believe that a fair amount of accuracy is lost during the training process. Likewise, even though FoRT addresses the accuracy issue, the training itself is time consuming, and questions have been raised about the long term effectiveness that FoRT can provide raters (Fiske & Neuberg, 1990; Stamoulis & Hauenstein, 1994). Therefore, the current approach will use an active intervention known as a Structured Free Recall Intervention (SFRI) in which raters are explicitly asked to remember and write down both positive and negative behaviors that they observed during the performance evaluation (Baltes & Parker, 2000). In doing so, raters would be able to recollect relevant information pertaining to the performance dimensions, reducing the impact of the halo error and improve the rating quality of the performance evaluations. In addition, whereas other forms of training require a formal training session to instruct and implement, SFRI is administered during each performance evaluation. This is important because the diminishing effects over time of error training programs and FoRT have been noted in previous research (Bernardin, 1978). Given that organizational training programs are usually administered a few times a year, raters tend to forget their training and revert back to committing psychometric errors when it comes time for annual performance evaluations. Since SFRI would be used while conducting each performance evaluation, it provides a novel application to the performance evaluation process.

In addition, the secondary goal of this study is to inspect the effects of SFRI on rating accuracy in addition to psychometric errors. Accuracy is a term used to describe the relationship between a set of measures and a set of appropriate and acceptable standards (Guion, 1965). It is

important to maintain high levels of accuracy throughout the performance evaluation process to preserve the quality and validity of the data obtained from the raters. Past research has emphasized that error training has the potential of decreasing accuracy at the same time that it decreases psychometric errors. Specifically, error training may reduce intercorrelations that result from the halo error, but it could also reduce rater accuracy as well (Borman, 1975; Cooper, 1981; Landy & Farr, 1980). However, since SFRI uses a rater's own observed positive and negative behaviors of the ratee, instead of specifying a correct rating distribution, it is believed that SFRI will increase rating accuracy while decreasing the halo error.

Moving forward, this paper will provide a history and analysis of performance evaluations, rater errors, rater training, and their intertwining relationships with one another. In addition, the hypotheses of the current study will be presented, followed by a detailed strategy that outlines the methods, procedures, and data analysis of this study.

### ***Performance Evaluations***

The performance evaluation procedure is a crucial component of any human resource management system within an organization and one of the most important responsibilities a manager can have (Miller & Cardy, 2000). Globally, performance evaluation procedures are used in numerous organizations to measure and evaluate employee performance and accomplishments over the course of their career with the organization (DeVries, Morrison, Shullman, & Gerlach, 1981). A performance evaluation is a formal structured system that is used to measure, evaluate, and influence employee attributions and behaviors within an organization (Bohlander & Snell, 2010). The data acquired from these performance evaluations can serve multiple purposes and are generally used to assess job performance, goals, and organizational priorities, as well as make personnel decision for promotions and bonuses, offer

career and employee development, and provide feedback. In addition, performance evaluations can identify training needs and help organizations achieve their goals and objectives (Cleveland, Murphy, & Williams, 1989; Latham, Skarlicki, Irvine, & Siegel, 1993; Sulsky & Keown, 1998).

### ***History***

The first performance evaluation procedure in the United States can be traced back to 1813 where an Army General submitted the first informal evaluation of his men to the War Department. The Army General used a global rating system that described his men as either a “knave despised by all”, or a “good-natured man” (Bellows & Estep, 1954). In the late 1800’s The Federal Civil Service of the United States began giving efficiency ratings, and Congress began to require these efficiency ratings, which included information about attention, faithfulness, and competence of all their clerks (Graves, 1948; Lopez, 1968; Petrie, 1950; White, 1954). Although this was the beginning of performance evaluations, these procedures were still not being used for employee selection or promotion purposes. However, this changed in the early 1900’s due to the need to select and promote top employees with outstanding performance records within large sized and hierarchical structured organizations such as the military and government (Wiese & Buckley, 1998). Simultaneously, industrial psychologists started using trait psychology to develop a rating system that was later used by the army in World War I and II to assess officer performance (Scott, Clothier, & Spriegel, 1941). The success of these performance evaluations caught the attention of everyday business leaders who wanted to use these new performance evaluations in their organizations. By the early 1950s, numerous performance evaluation procedures and techniques were created and developed for administrative purposes, and following World War II, almost 61% of organizations started incorporating regularly scheduled performance evaluations (Patten, 1977; Spriegel, 1962; Van

Riper 1958). Eventually, over the years, more and more organization began to adopt a systematic and formal process of evaluating the performance of their employees (Murphy & Cleveland, 1995).

For centuries, organizations were content with informal performance evaluation procedures. However, as they evolved towards large entities with professional management, a more formal performance appraisal system began to take shape (Wiese & Buckley, 1998). The advantages of a properly designed formal performance evaluation procedure has been noted in past studies (Murphy & Cleveland, 1995). A well designed performance evaluation can help organizations develop their employees, assist in everyday workforce decisions, and may even increase individual commitment and satisfaction through organizational communication (Wiese & Buckley, 1998). Regardless of the type of performance evaluation, in this day and age, they represent a universal and standard foundation of every organization that directly results in personnel decisions concerning raises, promotions, and terminations (Wiese & Buckley, 1998). As such, it is of the utmost importance to make sure that these evaluations remain error free. Nonetheless, formal performance evaluations continue to rely primarily upon human information processing and judgment. Thus, raters are susceptible to inaccurately rating a ratee that results in a rater error.

### ***Rater Errors***

A rater error is defined as an inaccurate rater evaluation due to an unconscious or conscious bias, and can be based on, but not limited to, factors that include age, race, gender, as well as ethnicity (Greenhaus & Callanan, 2006). Research has discovered numerous types of rater errors that can have a negative impact on the measurements obtained from performance evaluations. For example, leniency and severity error are generally used to describe the tendency

of a rater to consistently present inappropriate ratings that are either too high or too low regardless of the ratee's actual performance (Guilford, 1954; Myford & Wolfe, 2003). This introduces a few challenges into the measurements obtained from the performance evaluation due to the low variance between rating scores and the very high or low means that are concentrated at only the ends of the distribution (Berry, 2003). This type of assessment typically relies on sheer luck of the ratee receiving a either a lenient or severe rater and can produce unfair evaluation results that can affect the validity of decisions made from the obtained ratings. In contrast, a central tendency error occurs when a rater deliberately avoids the extreme ends of the scale and rates all the ratees as average (Linn & Gronlund, 2000). This type of rating behavior also causes a lot of problems. First, it destroys the credibility of the obtained ratings, and second, it fails to distinguish between competent and incompetent ratees (Anastasi, 1988; Linn & Gronlund, 2000). In addition, raters may develop a general impression after witnessing a limited number of performances, and allow these observations to affect future judgments about the individual. This type of rating behavior is known as the halo effect, and is characterized as having high intercorrelations between independent traits (Thorndike, 1920). This poses a major concern due to the decrease in the amount of opportunities a ratee has to display their proficiency in each performance dimension (Bechger, Maris, & Hsiao, 2007). Likewise, if a rater rates a ratee highly on all performance dimensions based on their initial assessment, then correlations between performance dimensions may be inflated. This would lead organizations to make incorrect employee decisions based on unreliable ratings.

Although there are several other types of rater errors such as logical and contrast error, as well as proximity and recency error, they are very hard to detect and therefore not as commonly studied (Myford & Wolfe, 2003). Hence, this study focused on the halo error, which has been

referred to as the longest recognized and most pervasive rater error to date (Nisbett & Wilson 1977).

### ***Halo Error***

Wells (1907) initially observed a rater error that demonstrated a rater's tendency to consider a ratee's one specific trait during the performance evaluation and allow it to influence their ratings in other areas. Later, Thorndike (1920) labeled this occurrence a halo error and pointed out the unrealistically high intercategory correlation between independent traits. Over the years, halo error has gone through numerous conceptual definitional changes and has been interpreted and defined in terms of attribute variance, working across raters, and conceptual operational states (Beckwith & Lethmann, 1975; Brown, 1968; Guilford, 1954). For instance, one idea is that halo error occurs due to a rater's general impression of a ratee that influences the rating of individual characteristics (King, Hunter, & Schmidt, 1980; Linn & Gronlund, 2000). Others have suggested that a halo error occurs when a rater's assessment of a ratee's performance on one dimension influences their assessment of that ratee on other dimensions (Anastasi, 1988; Robbins, 1989). Other researchers have defined a halo error as the result of a rater's failure to discriminate across conceptually independent features of a ratee's behavior (Saal, Downey, & Lahey, 1980). Over the years, more than 100 operational definitions of the halo error have been identified and used (Balzer & Sulsky, 1992), and having numerous conceptual and operational definitions about a singular construct can be quite problematic and confusing.

As previously mentioned, there are large amounts of conceptual and operational definitions for the halo error. Even so, the studies conducted over the years have come to a mutual agreement regarding the six of the more important features of the halo error. First, the

halo error is referred to as ubiquitous, and is thought to be quite common (Bernardin & Beatty, 1984; Blum & Naylor, 1968; Cascio, 1991; Cooper, 1981; Feldman, 1986; Jacobs & Kozlowski, 1985). Meaning, it is found everywhere and is constantly encountered. Second, it is believed that the cause of the halo error begins with the rater's overall evaluation of the ratee impacts their evaluations of specific traits (Bernardin & Beatty, 1984; Cooper, 1981; Feldman, 1986; Fisticaro & Lance, 1990; Landy, 1989; Muchinsky, 1987; Murphy, 1982). Therefore, the halo error is considered to go in a top/down direction, with general evaluations shaping specific ratings. Third, the halo error is often seen as the rater's inability or unwillingness to distinguish between multiple traits and attributes of the ratee, and as such, it is characteristically viewed as a rater error (Banks & Murphy, 1985; Cooper, 1981; Feldman, 1986; Lance & Woehr, 1986; Murphy & Jako, 1989; Nathan & Lord, 1983; Saal, Downey, & Lahey, 1980; Vance, Winne, & Wright, 1983). Fourth, due to the rater's inability to discriminate between performance dimensions, the halo error can lead to inflated correlations among the dimensions that were rated. (Bernardin & Beatty, 1984; Cascio, 1991; Cooper, 1981; Lance & Woehr, 1986; McCormick & Ilgen, 1985; Nathan & Lord, 1983; Pulakos, Schmitt, & Ostroff, 1986). Finally, the observed halo error has been split into two separate and distinct entities called true halo and illusory halo (Bartlett, 1983; Bingham, 1939; Cooper, 1981b; Lance & Woehr, 1986; Murphy, 1982; Pulakos, Schmitt, & Ostroff, 1986). True halo indicates a significant correlation between distinct performance dimensions due to a general impression. Specifically, true halo, also referred to as valid halo is a reflection of the genuine overlay between performance dimensions that are being rated. In contrast, illusory halo, also referred to as invalid halo is an error in measurement. Specifically, illusory halo is an error committed by the rater due to other factors that results in the correlation between performance dimensions (Murphy et al., 1993; Pulakos, Schmitt, & Ostroff, 1986). As

is customary in classical measurement theory to assume that observed scores are comprised of true scores and errors in measurement, the same is believed to be true about the observed halo measurement when split into true and illusory halo (Lord & Novick, 1968). Thus, researchers have proposed that theoretically, observed halo contains a bit of true and illusory halo, and the best way to find the amount of each in observed halo would be to subtract true halo from illusory halo (Lance, Fisicaro, & LaPointe, 1990; Pulakos et al., 1986). However, this study will not separate true and illusory halo from observed halo. The division of halo into true and illusory halo assumes that ratings by participants reflect true halo, and implies that raters could be sensitive to the true correlations between rating dimensions. Yet, previous research suggests that raters have a difficult time assessing or detecting the true covariation between rating dimensions (Peterson & Beach, 1967; Ward & Jenkins, 1965). Additionally, there is support that true halo levels are dependent on specific behaviors that the rater observes, which can vary among each rater (Murphy & Anhalt, 1992; Murphy & Jako, 1989; Murphy & Reynolds, 1988). However, for this study, raters will only be able to observe the ratee's behavior for 15 minutes, and the managers (ratee's) will not depict any particular behaviors, rather they act neutral and only provide enough information for the given dimension. Fifth, it has been argued that it is nearly impossible to obtain true halo scores in most settings, and when they are obtained, even in the most extreme conditions, the effects are minimal (Murphy & Reynolds, 1988; Murphy & Jako, 1989). Thus, although there might be a slight benefit to separate illusory and true halo from observed halo in theory, in practice, it is generally not done (Murphy, Jako, & Anhalt, 1993). Sixth, it is universally accepted that the halo error does lead to a negative impact on the quality of the evaluations, and thought to decrease the usefulness of the obtained evaluations (Cooper, 1981; Landy, Vance, Barnes-Farrell & Steele, 1980; Kinicki, Bannister, Horn, & DeNisi, 1985;



Saal & Knight, 1988). The final, and more important feature is that it is explicitly established that removing the halo error is possible and valuable (Bartlett, 1983; Bernardin & Beatty, 1984; Cooper, 1981; Holzbach, 1978; Kenny & Berman, 1980; Landy, Vance, Barnes-Farrell & Steele, 1980; Myers, 1965).

Although there are numerous definitions of the halo error, according to the literature, it is customary to use Thorndike's original definition when conceptually and operationally defining the halo error. Thorndike defined halo as a "marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of the [specific performance dimensions] by this general feeling" (Thorndike, 1920, p. 25). Based on this definition, the halo error becomes a within rater occurrence. As such, for this study, the halo error is defined as a cognitive bias that occurs when a rater's overall impression about a ratee during the performance evaluation has a strong influence on specific attributes across several performance dimensions (Cooper, 1981).

***General Impression Model.*** Three distinct models have been developed that correspond with the three conceptual definitions of the halo error. The most relevant of the models for this study is the general impression causal model developed by Fisiocar and Lance, 1990 that demonstrated the halo rater error effects. This model includes two dimensional performance ratings and their corresponding dimensional true scores, as well as the rater's general impression of a ratee, and disturbance terms. The most important idea of the general impression model is that dimensional ratings may be influenced by the rater's general impression and in turn effect the ratee's performance evaluation.

In social cognitive psychology, the impression formation procedure proposes that individuals naturally form consistent impressions of others early on (Fiske & Neuberg, 1990;

Srull & Wyer, 1989). This notion fits in well with the above mentioned general impression halo since the rater will experience the halo error during the performance evaluation process. As specific behavioral information is forgotten, rater's will begin to rely on overall impressions, but the extent to which rater's rely on these biases depends on numerous factors such as the availability and format of the rater training program (Feldman, 1981; Lance, Woehr, & Fisicaro, 1991; Nathan & Lord, 1983; Woehr 1991).

***Summary.*** Rater errors can have a significant negative effect on performance evaluations by skewing, restricting, or intercorrelating the data. These psychometric errors have been understood as an indication that performance evaluations can contain error. Specifically, research has concluded that when a halo error occurs, it implies a decrease in the number of independent opportunities for the ratee to demonstrate their proficiency to the rater (Bechger, Maris, & Hsiao, 2007). In a real world context, when these rater errors effect performance evaluations, the organizational consequences may include unreliable performance evaluations, incorrect employee selection, and improper tenure decisions (Kanavy, Mubeena, Chitalwalla, Champion, McCafferty, Gangone, & Duarte, 2007). Therefore, several researchers have promoted the removal of these rating errors through rater training. Particularly, the halo error has been studied for nearly a century and researchers are in agreement that the occurrence of a halo error is a threat to the validity of the measurements obtained through performance evaluations (Downing & Haladyna, 2004). Given that inaccurate evaluations occur due to rater errors, they can be detrimental to the performance evaluation process, and researchers have focused on developing training techniques that organizations can use to effectively train their raters to avoid the halo error (Bernardin & Buckley, 1981).

### ***Rater Training***

The main goal of rater training is to reduce psychometric errors in performance evaluations so that the overall quality and accuracy of performance evaluations increase. Over the years, researchers have advocated the idea of rater training to improve the quality of ratings obtained from performance evaluations (DeCotiis & Petit, 1978; Dunnette & Borman, 1979). As such, several experiments have demonstrated that rater training can decrease psychometric errors and improve rating quality during performance evaluations (Bernardin, 1978; Bernardin & Walter, 1977; Borman, 1975; Ivancevich, 1979; Latham, Wexley, & Pursell, 1975). Moreover, research has provided numerous examples that support the success of rater training programs in decreasing rater errors such as the halo error, and increasing accuracy in performance evaluations (Bernardin, 1978; Bernardin & Waler, 1977; Borman, 1975; Ivancevich, 1979; Latham, Wesley, & Pursell, 1979; Levine & Buter, 1952). Rater training usually involves exercises designed to provide raters with the knowledge and skills to successfully complete performance evaluations without committing psychometric and accuracy errors. For example, rater error training achieves this goal through training sessions that demonstrate and offer examples of ratings that portray common rater errors (Woehr & Huffcutt, 1994). A simulation study conducted by Latham, Wexley, and Pursell (1975) demonstrated common rating errors to participants and explained how to avoid those errors. Afterwards, they found that a trained group of managers in an experimental group committed less rater error than participants in the control group. Likewise, Borman (1975) implemented a 5-minutes training program that asked participants to read a carefully prepared description of the halo error, and showed them two examples of ratings, where the first rating contained halo error, while the second one did not. He discovered that even a brief training session that informed participants about rating errors and how to avoid them could significantly decrease rater error. Furthermore, researchers examined

the outcomes of a diary based rater training program on decreasing psychometric errors. They discovered that participants in the group who recorded behaviors and critical incidents during the performance evaluation showed significantly less halo error and superior rating quality than groups that did not receive the training (Bernardin & Walter, 1977). Lastly, other researchers have used a frame of reference training approach to improve the accuracy and quality of performance evaluations. For instance, in a study conducted by Uggerslev and Sulsky (2008), participants received information about performance expectations and definitions of the performance dimensions. In addition, participants watched a practice video and provided practice ratings that were later discussed out loud and given appropriate feedback by the researchers on the rating choices they made. The researchers discovered that participants who underwent the FOR training were significantly more accurate than the control group during performance evaluations. More recently, a structured free recall intervention was introduced by Baltes & Parker (2000) to improve rating quality, which later showed the ability to reduce a variety of stereotypes during performance evaluations (Baltes & Parker, 2000; Bauer & Baltes, 2002; Baltes, Bauer, & Frensch, 2007; Rudolph, Baltes, Zhdanova, Clark, & Bal, 2012). In these studies, participants were instructed to recall positive and negative behaviors they observed during the performance evaluation. They were given 5 minutes to write down relevant positive behaviors that related to the performance dimensions, and another 5 minutes to write down relevant negative behaviors related to the performance dimensions. Afterwards, participants were allowed to use the list they created during the rating process. Separately, each study discovered that the SFRI is effective in reducing the impact of varying stereotypes during performance evaluations in participants that received the SFRI compared to participants that did not.

Although there are other types of rater training programs such as performance dimensions training and behavioral observational training, this study will focus on the most popular and effective method of rater training, FoRT, which trains raters to accurately assess and distinguish between performance dimensions according to a set of standards. In addition, this study will present an intervention-based approach called SFRI to improve rating quality in performance evaluations. (Baltes & Parker, 2000; Balzet & Sulsky, 1992; Sulsky & Day, 1992, 1994).

The following pages will discuss the advantages and disadvantages of each type of training program in relation to rater errors, accuracy, and overall rating quality.

### ***Error Training***

Error training typically involves exercises designed to produce variability in the rater's evaluations of the ratee. Generally, raters are given the definition of the rater error and then presented with ratings that represent the rater error. The earliest focus was to reduce the occurrence of skewed, intercorrelated, and range restricted psychometric properties of subjective performance ratings which indicated leniency, halo, and central tendency errors (Cooper, 1981; Landy & Farr, 1980; Saal, Downey & Lahey, 1980). Generally, error training programs explain the different types of rater errors to the raters and then urge them to avoid those psychometric errors. If successful, rater errors such as the halo error would decrease, and in turn, it is believed that the accuracy of the performance evaluation process would increase. For example, Borman (1975) investigated a short rater error training session that he designed to specifically reduce the halo error in performance evaluations. Ninety participants were educated on the halo error and instructed to rate one of six vignettes. After the training session, participants began to spread out their ratings and were able to differentiate between performance dimensions. The results of this study indicated that a short error training session where participants were instructed to avoid the

halo error could significantly reduce this rater error. In addition, Bernardin (1978) examined the effects of an error training program on reducing the halo error on eighty undergraduate students who rated their instructors. The students were randomly assigned to four groups, three experimental groups that had a training emphasis, and one control group. The results showed that the quality of ratings significantly improved for participants who were in the training groups.

***Error Training and Accuracy.*** Accuracy is a term used to describe the relationship between a set of measures and another set of corresponding measures commonly known as benchmarks that are considered an acceptable standard of comparison (Guion, 1965). It is important to maintain high levels of accuracy throughout the performance evaluation process to preserve the quality and validity of the data obtained from the raters. Continued research has shown that although error training can successfully reduce halo errors, it generally may not increase accuracy, or worse, can even decrease accuracy (Borman, 1975; Cooper, 1981, Landy & Farr, 1980). For example, Bernardin & Pence (1980) administered a training program to students to rate teacher performance in hopes of reducing the halo error. The results indicated that even though error training was significantly reducing psychometric error, it was also reducing the accuracy of the evaluations. Due to this, and other similar findings, many researchers concluded that error training was not an appropriate rater training tool (Hedge & Kavanagh, 1988).

***Summary.*** Error training has gone through some growing pains, but early on, it was the most widely accepted, effective, and frequently evaluated training strategy when it came to decreasing the halo error in performance evaluations. Unfortunately, the decrease in accuracy that accompanied the decrease in psychometric errors was too high a cost for researchers to keep employing this method of training. Therefore, more recently, error training methods of

identifying and avoiding rater errors were set aside for a more favorable training program that focused on rating accuracy. Specifically, this new training method was able to influence how raters encode, characterize, organize, and recollect information (Roch, Woehr, Mishra, & Kieszczyńska, 2012).

### ***Frame of Reference Training (FoRT)***

As time went on, merely decreasing rater errors such as halo was not enough to guarantee accuracy. So, FoRT was created due to the shortcomings of traditional rater training programs that were unsuccessful in increasing rater accuracy even though they were decreasing psychometric rater errors (Stamoulis & Hauenstein, 1993; Hedge & Kavanagh, 1988; Woehr & Huffcutt, 1994). FoRT uses a rater training method that is based on a social cognitive approach that focuses on performance standards and their dimensions during a performance evaluation. This type of training characteristically involves emphasizing the multidimensionality of performance, defining performance dimensions, providing a sample of behavioral incidents representing each dimension and finally practicing using these standards to evaluate performance while receiving feedback (Bernardin & Buckley, 1981). Research has concluded that FoRT training has the ability to improve rating quality and accuracy by helping raters match behaviors with specific levels of performance and dimensions, as well as establish performance standards that help raters combat potential information loss (Hauenstein & Foti, 1989; Ilgen & Feldman, 1983; Sulsky & Day, 1992, 1994; Woehr, 1994).

There are two well-known meta analyses described the steps needed to implement a successful FoRT program, and provided support that FoRT does increase rating accuracy (Roch, Woehr, Mishra, & Kieszczyńska, 2012; Woehr & Huffcutt, 1994). To successfully present FoRT to participants, researchers employed numerous steps to facilitate the performance evaluation

process. First, participants were informed that performance is multidimensional and is made up of many parts. Second, researchers emphasized the fact that ratee performance needed to be evaluated separately and specifically to the given performance dimension. Third, participants watched a video of a manager and subordinate interaction, where the manager was always considered the ratee. Fourth, behaviorally anchored rating scales that consisted of four main performance dimensions such as motivating employees, developing employees, establishing and maintaining rapport, and resolving conflicts were used to rate the managers (Stamoulis & Hauenstein, 1993; Woehr & Huffcut, 1994; Sulsky & Day, 1992; Sulsky & Day, 1994; Woehr, 1994). Fifth, researchers went over the rating scales before the raters began the performance evaluation process, and finally, researchers provided feedback to the participants to let them know what rating they should have given for a specific performance dimension (Stamoulis & Hauenstein, 1993; Hedge & Kavanagh, 1988; Woehr & Huffcut, 1994; Sulsky & Day, 1992; Sulsky & Day, 1994; Woehr, 1994).

***FoRT and Accuracy.*** A key element of FoRT is its ability to increase rating accuracy, and as such, researchers have spent many years discussing what constitutes an accurate rating. The majority of studies have incorporated accuracy component indexes that include differential elevation, differential accuracy, elevation, and stereotype accuracy (Cronbach, 1955). Differential Elevation is measured by using the accuracy of the mean evaluation of each ratee within all performance dimensions, and differential accuracy is when raters rank ratees on a given performance dimension. Elevation is measured by using the accuracy of the mean rating over all the dimensions and ratees, while stereotype accuracy takes the mean rating of each dimension across all ratees (Woehr, 1994). Specifically, FoRT, and the performance evaluation literature focuses on all four accuracy component indexes (Day & Sulsky, 1995; Stamoulis &



Hauenstein, 1993; Sulsky & Day, 1992; Sulsky & Day, 1994; Woehr, 1994). However, out of these accuracy components, differential elevation is recognized as a significantly important type of accuracy in regards to FoRT and performance evaluations since it indicates how accurately a rater can differentiate between performance dimensions among individual ratees (Murphy & Cleveland, 1995; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992; Sulsky & Day, 1994; Woehr, 1994; Sulsky & Day, 1995). However, since this study instructs numerous raters to evaluate one single ratee, accuracy scores will be calculated using distance accuracy measures (Sulsky & Balzer, 1988). Distance accuracy is defined as the difference between each dimension rating and the corresponding true score for that dimension across dimensions for a single ratee. Although it is beyond the scope of this paper to delve deeper into the different types of relationships between accuracy indices and FoRT, it is commonly accepted that understanding these relationships can provide researchers with insight into the cognitive categorization methods used by raters.

***Cognitive Mechanism.*** There is a common consensus that raters tend to categorize ratee's on the basis of preexisting biases, and then use this general categorization to make rating decisions during performance appraisals (Feldman, 1981). The goal of FoRT is to provide raters with accurate and appropriate prototypes across low, moderate, and high performance levels for each performance dimension (Sulsky & Day, 1992). Thus, FoRT introduces a correct categorization process to raters that they can use to make accurate performance appraisal ratings. However, just like with any type of cognitive shortcut, this type of categorization may become automatic, and raters can, and do become too reliant on this process and begin to neglect specific behavioral recall. Previous research by Sulsky & Day (1992) hypothesized that the success of FoRT is due to the formation of valid prototypes for different levels of effectiveness on separate

performance dimensions. If valid and accurate prototypes of performance dimensions are used in categorizing ratees, then raters seem to improve in their ability at classification. However, this leads to a significantly greater bias within raters towards impression consistent behaviors that correspond with the trained prototypes. Meaning, once these impressions are formed within raters, they treat it just like any other bias and are quick to identify and categorize behaviors that occurred if its impressions were consistent. Therefore, Sulsky & Day (1992) concluded that raters who received FoRT ended up using their trained impressions to guide responses. Meaning, the success and rating accuracy increase that FoRT provides may not be due to greater memory for behavioral information, but rather, a result from the correct categorization of ratee performance.

Although FoRT leads to an increase in rating accuracy in comparison with untrained raters, raters who receive FoRT have shown lower memory recall, and a general decrease in identifying the occurrence of specific behaviors (Sulsky & Day, 1992; Sulsky & Day, 1994). Due to the decrease in behavioral recall, raters may not be able to provide ratees with specific examples when going over their performance appraisals. This is problematic because specific behavioral feedback is needed to improve employee performance (Roch & O'Sullivan, 1999). So, even though FoRT provides a benchmark for raters to increase the accuracy of performance appraisal ratings, raters might be relying on this process too heavily and in exchange losing actual behavioral recall of critical incidents.

**Summary.** FoRT has been widely regarded as the go to rater training program, and has been consistently used in recent performance evaluation literature since it is a significant improvement in rating quality and accuracy from other forms of rater training. (Roch, Woehr, Mishra, & Kieszczynska, 2012). However, the effectiveness of FoRT varies depending on the

operational definition of the type of accuracy index that is used. In addition, research has revealed that FoRT is time consuming, and the lack of understanding of its cognitive mechanisms may decrease a rater's behavioral accuracy, and their ability to identify certain behaviors that ratees portrayed (Sulsky & Day, 1992; Sulsky & Day, 1994; Sulsky & Day, 1995). Finally, questions have been raised about the long term effectiveness that FOR training can provide raters (Fiske & Neuberg, 1990; Stamoulis & Hauenstein, 1993). Thus, research on rating quality and accuracy during performance evaluations has been shifting its focus towards investigating rater cognitive processes and other forms of training programs (Woehr & Huffcutt, 1994).

### ***Structured Free Recall Intervention***

SFRI is an active intervention in which raters are asked to recollect and write down both positive and negative events during the performance evaluation. This cognitive process makes specific observed behaviors more salient and accessible to the rater, which in turn helps them avoid any conscious or unconscious biases towards the ratee (Baltes & Parker, 2000; Bauer & Baltes, 2002; Baltes, Bauer, & Frensch, 2007; Rudolph, Baltes, Zhdanova, Clark, & Bal, 2012). This notion was based on research conducted by Feldman and Lynch (1988) that examined the varying conditions that effected a raters decisions and memory processes to generate an evaluation of the ratee. It was discovered that two components, accessibility and diagnosticity, directly impact the connection between decisions and memory. Accessibility describes how easily a rater can become aware and remember the cognitive constructs and observations that were presented during the performance evaluation, and diagnosticity signifies if the rater believes that the observations and cognitive constructs are important to the performance evaluation (Feldman & Lynch, 1988). Specifically, Feldman and Lynch (1988) state "the most

accessible cognition sufficient to determine a response is used” (p. 429). This directly impacts performance evaluations and halo, as well as rating quality and accuracy since raters do not cycle through all available information that may be of importance to the ratee’s performance evaluation (Feldman & Lynch, 1988). Meaning, raters will stop thinking once they find a piece of information that is suitable for the given performance evaluation, regardless if it is correct or not, which consequently effects the following performance dimensions.

To combat this process, a structured free recall intervention is used to instruct raters to recall specific positive and negative details that they have seen to conduct the performance evaluation. This modifies the way that raters recollect observed behaviors and allows for relevant and non-relevant information to become salient. In theory, since raters have to focus their energy and time on remembering the positive and negative performance behaviors, these memories should not be affected by their biases. Therefore, it is believed that this will reduce the reliance on biases that raters commonly exhibit during performance evaluations (Baltes & Bauer 2002). In addition, Baltes and Parker (2000) agree that the SFRI should aid raters in accessing specific performance behaviors that ratees demonstrated during the performance evaluation. In turn, this could increase the probability of raters using those relevant memories when they conduct the performance evaluation. Ultimately, SFRI is trying to force the rater into using observed and relevant information to successfully conduct a performance evaluation by decreasing a rater’s dependence on an overall judgment or bias of the ratee.

***SFRI and Accuracy.*** Previous research on the relationship between SFRI and accuracy is limited. However, there is one study conducted by Baltes and Parker (2000) that examined the effects of SFRI on the accuracy of participant ratings when given performance cues. They expected that rating accuracy should improve in participants in the SFRI groups as raters relied

less on performance cues. Results indicated that SFRI successfully removed the negative effects of performance cues in the experimental group, and increased rating accuracy. Building on the previous work of Baltes and Parker (2000), it is expected that since raters explicitly write down the positive and negative events during the performance evaluations, they should decrease their reliance on conscious or unconscious biases, and the accuracy of their ratings should improve.

***SFRI Efficacy.*** The effectiveness of SFRI has been demonstrated on numerous occasions against prevalent rater biases such race, gender, and bodyweight (Baltes & Bauer, 2002; Baltes et al., 2007; Rudolph et al., 2012). For example, Bauer and Baltes (2002) wanted set out to understand if SFRI could reduce stereotypical behavior within raters who exhibited a bias towards women. In this study, participants evaluated vignettes describing the performance of male or female college professors and provided performance ratings. Results showed that without SFRI, raters who exhibited a gender bias evaluated women more negatively, but with SFRI, the effects of gender biases on ratings were effectively removed from participants. Moreover, they demonstrated the accuracy of raters that underwent the SFRI by measuring one of Cronbach's (1955) components, differential elevation. Results showed that SFRI had a positive impact on accuracy with raters who held gender stereotypes during performance evaluations.

In addition, Baltes et al. (2007) built on the research of Bauer and Baltes (2002) and analyzed the effectiveness of SFRI on negative racial biases. Results indicated that participants that did not receive the SFRI and held negative racial biases evaluated Black men more negatively, but these biases were reduced in participants who did receive the SFRI. Furthermore, this study examined the cognitive mechanisms that underlie the SFRI process. It was revealed that the reduction in biases is due to a modified strength threshold for retrieval of specific

performances from memory. Furthermore, Rudolph et al. (2012) extended and developed previous research by investigating the effect that SFRI had on body weight stereotypes and time delays between the observation and performance evaluation. Results showed that the SFRI effectively and significantly reduced body weight based stereotypes that impacted performance evaluations.

The goal of this study is to extend the range of the SFRI, and demonstrate its effectiveness in reducing psychometric errors such as the halo error. Given the above methods and descriptions about the SFRI, and the previous explanations about the halo error, it is believed that the SFRI can be effective in reducing halo errors in raters. When a rater commits a halo error, it is because the rater allowed a general impression of the ratee on one performance dimension to influence his ratings on subsequent independent performance dimensions (Beckwith & Lethman, 1975; Schmidt, 1980; Linn & Gronlund, 2000). The logical solution for this problem would be to provide raters with information about the ratee's performance on each dimension to force raters to consider each performance separately. When raters receive SFRI training, they are instructed to recall all the positive and negative behaviors they observed during the performance evaluation process (Baltes & Parker, 2000; Bauer & Baltes, 2002). This recollection should force the raters to consider each performance dimensions separately according to the positive and negative behaviors they listed. Ultimately, raters would evaluate ratee's independently on each performance dimensions according to their own memories and not let their general impression influence the rating process.

**Summary.** Overall, the above studies by Bauer and Baltes (2002), Baltes et al. (2007), and Rudolph et al. (2012) provide evidence of the efficacy of the SFRI. Unlike traditional rater training programs that might be conducted annually or bi-annually, SFRI is administered during

the time of the performance evaluation, and uses a cognitive process that allows raters to use their own observations to make better evaluations of the ratees. In theory, these observations should allow raters to rate ratee's independently on varying performance dimensions, and as such, it is believed that a SFRI should be able to reduce halo errors in performance evaluations. In addition, unlike traditional rater training programs that might be expensive and time consuming, a SFRI can be administered very easily to raters, and can be cost effective for organizations. Given this information, this study believes that a SFRI could be a useful instrument in reducing the halo error and increasing rating accuracy during the performance evaluation process.

### *Hypotheses*

The motivation for conducting the present study is to introduce an alternate method to rater training, known as a structured free recall intervention, which has been shown to significantly reduce a number of biases during performance evaluations. The main goal of this study is to test the efficacy of SFRI at improving a rater's rating quality during performance evaluations by reducing the effects of the halo error, while maintaining and increasing rating accuracy. As discussed previously, while research has demonstrated the efficacy of SFRI for race, gender, and bodyweight, it is crucial to provide ongoing support for the validity of SFRI by generalizing across diverse biases (Fontenelle, Phillips, & Lane, 1985). In addition, research has discussed that SFRI can directly reduce a rater's reliance on external judgments such as stereotypes and biases by forcing raters to recall behaviors displayed by the ratee during the time of the performance evaluation (Baltes & Parker, 2002), whereas FoRT is primarily concerned with establishing appropriate ratings for various levels of performance (Roch, Woehr, Mishra, & Kieszczynska, 2012). In this specific situation, participants in the FoRT group were at an

advantage since the performance evaluation came directly after the training. Whereas in the workplace, months could pass between FoRT and the actual evaluations that employers would have to provide. Still, since SFRI forces an individual to recall positive and negative behaviors that they experienced instead of relying on FoRT's previously established rating scales for certain behaviors, it is expected that SFRI would increase accuracy and reduce halo more than FoRT.

As such, the first formulated hypothesis for this experiment was that (H1a) raters in the SFRI group would have lower halo rating error scores than raters in the control group, and (H1b) raters in the SFRI group would have lower halo rating error scores than raters in the FoRT group. The second formulated hypothesis for this experiment was that (H2a) raters in the SFRI group would have higher accuracy rating scores than raters in the control group, and (H2b) raters in the SFRI group would have higher accuracy rating scores than raters in the FoRT group.

Ultimately, the goal of this study is to expand upon previous research in the rater training literature and provide a novel, easy to use, rater training intervention to improve the quality and accuracy of performance evaluations.

## **Chapter 2: Methods**

### ***Participants and Procedure***

Participants were recruited from a student body pool of a large, urban, Midwestern U.S. university and were awarded extra credit in their psychology courses in exchange for their participation. The inclusion criteria included individuals who are over the age of 18. Participants were able to sign up for the study online, and there were 1 to 6 slots available for each time slot. In order to determine the appropriate number of participants, a power analysis was conducted using G\*Power that established a total sample size of 300 for 3 groups (100 per group) with an



effect size of .18, which was determined by adding a medium (.25) and small (.10) effect size and dividing by 2, and .8 power (Erdfelder, Faul & Buchner, 1996). However, due to above average participation from the university's student body, the final total sample size was 429 for 3 groups (143 per group). Afterwards, all participants that were signed up for the study were randomly assigned as groups to either two experimental groups, (1) frame of reference training and (2) structured free recall, or a (3) control group, prior to beginning the experiment.

Upon entering the laboratory, participants were greeted by a researcher and asked to sign a consent form (see Appendix A). Next, the researcher provided the experimental packet and read a brief script (see Appendix B) that conveyed minor deception to the participants. This deception was necessary since previous research has shown that the reason of the assessment can affect the performance evaluations. In addition, participants provided more realistic and accurate performance ratings during experiments when they believed that their ratings will be used for administrative purposes rather than just for research (Dobbins, Cardy & Truxillo, 1986; Dobbins, Cardy & Truxillo, 1988; Zedeck & Cascio, 1982). Even though there was a possibility that some participants may not believe this statement, past research has provided evidence that this technique does work (Bauer & Baltes, 2002; Dobbins et al., 1986; Dobbins et al., 1988; Maurer & Taylor, 1994).

Next, depending on the group that the participants were in, researchers administered the appropriate training program. Participants in the frame of reference training group (G1) received the training first and then watched the video. In this group, a separate and different video called "Nick" was used during the training segment to train participants. Once complete, participants used a behaviorally anchored rating scale to evaluate the performance of the managers on three distinct performance dimensions (Smith & Kendall, 1963). Participants in the structured free

recall intervention group (G2) began by watching the video, followed by a 10-minute intervention where participants recalled and wrote down positive and negative behaviors they observed. Afterwards, participants were able to use what they wrote down while evaluating the performance of the managers on three distinct performance dimension using the behaviorally anchored rating scales. Finally, participants in the control group (G3) did not receive any training, and directly completed the behaviorally anchored rating scales after watching the videos. Lastly, participants in each group were provided with two questions at the end of the study to act as manipulation checks, and to determine if participants were paying attention during the experiment (see Appendix C). When everything is completed, participants will be debriefed, as well made aware of the initial deception, and thanked for their time.

### ***Materials***

***Performance Videos.*** The video used in this study depicted a critical incident of managerial performance that was developed by Roberson and Banks (1986). The incidents were based on one of eight videotapes created by Borman (1977), and later used by Sulsky and Day (1992). Specifically, the videos named “Jim” and “Nick” illustrate a fictitious manager, who are the ones being evaluated by participants, interviewing a problematic subordinate. Again, the video “Nick” will only be used for the training portion for participants in the FOR group (G2). The manager displayed behaviors that pertained to one of four performance dimensions of (1) motivating employees, (2) developing employees, (3) establishing and maintaining rapport, and (4) resolving conflict. Although results may be less generalizable when compared to similar procedures that are conducted in a field setting, the videotape method was preferable here due to the available true scores, which are generally difficult to obtain or estimate accurately in most field settings (Murphy, Jako & Anhalt, 1993).

***Performance Rating Scales.*** For this study, the performance measures were obtained by three 7-point behaviorally anchored rating scales that were developed by Borman (1978). These scales are separated according to low performance (1 & 2), average performance (3, 4, & 5), and high performance (6 & 7). These criteria will be used to assess employee performance on (1) motivating employees, (2) developing employees, and (3) establishing and maintaining rapport (see Appendix D). The resolving conflicts dimension will be used as a distractor to enhance generalizability of the appraisal task, since research has shown that raters often observe behavior that is irrelevant to the performance dimension (Sulsky & Day, 1995). Finally, true scores were acquired by Sulsky and Day (1992), and consist of (1), (1), (7) for manager “Jim” and (7), (7), (1) for manager “Nick”.

#### ***Rater Training and Intervention***

***Frame of Reference.*** Researchers provided participants in the FoRT group with instructions that were adopted from experiments by Bernardin, Buckley, Tyler, and Wiese (2000). First, researchers instructed participants to independently evaluate performance dimensions and not let one judgment on a certain dimension carry over to the other. Second, researchers discussed the behaviorally anchored rating scales with the participants and specifically focused on the differences and definitions between each dimension. Third, participants watched a practice video “Nick” and used a behaviorally anchored rating scale to provide practice ratings based on the above-conveyed frame of reference. Fourth, researchers provided feedback to the participants regarding their rating accuracy. Finally, each participant described the reasoning behind their practice ratings, and discussed it with the researcher out loud.

**Structured Free Recall Intervention.** Researchers provided instructions on how to use the SFRI based on previous experiments by Baltes and Bauer (2002), Baltes et al. (2007), and Rudolph et al. (2012). After participants watched the video, they were instructed to explicitly remember and document both positive and negative behaviors they observed that were relevant to the performance dimensions after watching the video. Next, participants were given five minutes to remember and record as many positive behaviors relevant to the performance dimensions that are being evaluated. After the initial five minutes, participants received another five minutes to remember and document as many negative behaviors that are relevant to the performance dimensions being evaluated. Previous research suggested to counterbalance this process to avoid any type of order effects (Baltes & Bauer, 2002; Baltes et al., 2007; Rudolph et al., 2012). After 10 minutes, participants finished recording their responses on a sheet of paper., and they were able to use this information during the rating process.

### **Measures**

**Rating Accuracy.** As previously mentioned, performance evaluations are an integral part of an organizations process of allocating promotions, terminations, and training needs.

Therefore, it is important to make sure that raters are accurate in these evaluations. Since this study only had one ratee during the performance evaluations, the measure of accuracy that was used is distance accuracy. Distance accuracy is defined as the average absolute value of the deviation of the obtained ratings from the true scores across dimensions for a particular ratee

(McIntyre, Smith, & Hassett, 1984). The distance accuracy formula is defined as 
$$DSTA = (1/dn) \sum_{j=1}^d \sum_{i=1}^n |O_{ij} - T_{ij}|$$

and reflects the proximity of observed ratings to actual true scores.

**Halo.** As stated earlier, there are numerous operational and conceptual definitions of the halo error. However, this study operationalized the halo error according to Thorndike's (1920)

original definition for measurement purposes. As such, the measurement of the halo error was based on a single rater evaluating a single ratee, and was defined as the standard deviation across performance dimensions using this formula.

Thus, halo in this study is considered to be a relative measure, meaning performance ratings from raters could contain an increased or decreased level of halo. The standard deviation measure of halo error was obtained by computing the standard deviation of true scores across performance dimensions, and the standard deviation of observed ratings across dimensions. Then these numbers were subtracted, and the mean was calculated across the performance dimensions, where a positive value indicates the presence of halo error (Fisicaro, 1988). 
$$\text{HALO-S} = (1/n) \sum_{i=1}^n (ST_{i.} - SO_{i.})$$

### **Chapter 3: Data Analysis**

To test the first hypothesis proposed in this study, that (H1a) Raters in the SFRI group would have lower halo rating error scores than raters in the control group, and (H1b) Raters in the SFRI group would have lower halo rating error scores than raters in the FoRT group, an ANOVA was conducted between the experimental training groups and the control group to analyze the differences between the group means in regards to halo using an alpha of .05. Additionally, to test the second hypothesis proposed in this study that (H2a) Raters in the SFRI group had higher accuracy rating scores than raters in the control group, and (H2b) raters in the SFRI group had higher accuracy rating scores than raters in the FoRT group, another ANOVA was conducted between the experimental training groups and the control group to analyze the differences between the group means in regards to accuracy using an alpha of .05. In the event of a significant omnibus F, indicating that there was a significant difference between the groups, post-hoc analyses would be conducted. In the event of a non-significant omnibus F, indicating that there was not a significant difference between the groups, planned comparisons would be

conducted. It was expected that there might be significant difference between the groups, and a priori non-orthogonal comparisons would need to be conducted.

The initial planned comparison used univariate T-tests to analyze the significant difference between the (*H1a*) SFRI group and control in regards to the halo error, and the (*H1b*) SFRI group and FoRT group in regards to the halo error. It was expected that raters who undergo the SFRI would make significantly fewer halo errors than raters who do not. The following planned comparisons used univariate T-tests to analyze the significant difference between the (*H2a*) SFRI group and control groups in regards to accuracy, and (*H2b*) the SFRI group and FoRT groups in regards to accuracy. It was expected that SFRI trained participants would be more accurate than raters who do not. Since the number of comparisons being made did not exceed the number of degrees of freedom between groups ( $K-1$ ), the alpha level of each comparison stayed at .05. However, since the comparisons were non-orthogonal (SFRI+Control and SFRI+FoRT) for each dependent variable, a Bonferroni correction was made to correct for the multiple non-orthogonal statistical tests.

#### **Chapter 4: Results**

Data cleaning procedures were utilized through methods that analyzed missing data, participant manipulation checks, as well as outliers and normality (see Figure 2). First, a frequency analyses was conducted on the data set that identified no missing data for the Control ( $n = 143$ ), SFRI ( $n = 143$ ), and (FoRT  $n = 143$ ) groups. Next, participant manipulation checks were reviewed for each group to identify participants that were not paying attention during the experiment. This was accomplished by identifying participants who passed or failed the manipulation check. If one of the two manipulation checks were failed, then that participant was removed from the data set. According to a frequency analyses of manipulation check 1 and

manipulation check 2, a total of 7 participants failed at least one of the manipulation checks in the Control group (ID# 23, 30, 92, 97, 108, 125, and 135), 7 participants failed at least one of the manipulation checks in the SFRI group (ID# 3, 33, 45, 47, 59, 124, and 125), and 7 participants failed at least one of the manipulation checks in the FoRT group (ID# 35, 45, 48, 69, 98, 126, and 131). These participants were removed from the data set leaving a sample size of  $n = 136$  for the Control, SFRI, and FoRT groups. After removing participants that failed the manipulation check, univariate and multivariate analyses were conducted. Univariate outliers were determined by transforming the three rating dimension into  $z$  scores, and using a  $p$  value of less than .001 to establish a  $\pm 3.29$  cut off. A frequency analysis of the  $z$  scores concluded that there were no univariate outliers in the data set. Next, multivariate outliers were determined by Mahalanobi's Distance and interpreted by using a  $p$  value of less than .001, and a *chi square* value with the degrees of freedom equal to the number of values, in this case 3, to establish a  $\pm 16.266$  cutoff. A frequency analysis of the Mahalanobi's Distance scores also concluded that there were no multivariate outliers in the data set. Data was also inspected for distributional assumptions of normality by examining the skew and kurtosis of each evaluation within each group. To determine the acceptable range of skewness and kurtosis, a  $p$  value of less than .001 was used to establish a  $\pm 3.29$  cut off, and the data was interpreted by dividing the standard error of skew and kurtosis by the appropriate statistic of skew and kurtosis. Results indicated that employee motivation (4.41) in the SFRI group, as well as employee development (4.51) and establishing and maintaining rapport (5.00) in the FoRT group were moderately and positively skewed, and therefore, a square-root transformation was applied to the variables. However, significance levels did not change significantly when transformed variables were used, and thus, for clear interpretation, untransformed variables were continued to be used.

Next, an overall Accuracy measure was calculated for each participant by subtracting the average absolute value of the deviation of the obtained ratings from the true scores (1, 1, 7) across the dimensions for a particular rater (DSTA =  $(1/dn) \sum(d) \sum(n) |O_{ij}-T_{ij}|$ ), where n = number of raters (1), d = number of dimensions (3), O = observed score, as well as T = true score, and the closer the calculated accuracy score was to zero, the more accurate individuals were on each dimension. Again, univariate analyses were conducted, and outliers were determined by transforming the calculated overall accuracy into z scores, and using a p value of less than .001 to establish a +/-3.29 cut off. A frequency analysis of the z scores concluded that there were no univariate outliers in the data set. Next, an overall Halo measure was calculated for each participant by subtracting the obtained standard deviation across the three dimensions from the true score standard deviation (1, 1, 7) across the three dimensions (Halo-S =  $(1/n) \sum(n) (SD_{Ti}-SD_{Oi})$ ), where n = number of raters (1),  $SD_T$  = true standard deviation across the three dimensions, as well as  $SD_O$  = observed standard deviation across three dimensions, and the closer the calculated halo score was to zero, the less halo each dimension had.

In summary, after thoroughly analyzing the groups for missing data, participant manipulation, as well as outliers and distributional assumptions of normality, the data was ready to be analyzed to test the proposed hypotheses.

**Hypothesis 1:** *(H1a) Raters in the SFRI group had lower halo rating error scores than raters in the control group, and (H1b) Raters in the SFRI group had lower halo rating error scores than raters in the FoRT group.* A one-way between subjects ANOVA was conducted to compare the effect of rater training on halo rating error for the SFRI, FoRT, and Control groups. A significant effect of training type on levels of halo rating error at the  $p < .05$  level for the three



groups [ $F(2, 405) = 0.50, p = 0.952$ ] was not found, suggesting that rater training type did not have an effect on the increase or decrease in halo rating error (see Table 1).

**Hypothesis 2:** (*H2a*) *Raters in the SFRI group had higher accuracy rating scores than raters in the control group, and (H2b) raters in the SFRI group had higher accuracy rating scores than raters in the FoRT group.* A one-way between subjects ANOVA was conducted to compare the effect of rater training on accuracy ratings for the SFRI, FoRT, and Control groups. A significant effect of training type on levels of accuracy ratings at the  $p < .05$  level for the three groups [ $F(2, 405) = 2.615, p = 0.074$ ] was not found, suggesting that rater training type did not have an effect on the increase or decrease in accuracy ratings (see Table 2).

### **Additional Analyses**

Given previous findings on the positive effects of rater training on halo rating error and accuracy ratings, additional exploratory analyses were conducted to better understand these results this study. These analyses consisted of post-hoc examinations of mean differences between each group for both halo rating error and accuracy ratings, as well as mean differences between pre and post FoRT training for both halo rating error and accuracy ratings.

To investigate whether or not there were significant differences between rater training groups for halo error ratings, an independent sample's t-test was conducted. Figure 3 presents the means and standard deviations of rater training by halo error. Analyses between the Control and SFRI ( $t(270) = 0.330, p = .742$ ), Control and FoRT ( $t(270) = 0.080, p = .936$ ), as well as the SFRI and FoRT ( $t(270) = -0.215, p = .830$ ) groups yielded non-significant results.

Next, to investigate whether or not there were significant differences between rater training groups for accuracy ratings, an independent sample's t-test was conducted. Figure 4 presents the means and standard deviations of rater training by accuracy. Analyses between the

Control and SFRI ( $t(270) = -0.208, p = .836$ ), and SFRI and FoRT ( $t(270) = -1.874, p = .062$ ) groups also yielded non-significant results. However, there were significant differences between the Control and FoRT ( $t(270) = -2.045, p < .05$ ) group. Although this analysis suggested that there were significant differences between the FoRT group and the Control group, the direction of significance showed that individuals who did not receive training were significantly more accurate during rater assessments than individuals who did receive FoRT. This abnormal finding went against all previous research and literary findings on rater training.

Finally, to investigate whether or not there were significant differences between pre and post FoRT for halo error and accuracy ratings, a paired sample's t-test was conducted. Figure 5 presented the means and standard deviations, and showed a non-significant interaction between pre and post FoRT for halo error ratings ( $t(135) = 0.561, p = .576$ ). Figure 6 presented the means and standard deviations, and showed a significant difference between pre and post FoRT for accuracy ratings ( $t(135) = -3.168, p < .05$ ), but the significant result was again in the wrong direction, suggesting that individuals who did not receive FoRT were significantly more accurate in their rater assessments than after FoRT. One explanation of these results could be due to the length of time individuals spent in the FoRT experiment. The training itself was longer than the SFRI group, and participants could have lost interest in the study, and overtime became less engaged. Thus, the longer the training took, the less accurate they were post FoRT than pre FoRT

## **Chapter 5: Discussion**

The purpose of this study was to demonstrate the effect that rater training has on improving rating quality in performance evaluations. Specifically, this study examined the differences between participants in the SFRI and control group in regards to reducing

psychometric error and increasing accuracy in performance evaluations. In addition, this study made comparisons between training groups (SFRI and FoRT) to determine if SFRI was better suited at reducing the halo error and increasing accuracy than FoRT, the most current, and predominant forms of rater training.

The primary goal of this current study was to examine a new training program, called structured free recall intervention, to modify the cognitive retrieval process of raters to maximize the psychometric quality of performance evaluation measurements. Psychometric errors such as halo errors have a negative influence on the effectiveness of performance evaluations by impacting the rater's overall impression of a ratee, which can strongly influence ratings of specific attributes across multiple performance dimensions during the performance evaluation process (Cooper, 1981). Effective performance evaluations are foundational for an organizations success, and previous research has shown the positive effects of rater training on halo errors (Banks & May, 1999; MacLean & Chelladurai, 1995). In addition, the secondary goal of this study was to examine the effects of a structured free recall intervention to modify the cognitive retrieval process of raters to maximize the accuracy of performance evaluation measurements. Given the importance of performance evaluations and their role in personnel decisions, it becomes important to maintain a high level of accuracy throughout the performance evaluation process to preserve the quality and validity of rater evaluations (Banks & May, 1999; Guion 1965). Just as halo errors, past research has also shown the positive effects of rater training on accuracy (Bernardin, 1978; Bernardin & Walter, 1977). Moreover, research has provided an abundance of examples that support the success of rater training programs in decreasing rater errors such as the halo error, and increasing accuracy in performance evaluations (Borman, 1975; Ivancevich, 1979; Latham, Wesley, & Pursell, 1979; Levine & Buter, 1952).

Given the amount of support for rater training, and its positive effects on halo errors and accuracy, it was unfortunate to discover that the findings of the current study did not align with past research in the field. Surprisingly, no significant statistics were observed of the posited hypotheses for both dependent variables. Below, a discussion has been presented about possible explanations about the non-significant results. In addition, limitations of the study will be addressed, and future research will be identified.

### **Issue of Power**

The first potential explanation for non-significant results could be due to a limited sample size and indistinguishable effect sizes. For this study, the appropriate amount of participants was determined by a power analysis using G\*Power that established a total sample size of 300 for 3 groups (100 per group). The effect size used in this calculation was determined to be .18, which was calculated by multiplying a medium (.25) and small (.10) effect size and dividing by 2 (Erdfelder, Faul & Buchner, 1996). The total participant count was well over the recommended amount of participants, at 143 participants, and after data cleaning, 136 net participants were used for the data analyses. Meaning, the sample size was more than adequate to detect a significant difference.

Given that the sample size was more than enough to find statistically significant differences between groups, the second potential explanation for non-significant results could be that the manipulated variable, types of rater training, had no true effect on the dependent variables, accuracy, and halo.

### **SFRI and Halo**

This study used SFRI, a cognitive based active intervention where the rater explicitly remembers and documents both positive and negative events during the performance evaluation,

to reduce halo errors, a common type of rater bias. In theory, the SFRI process should make specific observed behaviors more accessible to the rater, which in turn should help them avoid any conscious or unconscious biases towards the ratee. Several studies have documented the effectiveness of SFRI in reducing a variety of biases such as race, gender, and bodyweight. In each study, the respective bias was successfully removed, and a significant difference was documented between individuals in the control and experimental groups. (Baltes & Parker, 2000; Bauer & Baltes, 2002; Baltes, Bauer, & Frensch, 2007; Rudolph, Baltes, Zhdanova, Clark, & Bal, 2012). However, in the current study, no such effects were found on halo with individuals in the SFRI group when compared to the control and FoRT groups. Therefore, it was believed that SFRI may not have an effect on halo, and alternate explanations were explored to try and explain this observation.

One such alternate explanation as to why SFRI may not have had an effect on halo could be found in the variety of biases that individuals may hold. For example, in cognitive science, there is a clear distinction between cognitive biases such as race, gender, and bodyweight, and attributional biases such as halo. Since cognitive biases are based on memories, SFRI could affect the memory recollection process during positive and negative memory retrievals. Alternately, halo, an attributional bias, is based on behaviors and an individual's tendency to consider a ratee's one specific trait or behavior during the performance evaluation and allow it to influence their ratings in other areas (Feldman, 1981; Wells, 1907). Given the method used for raters to evaluate ratees in this study, it is very plausible that no true halo bias even occurred with raters during the rater evaluation sessions. Specifically, participants in the structured free recall intervention group watched the experimental video, and immediately after, were asked to recall and write down positive and negative behaviors they observed. Since a halo error is an

attributional bias, it could be that not enough time, rapport, or information was available to successfully attribute any behaviors from the raters to the ratees in the video, and instead, participants focused on other factors of the ratees.

Another explanation could be that the video that was shown to participants was never intended as a device to manipulate or create halo in an experimental setting. Previous studies that have used these types of videos that capture the critical incidents of managerial performance solely focused on a variety of accuracy measures, but never any type of rater error or bias (Borman, 1977; Roberson and Banks, 1986; Sulsky and Day, 1992). Furthermore, studies that have used halo as a dependent variable used methods that specifically created halo during evaluations, such as raters that had a motivational or experiential connection to the ratee. Meaning, there was a clear purpose for the evaluation, and there had been enough interactions between the rater and the type of ratee, either physically or mentally to effectively attribute certain behaviors. For example, Feeley (2010) used 128 students from three communication courses to evaluate the effectiveness of a college professor, and the effects of halo on these evaluations.

***Halo Scores.*** Mean overall halo scores were calculated for each rater, where a score of 0 represented no halo effect, and higher scores represented the presence of halo, where the higher the score, the more severe the halo effect. Referring back to Table 1, which shows the means and standard deviations of overall halo scores for the control, SFRI, and FoRT groups, it is evident that the rating task could have been too short and simple, and ultimately may not have allowed halo to develop. Given that the total halo calculation could range from 0 (no halo present) to 2.82 (severe halo effect), the mean scores of 1.68 (control), 1.66 (SFRI), and 1.67 (FoRT) are representative of the lower end of the halo effect. Additionally, Halo scores did not significantly

differ between the control group and the two rater training groups, indicating that these numbers could have represented baseline halo for all participants.

**Summary.** Previous research has shown the effectiveness of rater training programs on reducing rater errors and biases. More specifically, SFRI has been shown to reduce a variety of biases such as race, gender, and bodyweight. This study attempted to further extend the efficacy of SFRI and explored its effects on a prominent psychometric rating error, halo. However, this study did not find the same significant effects of SFRI on a bias, as previous research has found. It was believed that the halo effect was not in the same category as other biases, and that there was no actual halo bias to get rid of and therefore SFRI was not able to have an effect on it.

### **SFRI and Accuracy**

Previous research on the relationship between SFRI and accuracy had been limited. One study by Baltes and Parker (2000) examined the effects of SFRI on the accuracy of participant ratings when given performance cues. They expected that rating accuracy would improve in participants of the SFRI groups as raters relied less on performance cues. Results indicated that SFRI successfully removed the negative effects of performance cues in the experimental group, and in turn increased rating accuracy. In this study, SFRI was hypothesized to directly impact rating accuracy, more specifically, distance accuracy. The reasoning for this hypothesis was due to the cognitive mechanisms that individuals use while making decisions, or in this case, rating others. According to Feldman and Lynch (1988), when raters make a judgment or evaluation, they will conduct only a quick and limited search for information before providing a response. Meaning, raters will most likely use information that is easy to remember and access when rating rates, even if it might not be the correct information that is remembered and accessed. Thus, when SFRI was used, not only were raters asked to write down the information that was easily

accessible in their memory, but they were also forced to recollect and write down the deeper and conflicting information as well. This allows the recollection of both positive and negative events during the performance evaluation, and made specific observed behaviors more accessible to the rater (Feldman & Lynch, 1988; Baltes & Parker, 2000). Given this process, individuals who received the SFRI should have increased accuracy ratings. However, this hypothesis was not supported.

One explanation as to why SFRI may not have had an effect on rating accuracy could be due to the rating accuracy measure that was chosen. According to Cronbach (1955), there were four accuracy component indexes that are referred to as differential elevation, differential accuracy, elevation, and stereotype accuracy. Differential Elevation measured the accuracy of the mean evaluation of each ratee within all performance dimensions, differential accuracy had raters rank ratees on a given performance dimension, elevation measured the accuracy of the mean rating over all the dimensions and ratees, and stereotype accuracy measured the mean rating of each dimension across all ratees (Woehr, 1994). Since this study instructed several raters to evaluate one single ratee, accuracy scores were calculated using Borman's (1979) distance accuracy measure (Sulsky & Balzer, 1988). Although Baltes and Parker (2000) also used the same distance accuracy measure as this study, their videos were specifically made to manipulate their variables. Thus, the reasoning for non-significant interactions between SFRI and accuracy in this current study could be due to the fact that no biases or inaccuracies existed to get rid of.

***Accuracy Scores.*** Mean overall accuracy scores were calculated for each rater, where a score of 0 represented very accurate ratings, and higher scores represented more inaccurate ratings according to the relevant true scores. Referring back to Table 1, which shows the means



and standard deviations of overall accuracy scores for the control, SFRI, and FoRT groups, it is evident that participants were not too inaccurate to begin with. Given that the total accuracy calculation could range from 0 (very accurate) to 6 (very inaccurate), the mean scores of 2.9 (control), 2.92 (SFRI), and 3.1 (FoRT) are representative of the middle point of the accuracy spectrum. Meaning, given the control groups accuracy score, there was potential room for improvement in accuracy that the rater training could have provided. However, accuracy scores did not significantly differ between the control group and the SFRI rater training group. Conversely, there was a significant difference between the FoRT and Control groups, but as is evident by the scores, participants who received FoRT, according to this study, were more inaccurate.

### **FoRT, Halo, and Accuracy**

Most unexpectedly, non-significant interactions were also discovered with FoRT and both dependent variables, halo and accuracy. Even though there may be some alternative explanations for the non-significant interactions between SFRI and halo and accuracy, it was difficult to explore cases and present alternate suggestions on why FoRT did not have a significant impact in the experimental group over the control group in this study. Previous research had documented the ability of FoRT to improve rating quality and accuracy through mechanisms that helped raters match behaviors with specific performance and dimension levels (Hauenstein & Foti, 1989; Ilgen & Feldman, 1983; Sulsky & Day, 1992, 1994; Woehr, 1994). Explanations of cognitive mechanism that were provided as support as to why SFRI did not work are unavailable when it comes to FoRT due to the fact that FoRT's training method was based on a cognitive approach that focuses on performance standards and their dimensions during a performance evaluation. Meaning, the training explicitly emphasizes the multidimensionality of

performance, defines performance dimensions, provides a sample of behavioral incidents representing each dimension, and finally, individuals had a chance to practice and receive feedback before the actual performance evaluations (Bernardin & Buckley, 1981). Furthermore, the present consensus, based on decades of research, has concluded that FoRT has been widely regarded as the prominent rater training program, and has been consistently and successfully used in performance evaluation literature and practice. (Roch, Woehr, Mishra, & Kieszczyńska, 2012).

### **Limitations and Future Research**

There were numerous supporting articles with a variety of manipulations that provided evidence that rater training, specifically SFRI and FoRT, had the ability to provide positive and favorable effects on biases and accuracy (Baltes & Parker, 2000; Bauer & Baltes, 2002; Baltes, Bauer, & Frensch, 2007; Hauenstein & Foti, 1989; Ilgen & Feldman, 1983; Rudolph, Baltes, Zhdanova, Clark, & Bal, 2012; Sulsky & Day, 1992, 1994; Woehr, 1994). As with all studies, there are potential limitations that need to be addressed. Due to the non-significant findings, the first limitation may be the simplicity and straightforwardness of the stimuli in this current study. Participants were rating individuals that they had never met before, and the ratings were based on three performance criteria. Meaning, there may not have been enough time for halo to present itself. In its current form, the rater did not have enough experience with the ratee, and only had a brief amount of time to form an impression of their performance. In organizational settings, individuals have ample time to develop a halo effect towards individuals before evaluating them, and since that was not possible during the experimental study for lack of time, and without other immediate biases that were measured (i.e., race, gender, etc.), could have led to the current results.

Additionally, there may have been issues with the cognitive categorization of different biases with participants, as well as the notion that the stimuli or situation may not have created a halo effect. Therefore, future studies should add more cognitive relevant biases for individuals to pick up on, as well as create situations for halo to present itself. Likewise, prior participant skillset for accurately rating performance evaluations was not considered. Due to the simplicity and straightforwardness of the rating task, participants may not have needed to be as accurate across each group, which led to the same accuracy scores, without any improvement, across each rating group.

Next, all research that utilized student subjects may be prone to generalizability and effectiveness issues (Gordon, Slade, & Schmitt, 1986). This study did not posit that it had any more or less student subject issues than any other study, but in this case, not only does this study not generalize to a work setting, but it also does not generalize to any academic setting. Nonetheless, future studies should focus on actual work settings to better understand the effects of SFRI. Also, the use of measurable variables in this research could have been stronger. This study only explored the direct effects of training on accuracy and halo, without any regard to other indirect pathways that training could reach the final outcome of decreased halo and increased accuracy.

Future research should consider pursuing opportunities to replicate these findings in an actual work environment. In addition, more than one ratee could be included to understand how SFRI affects different accuracy indexes. Also, different types of rater training programs could be combined with the SFRI to further strengthen the efficacy of performance evaluations. Furthermore, it would be beneficial to make the rating task more complex, as to invest individuals into the scenario, and to hold sessions or provide background info of the ratee, as to

build rapport with the raters before the evaluation process. This may allow the formation of a raters first impression towards the ratee's performance. The additions of this type of information would also aid in bringing the experimental study closer to replicate real world situations during performance evaluations. Lastly, the effectiveness of a self-managed approach to SFRI has yet to be tested. Since the SFRI did not require a formal training session, future research could study how a self-managed SFRI can affect the quality of performance evaluations. More specifically, future studies should test SFRI as a self-managed rateer training intervention, rather than an experimentally managed one.

### **References**

Anastasi, A. (1988). *Reliability in psychological testing*. New York: MacMillan Publishing

- Baltes, B. B., & Parker, C. P. (2000). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes*, 82(2), 237-267.
- Baltes, B. B., Bauer, C. B., & Frensch, P. A. (2007). Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism?. *Journal of Applied Psychology*, 92(1), 151.
- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, 77(6), 975.
- Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology*, 38(2), 335-345.
- Bauer, C. C., & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles*, 47(9-10), 465-476.
- Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, 34(8), 607-619.
- Beckwith, N. E., & Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *Journal of Marketing Research*, 265-275.
- Bellows, R. M., & Estep, M. F. (1954). *Employment psychology: The interview*. New York: Rinehart.
- Bernardin, H. J., & Walter, C. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*, 62(1), 64.
- Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 63(3), 301-308. doi: 10.1037/0021-9010.63.3.301

- 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. doi: 10.1037/0021-9010.65.1.60
- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205-212. doi: 10.5465/AMR.1981.4287782
- Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston: Kent Publishing Company.
- Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A reconsideration of strategies for rater training. *Research in Personnel and Human Resources Management, 18*, 221-274.
- Bingham, W. V. (1939). Halo, invalid and valid. *Journal of Applied Psychology, 23*(2), 221.
- Bitner, R. H. (1948). Developing an industrial merit rating procedure. *Personnel Psychology, 1*, 403-432.
- Blum, M. L., & Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundations*. New York: Harper & Row.
- Bohlander, G., & Snell, S. (2010). *Managing Human Resources. South-Western Cengage Learning, USA*.
- Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60*(5), 556.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*(2), 238-252.
- Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135.

- Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410.
- Cascio, W. F. (1991). *Costing human resources*. South-Western Educational Publishing.
- Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*(1), 130.
- Cooper, W. H. (1981). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology, 66*(3), 302-307. doi: 10.1037/0021-9010.66.3.302
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological bulletin, 90*(2), 218.
- Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity.". *Psychological Bulletin, 52*(3), 177.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349.
- 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.
- DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of management review, 3*(3), 635-646.
- DeNisi, A. (1997). *A Cognitive Approach to Performance Appraisal: A Program of Research*. Routledge.
- DeVries, D. L., Morrison, A. M., Shullman, S. L., & Gerlach, M. L. (1981). Performance appraisal on the line. New York: Wiley.
- Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1986). Effects of ratee sex and purpose of

- appraisal on the accuracy of performance evaluations. *Basic and Applied Social Psychology*, 7(3), 225-241.
- Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology*, 73(3), 551.
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333.
- Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. *Annual review of psychology*, 30(1), 477-525.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior research methods, instruments, & computers*, 28(1), 1-11.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66(2), 127.
- Feldman, J. M. (1986). A note on the statistical correction of halo error. *Journal of Applied Psychology*, 71(1), 173.
- Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, 73(3), 421.
- Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, 14(4), 419-429.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category—based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, 23, 1-74.
- Fontenelle, G. A., Phillips, A. P., & Lane, D. M. (1985). Generalizing across stimuli as well as



- subjects: A neglected aspect of external validity. *Journal of Applied Psychology*, 70(1), 101.
- Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review*, 92(3), 317.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7, 7-28.
- Graves, M. (1948). *Design judgment test*. New York: Psychological Corp.
- Greenhaus, J. H., & Callanan, G. A. (Eds.). (2006). *Encyclopedia of Career Development*. Sage Publications.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- Guion, R. M. (1965). *Personnel testing* (pp. 302-353). New York: McGraw-Hill.
- Hauenstein, N., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. *Personnel Psychology*, 42(2), 359-378.
- 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.
- Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, 63(5), 579.
- Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. *Research in Organizational Behavior*, 5, 141-197.
- Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, 64(5), 502.

- Jacobs, R., & Kozlowski, S. W. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal*, 28(1), 201-212.
- Kenny, D. A., & Berman, J. S. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin*, 88(2), 288.
- King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65(5), 507.
- Kinicki, A. J., Bannister, B. D., Hom, P. W., & Denisi, A. S. (1985). Behaviorally anchored rating scales vs. summated rating scales: Psychometric properties and susceptibility to rating bias. *Educational and Psychological Measurement*, 45(3), 535-549.
- Kruglanski, A. W., Atash, M., DeGrada, E., Mannetti, L., Pierro, A., & Webster, D. M. (1997). Psychological theory testing versus psychometric nay-saying: Comment on Neuberg et al.'s (1997) critique of the Need for Closure Scale. *Journal of Personality and Social Psychology*, 73(5), 1005-1016.
- Lance, C. E., & Woehr, D. J. (1986). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal Of Applied Psychology*, 71(4), 679-685. doi:10.1037/0021-9010.71.4.679
- Lance, C. E., Fiscaro, S. A., & Lapointe, J. A. (1990). An examination of negative halo error in ratings. *Educational and Psychological Measurement*, 50(3), 545-554.
- Lance, C. E., Woehr, D. J., & Fiscaro, S. A. (1991). Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior*, 12(1), 1-20.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87(1), 72.
- Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo

- error in performance ratings. *Journal of Applied Psychology*, 65(5), 501.
- Landy, F. J. (1989). *Psychology of work behavior*. Thomson Brooks/Cole Publishing Co.
- Larkin, J. E., & Pines, H. A. (1994). Affective consequences of self-monitoring style in a job interview setting. *Basic and Applied Social Psychology*, 15(3), 297-310.
- 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.
- Latham, G. P. (1986). Job performance and appraisal. *International review of industrial and organizational psychology*, 1, 117-53.
- Latham, G. P., Skarlicki, D., Irvine, D., & Siegel, J. P. (1993). The increasing importance of performance appraisals to employee effectiveness in organizational settings in North America. *International review of industrial and organizational psychology*, 8, 87-132.
- Levine, J., & Butler, J. (1952). Lecture vs. group decision in changing behavior. *Journal of Applied Psychology*, 36(1), 29.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Columbus, OH: Merrill, an imprint of Prentice Hall.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison.
- MacLean, J. C., & Chelladurai, P. (1995). Dimensions of coaching performance: Development of a scale. *Journal of Sport Management*, 9(2), 194-207.
- Maurer, T. J., & Taylor, M. A. (1994). Is sex by itself enough? An exploration of gender bias issues in performance appraisal. *Organizational Behavior and Human Decision Processes*, 60(2), 231-251.
- Miller, J. S., & Cardy, R. L. (2000). Self-monitoring and performance appraisal: rating outcomes

- in project teams. *Journal of Organizational Behavior*, 21(6), 609-626.
- Muchinsky, P. (1987). *Validation documentation for the development of personnel selection test batteries for telecommunications service jobs*. Ames: Iowa State University.
- Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, 67(3), 320.
- Murphy, K. R., & Jako, R. (1989). Under what conditions are observed intercorrelations greater or smaller than true intercorrelations?. *Journal of Applied Psychology*, 74(5), 827.
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78(2), 218.
- Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–422.
- Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, 68(1), 102.
- Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250-256. doi: 10.1037/0022-3514.35.4.250
- Patten, T. H. (1977). *Pay: Employee compensation and incentive plans*. New York: Free Press.
- Petrie, F. A. (1950). Is there something new in efficiency rating?. *Personnel Administrator*, 13, 24.

- Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within rates to measure halo. *Journal of Applied Psychology, 71*(1), 29.
- Roberson, L., & Banks, C. G. (1986). Beyond job knowledge: Assessment skill training to increase rating accuracy. In *94th Annual Convention of the American Psychological Association, Washington, DC*.
- Robbins, S. B. (1989). Validity of the superiority and goal instability scales as measures of defects in the self. *Journal of Personality Assessment, 53*(1), 122-132.
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370-395.
- Rosenberg, M. (1965). The measurement of self-esteem. *Society and the Adolescent Self Image, 297*, 307.
- Rudolph, C. W., Baltes, B. B., Zhdanova, L. S., Clark, M. A., & Bal, A. C. (2012). Testing the structured free recall intervention for reducing the impact of bodyweight-based stereotypes on performance ratings in immediate and delayed contexts. *Journal of Business and Psychology, 27*(2), 205-222.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413.
- Saal, F. E., & Knight, P. A. (1988). *Industrial/organizational psychology: Science and practice*. Thomson Brooks/Cole Publishing Co.
- Scott, W. D., Clothier, R. C., & Spriegel, W. R. (1941). *Personal management*. New York: McGraw-Hill

- 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.
- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, 30(4), 526.
- Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: matters of assessment, matters of validity. *Journal of Personality And Social Psychology*, 51(1), 125.
- Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology*, 31(4), 853-888.
- Spiegel, W. R. (1962). Company practices in appraisal of managerial performance. *Personnel*, 39(3), 77-83.
- Strull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, 96(1), 58.
- Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, 78(6), 994.
- Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 28 (2), 191.
- Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: an empirical investigation of rater memory issues. *Journal of Applied Psychology*, 77(4), 501.
- Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology*, 79(4), 535.

- Sulsky, L. M., & Keown, J. L. (1998). Performance appraisal in the changing world of work: Implications for the meaning and measurement of work performance. *Canadian Psychology/Psychologie canadienne*, 39(1-2), 52.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), 25-29. doi: 10.1037/h0071663
- Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, 93(3), 711.
- Vance, R. J., Winne, P. S., & Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. *Personnel Psychology*, 36(3), 609-620.
- Van Riper, P. P. (1958). The senior civil service and the career system. *Public Administration Review*, 189-200.
- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality And Social Psychology*, 67(6), 1049.
- Wells, F. L. (1907). *A statistical study of literary merit: With remarks on some new phases of the method* (No. 7). The Science Press.
- Wiese, D. S., & Buckley, M. R. (1998). The evolution of the performance appraisal process. *Journal of Management History (Archive)*, 4(3), 233-249.
- Woehr, D. J. (1991). Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior*, 12, 1– 11.
- Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79(4), 525.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology*, 67(3), 189-205.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, 67(6), 752.



Table 1

Means and Standard Deviations for Structured Free Recall, Frame of Reference Training and Control Groups for Halo

	Halo	
	M	SD
SFRI	1.66	0.58
FoRT	1.67	0.73
Control	1.68	0.61

Table 2

Means and Standard Deviations for Structured Free Recall, Frame of Reference Training and Control Groups for Accuracy

	Accuracy	
	M	SD
SFRI	2.92	0.77
FoRT	3.10	0.81
Control	2.90	0.79

\

**Figures**

Figure 1: Methods

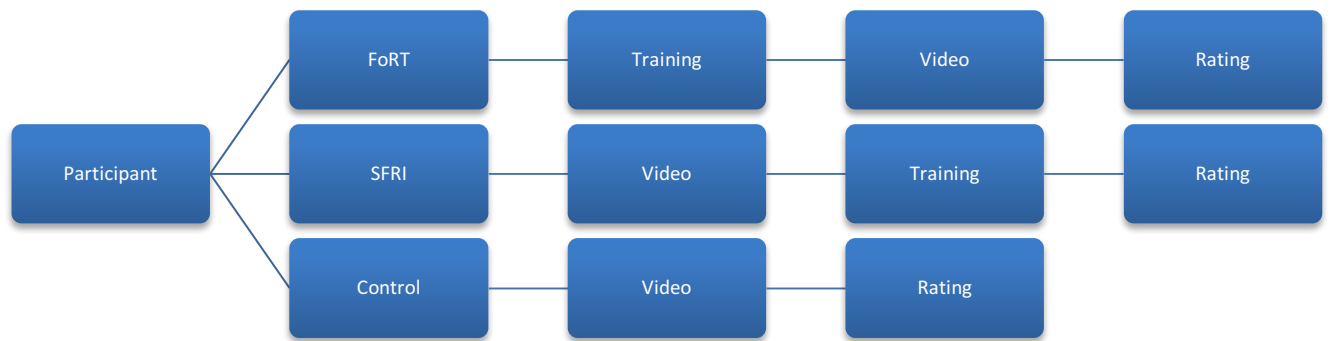


Figure 2: Data Cleaning

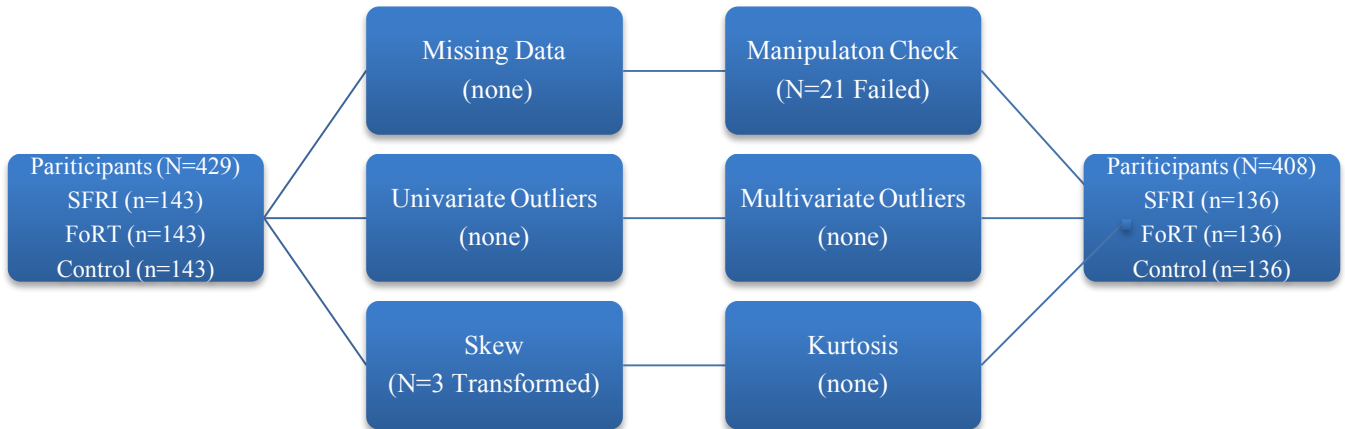


Figure 3: Additional Analysis Overall Halo

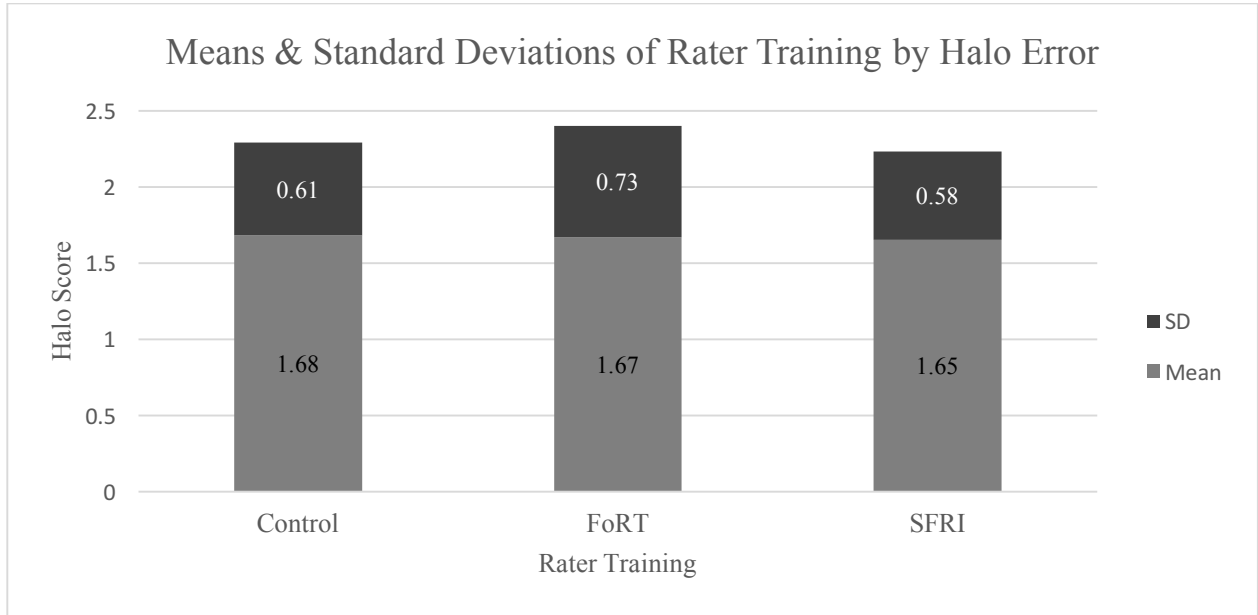


Figure 4: Additional Analysis Overall Accuracy

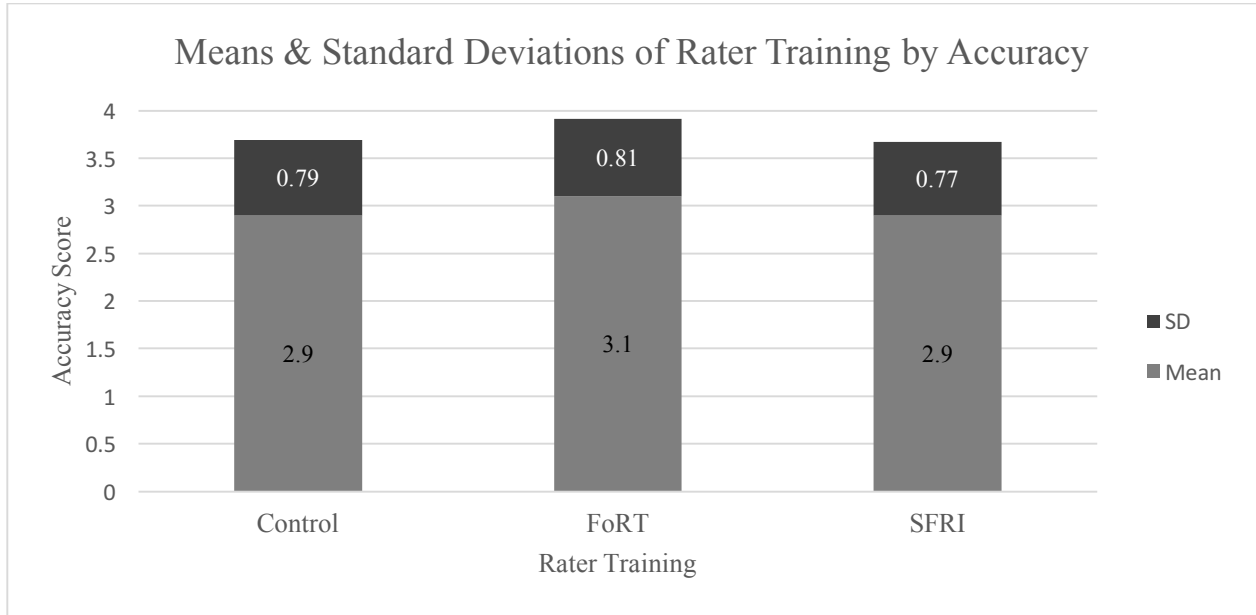


Figure 5: Additional Analysis Pre and Post FoRT Accuracy

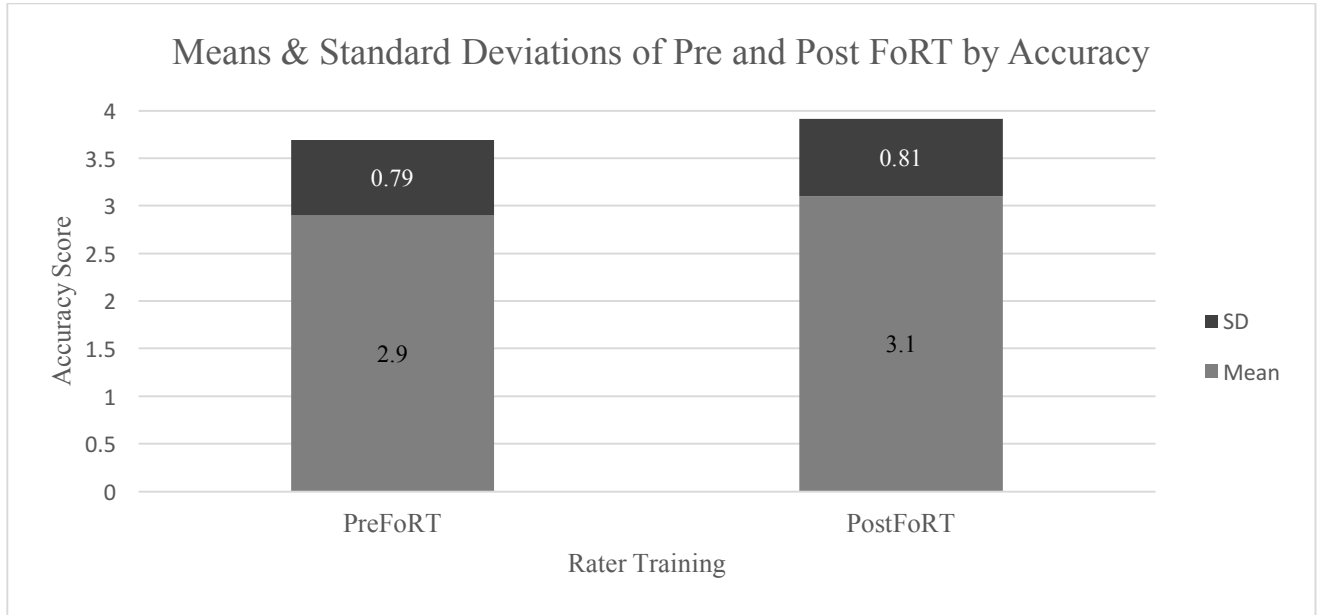
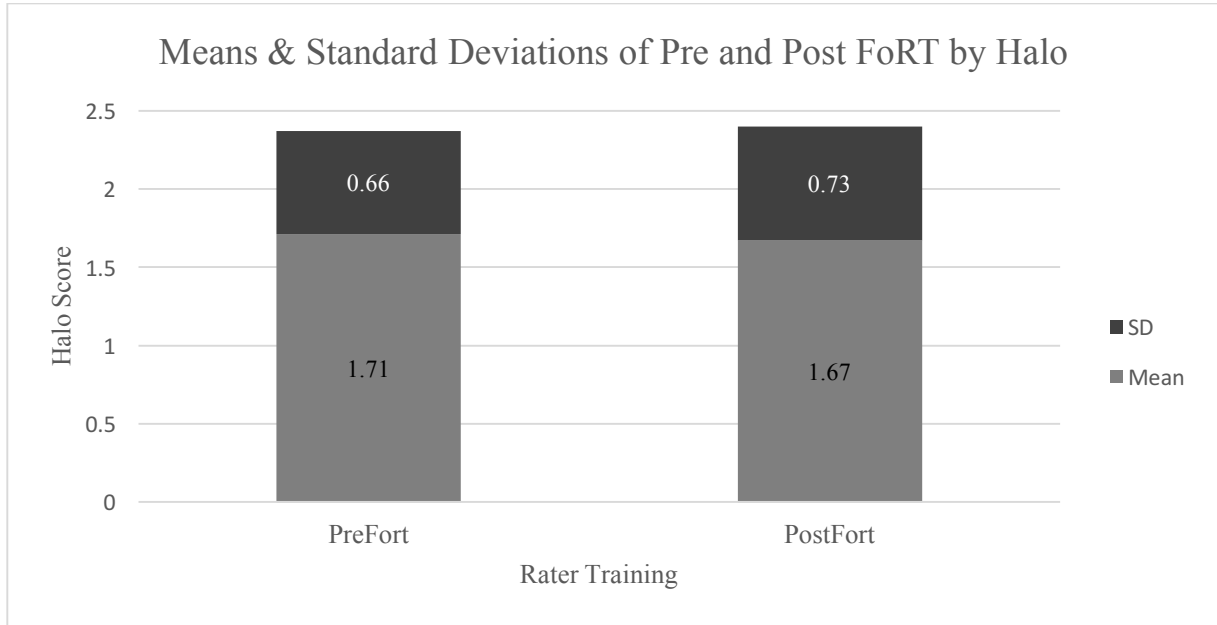


Figure 6: Additional Analysis Pre and Post FoRT Halo





## Information Sheet

### Research Information Sheet

Title of Study: Efficacy of a Structured Free Recall Intervention to Improve Rating Quality in Performance Evaluations

Principal Investigator (PI): Mgrdich A. Sirabonian  
 Psychology Department  
 Wayne State University  
 (313) 577-2424

#### **Purpose:**

You are being asked to be in a research study of performance evaluation because you have signed up through the SONA Systems. This study is being conducted at Wayne State University. The estimated number of study participants to be enrolled is about 300. **Please read this form and ask any questions you may have before agreeing to be in the study.**

In this research study, we are looking at the process of rating another's performance. Specifically, we will be looking at how various individuals rate the performance of managers.

#### **Study Procedures:**

If you agree to take part in this research study, you will be asked to view videotapes of managers and rate their performance. You may be asked to recall the positive and negative behaviors exhibited by the manager before you provide ratings of his/her performance. In addition, you may receive other forms of training. You have the option to not answer any questions you choose not to. This study will take approximately 45-60 minutes.

#### **Benefits:**

As a participant in this research study, there may be no direct benefit for you; however, information from this study may benefit other people now or in the future.

#### **Risks:**

There are no known risks at this time to participation in this study

#### **Costs:**

There will be no costs to you for participation in this research study.

#### **Compensation:**

You will not be paid for taking part in this study.

However, you will receive 1.0 research participation credits towards your psychology course(s).

#### **Confidentiality:**

All information collected about you during the course of this study will be kept without any identifiers.

#### **Voluntary Participation /Withdrawal:**

Taking part in this study is voluntary.

You are free to not answer any questions or withdraw at any time.

Your decision will not change any present or future relationships with Wayne State University or its affiliates.

#### **Questions:**

If you have any questions about this study now or in the future, you may contact Mgrdich Sirabonian or one of his research team members at the following phone number 313-577-2424. If you have questions or concerns about your rights as a research participant, the Chair of the Human Investigation Committee can be contacted at (313) 577-1628. If you are unable to contact the research staff, or if you want to talk to someone other than the research staff, you may also call (313) 577-1628 to ask questions or voice concerns or complaints.

#### **Participation:**

By completing the surveys you are agreeing to participate in this study.

## Appendix B

## Script

This study is designed to help the business school improve and develop the skills of MBA students as future managers. So, we are going to show you a video recording of a MBA student interacting with an employee. After watching the videotape, you will be asked to rate the MBA student on various performance dimensions. Your performance ratings of the MBA student will be used for both training and grading purposes.

**Appendix C**

## Manipulation Check

**Instructions:** Please read the following four questions and circle the correct response.

1. What was the **gender** of the MBA student?

Male

Female

2. What was the **ethnicity** of the MBA student?

Black

White

Other

**Appendix D**

## Rating Scales

**Instructions:** Please circle one number per page.

**EMPLOYEE MOTIVATION***Rating**Performance Examples***HIGH PERFORMANCE**

- 7 Can be expected to tell the employee that the company needs him because of his impressive expertise and proven ability to get the job done.
- 6 Would be expected to re-state commitments he made to the employee about helping him acquire a better position in the company.

**AVERAGE PERFORMANCE**

- 5 Would be expected to offer the employee a tough job assignment in such a way that the employee would agree to take it on, and then say that he knew the employee would do a good job because of his success in the past.
- 4 Throughout the interview, this manager can be expected to emphasize his desire to keep the employee in the company.
- 3 Can be expected to tell the employee that he appears to be doing an adequate job in his department, but that he could probably be doing better.

**LOW PERFORMANCE**

- 2 This manager could be expected to tell the employee to “keep plugging” on his job because the company needs to increase its earnings.
- 1 After discussing the employee’s problems with the company, this manager would suggest that the employee leave the company since he was so dissatisfied.

## EMPLOYEE DEVELOPMENT

*Rating*

*Performance Examples*

### HIGH PERFORMANCE

- 7      Could be expected to say that he would gladly review the employee's professional development on a regular basis. Could be expected to offer to attend a professional development course (e.g., the Dale Carnegie program) with the employee and suggest that they both could benefit from it.
- 6      This manager would suggest some particular professional development courses that would help the employee.

### AVERAGE PERFORMANCE

- 5      Could be expected to offer to talk with the employee about professional problems as they arise.
- 4      Expected to help the employee in his general development.
- 3      This manager would suggest that the employee obtain a list of courses from the personnel department and take the ones he felt he needed.

### LOW PERFORMANCE

- 2      Manager would be expected to state that the employee would have to work on his own to accomplish changes in his managerial style.
- 1      If the employee asked this manager for a list of things he could improve on in order to get promoted, the manager would be unable to come up with anything, and also state that he didn't believe in training and development anyway.

**ESTABLISHING AND MAINTAINING RAPPORT**

*Rating*

*Performance Examples*

**HIGH PERFORMANCE**

- 7     Would expect the manager to project considerable warmth and sincerity throughout the interview. Manager is expected to discuss the employee's job-related problems candidly and in a non-threatening manner.
  
- 6     Would be expected to begin the interview by saying that it was nice to talk with the employee in an informal setting and that he hopes they would have a good working relationship.

**AVERAGE PERFORMANCE**

- 5     Manager can be expected to draw out the employee by telling him about similar problems experienced in a previous job.
  
- 4     Can expect this manager to greet the employee cordially at the door and offer the employee a chair.
  
- 3     Can be expected to begin the interview by slapping the employee on the back and asking him how things are going in such a manner that the employee feels somewhat uneasy.

**LOW PERFORMANCE**

- 2     This manager would be expected to begin the interview somewhat abruptly by telling the employee he had arranged the meeting to talk about the employee's problems in the company.
  
- 1     This manager could be expected to tell the employee, without any initial small talk, "I suppose we both know that you are here because we have been getting reports about your not being able to get along with people on the job."

**Abstract**

This experiment investigated the effects of a rater training on halo errors and accuracy during performance evaluations. 408 participants were randomly assigned to three groups (n=136) where they were either presented with a structured free recall intervention (SFRI), frame of reference training (FoRT), or no training. The purpose of this study was to further investigate the efficacy of SFRI against prominent training methods and no training at all. Results were not significant, and did not support previous finding in the literature. Further explanations are offered and a discussion is presented as to why these results were obtained.

**EFFICACY OF A STRUCTURED FREE RECALL INTERVENTION TO INCREASE  
RATING QUALITY IN PERFORMANCE EVALUATIONS**

by

**MAXIMUM MGRDICH-ARARAT SIRABIAN**

2017

Advisor: Dr. Boris B. Baltes

Major: Psychology (Industrial and Organizational)

Degree: Master of Arts

## **Autobiographical Statement**

Maximum Mgrdich-Ararat Sirabian

I am a graduate student pursuing a doctoral degree in Industrial and Organizational Psychology. My primary research interests lie in the areas of technology, and their use in training and development, occupational health, selection, and education. I have actively presented at the Society for Industrial and Organizational Psychology, and published in journals and books. Additionally, I had the opportunity to work on the competency validation team of a Fortune 1000 company, and interned with a large health insurance company as a human performance analyst. Currently, I am exploring new emerging technologies and rich digital media to better the social, personal, and professional lives of employees, as well as use these tools to provide novel and innovative solutions to common and unique organizational problems.