

5-2010

The Effects of Rater Training on the Relationship between Item Observability and Rater Agreement

Keaton Edwin Montgomery

Western Kentucky University, keaton.montgomery@wku.edu

Follow this and additional works at: <http://digitalcommons.wku.edu/theses>



Part of the [Cognition and Perception Commons](#), and the [Industrial and Organizational Psychology Commons](#)

Recommended Citation

Montgomery, Keaton Edwin, "The Effects of Rater Training on the Relationship between Item Observability and Rater Agreement" (2010). *Masters Theses & Specialist Projects*. Paper 168.
<http://digitalcommons.wku.edu/theses/168>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.

THE EFFECTS OF RATER TRAINING ON THE RELATIONSHIP BETWEEN ITEM
OBSERVABILITY AND RATER AGREEMENT

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky


In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

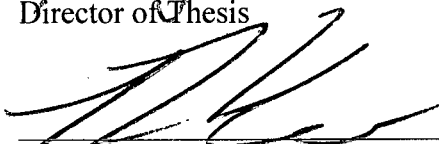
By
Keaton Edwin Montgomery

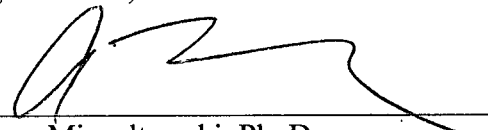
May 2010

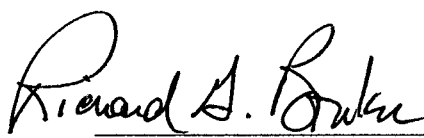
THE EFFECTS OF RATER TRAINING ON THE RELATIONSHIP BETWEEN ITEM
OBSERVABILITY AND RATER AGREEMENT

Date Recommended 4/30/10


Anthony Paquin, Ph. D.
Director of Thesis


Reagan Brown, Ph. D.


Andrew Mienaltowski, Ph. D.

 June 7, 2010
Dean, Graduate Studies and Research Date

Acknowledgments

I would like to thank my thesis Chair, Dr. Anthony Paquin, for his help throughout the thesis process. I would also like to thank my committee members, Dr. Reagan Brown and Dr. Andrew Mienaltowski for their efforts in reviewing my thesis. Finally, I would like to thank my friends and family for their support throughout my academic career and in life.

Table of Contents

Abstract	iii
Literature Review	3
Interrater Reliability versus Interrater Agreement	3
Behavioral Observability	5
Frame of Reference Training	8
Current Study	14
Method	15
Participants	15
Stimulus Performance.	15
Rating Form	16
Item Observability	16
Procedure	16
Results.	18
Discussion	19
Limitations	20
Future Research	21
References	24
Appendix A	26
Appendix B	31

THE EFFECTS OF RATER TRAINING ON THE RELATIONSHIP BETWEEN ITEM
OBSERVABILITY AND RATER AGREEMENT

Keaton E. Montgomery

May 2010

37 Pages

Directed by: Anthony Paquin, Reagan Brown, Andrew Mienaltowski

Department of Psychology

Western Kentucky University

This study was an extension of a study conducted by Roch, Paquin, and Littlejohn (2009). They investigated the relationship between rater agreement and the observability of items on a rating form. The current study found similar results in that, as items became less observable, interrater agreement increased. The purpose of this study was to introduce frame of reference training as an extension to the Roch et al. study in order to reverse their findings. In other words, trained raters would be less likely to default to a general impression on less observable items and thus would demonstrate higher rater agreement on more observable items than untrained raters. The results, based on 66 raters, replicated the findings of the Roch et al. study. The frame of reference training appeared to have no impact on the results. Results are discussed.

The Effects of Rater Training on the Relationship Between Item Observability and Rater Agreement

Performance ratings are often used in business and academia to determine levels of leadership, team skills, communication skills, problem solving skills, promotion decisions, and various other applications. Often times, the ratee is observed by multiple raters such as in 360 feedback scenarios in which a ratee receives ratings in the form of self, peer, supervisor, and subordinate ratings. One concern associated with the use of multiple raters is a lack of agreement across raters (Van Hooft, Van Der Flier, & Minne, 2006). This lack of agreement has been attributed to several sources. Hooft et al. noted that self-ratings can suffer from self-serving bias, peer-ratings may not be taken seriously by the peers, and any rater may suffer from one or more of several rating errors. Hooft et al. found that supervisors were more severe raters than peers or the self, however they also found moderate levels of agreement among the three ratings (self, peer, and supervisor). One possible solution for improving rater agreement is by training the raters in how to properly assess performance.

The purpose of the present study is to investigate the effect of rater training on the relationship between the behavioral observability of items and rater agreement. The following sections include discussions on the effect of interrater reliability, interrater agreement, behavioral observability, common rating errors, and frame of reference training (FOR) on rater agreement.

Interrater Reliability versus Interrater Agreement

Interrater reliability and interrater agreement are major components of performance ratings. These terms, however, are often confused. According to Brown

(2006) interrater reliability is obtained by correlating the ratings of multiple raters on one individual's performance. A strong correlation indicates that different raters provide the same highest rating. Wu, Whiteside, and Neighbors (2007) described interrater reliability as measuring the consistency of ratings between raters by comparing their personal scores with the scores of the other raters. For example, if Rater 1 provides ratings of 5, 4, 3 on Items 1, 2, and 3 respectively and Rater 2 provides ratings of 4, 3, 2 on Items 1, 2, and 3 respectively, there is a correlation of 1.00. This does not, however, indicate whether or not the raters used the scale correctly or necessarily agreed on the ratings. One explanation for this lack of agreement despite high reliability is offered by Roch, Paquin, and Littlejohn (2009). These authors believed that the wording of the items on the rating form could affect interrater reliability. If true, this could mean that raters are agreeing more on some items rather than others due to characteristics of the items themselves and not due to the characteristics of the ratee. Likewise, Littlefield and Troendle (1986) noted that rating forms that parallel cognitive processes of the rater could result in more reproducible and reliable scores, although this does not guarantee better agreement. In summary, interrater reliability indicates consistent ratings from multiple raters, but reliability does not indicate to what extent the raters agree on scoring behaviors.

In contrast, interrater agreement is the degree to which raters provide the same exact ratings. Although there are several methods for computing agreement, it is often measured using r_{wg} . Interrater agreement is more challenging to achieve than interrater reliability, and if interrater agreement is good, interrater reliability is not a large concern (Brown, 2006). When comparing ratings across raters, interrater agreement becomes crucial (Roch et al., 2009). If you have different raters providing vastly different ratings,

it is impossible to get even a moderately accurate measure of performance. This will lead to false feedback for the ratee. Agreement among raters is not, however, easily accomplished, as interrater agreement can be effected by numerous influences. These include, how individual raters interpret an item, how individuals interpret a behavior as beyond expectations, moderate, or poor, and the level/type of rater training the rater has received. One explanation for lack of agreement offered by Roch et al. is that error in agreement hides true performance by giving the ratee ratings that do not reflect the behaviors that were actually performed. Similarly, Littlefield and Troendle (1986) argued that raters' use of inappropriate criteria in making decisions on detailed items on rating forms may result in low levels of agreement. While it is unlikely multiple raters will provide perfect agreement prior to a consensus rating, the more rater errors are controlled for and raters are able to interpret items and behaviors in the same manner (via rater training), the more likely there will be high interrater agreement.

Behavioral Observability

Another key factor influencing performance ratings is the behavioral observability (also known as specificity) of the items. Brutus and Facticeau (2003) defined a specific behavior as being one "that not only narrowly defines the behavior to be evaluated but also provides, when possible, a contextual frame within which the target behavior is expected to occur" (p. 315). Not every item on a performance rating will be highly specific. Some items have to be more generalized simply due to the content/construct the item is assessing. In addition, Wohlers and London (1989) suggested that some performance dimensions may be easier to rate because they are more observable. An example of an observable item would be, "Involved others in the task." This would be

easy to observe for a rater as a poor performance on this item would be represented by passive behavior by the ratee. Likewise, if the performance were good, it would be easily observed by the ratee's assertiveness and actively asking group members their opinions. Wohlers and London believed that because the more observable items were easier to rate, that those items would have higher interrater reliability. They did, however, fail to find consistent support for this.

Roch et al. (2009) investigated the relationship between behaviorally based ratings (rather than subjective judgments) and rater agreement in two studies. The first study investigated the relationships amongst behavioral observability, perceived rating difficulty, rater agreement, and rater reliability. The authors defined behavioral observability as being the extent to which judgment was involved in determining if a behavior occurred or not. The behavioral observability was scored using a scale ranging from "observable behavior" to "subjective judgment". In this study, participants viewed a videotaped assessment center activity and were instructed to evaluate one of the individuals in the video. After viewing the video, the participants rated the target individual's performance on an 86 item rating form. This rating form included four dimensions of assessment: team skills, oral communication, professionalism, and problem solving.

The authors found a positive correlation between rater agreement (r_{wg}) and performance ratings and perceived rating difficulty. However, they found a negative correlation between rater agreement and behavioral observability. This finding led to the conclusion that rater agreement increased for items with less observability. Findings also

showed that the more subjective an item was, the more difficult the participants found that item to rate.

The second study by Roch et al. (2009) was designed to replicate the results of the first study as well as to investigate whether or not prior exposure to the rating items impacted the rater agreement and/or perceived difficulty. The only difference in procedure from the first study was the introduction of the performance dimensions and rating items to the treatment group. The data from the control group in the second study replicated that of the first study. Specifically, the authors again found that rating agreement increased as behavioral observability decreased and that more observable items were perceived as easier to rate. Similar to the control group, the experimental group produced higher levels of agreement as items became less observable. The authors suggested that this may occur because as rating items become less observable (and thus perceived to be harder to rate), the participants might be referring to a “default” answer based on their general impression of the target individual’s overall performance. This explanation indicates that general evaluations of the ratee, potentially including first impression bias and halo error, may be used by raters when they face rating an item that is perceived to be difficult to rate. The authors do note in their follow-up study, that difficulty in assessing specific behaviors from memory cannot completely account for the initial study’s results.

Ultimately, Roch et al. (2009) concluded their results suggested that high levels of interrater agreement and reliability may not be indicative of high rating quality. Based on their results it seems that the less observable an item is, the more a rater will refer to the general impression of the target individual when rating that item. This would potentially

lead to higher rater agreement based on a common general impression of the ratee rather than a specific rating item. The authors also suggested that future research should address whether or not rater training programs could reduce this tendency for raters to refer to a default impression for less observable rating items. The advised method of rater training was frame of reference training.

Frame of Reference Training (FOR)

A performance rating is often given based on an observation of a peer or subordinate's performance. As this is a subjective measure of performance, several errors may affect the ratings given by the supervisor (Woehr & Huffcutt, 1994). Halo error, for example, is a common rater error in which the rater scores performance in all dimensions similar to a performance in one dimension. Central tendency error can happen when a rater provides average ratings across dimensions. This can occur because the rater fears giving poor or above average ratings, the rater did not observe an adequate amount of behavior and defaults to an average rating, or various other reasons. Woehr and Huffcutt also noted that leniency and severity errors are also common among raters. A lenient rater typically provides ratings that are generally higher than the performance deserved, while severe raters typically provide ratings that are generally lower than the performance deserved.

Several training techniques have been developed in order to attempt to alleviate some of these problems. Woehr and Huffcutt (1994) conducted a meta-analysis of four commonly used training techniques: rater error training, performance dimension training, frame of reference training (FOR), and behavioral observation training. In rater error training, the trainees are taught how to guard against common rating errors such as halo,

leniency, and severity. Performance dimension training introduces trainees to the dimensions of performance being used in the ratings. In behavioral observation training, the focus is on the raters' observation of behavior rather than their evaluations of behavior. In FOR training, trainees learn about the multidimensionality of performance, the performance dimensions being rated receive a sample behavior from each dimension being rated, and practice making ratings and receive feedback on those ratings.

Woehr and Huffcutt (1994) located studies which empirically tested the effectiveness of these types of rater training. Each study was then coded by one of the authors based on the type of rater training that was investigated and the dependent measure used in that study. The dependent measures investigated were halo error, leniency error, rating accuracy, and observational accuracy. By combining these four dependent measures with the four types of training, the authors created a grid of 16 unique combinations. This allowed for each cell in the grid to contain one training type and one dependent measure.

All four of the training types were shown to decrease the raters' incidence of halo, central tendency, leniency, and severity errors. FOR training, however, was found to result in the largest increase in rating accuracy (Woehr & Huffcutt 1994). The authors suggested that this result could be a consequence of the raters being trained on a specific theory of performance which increases rater accuracy when this theory is applied to actual evaluation. This is possibly caused by the fact that those who received FOR training were evaluating performance using what Woehr and Huffcutt called "expert rater" standards, and thus the raters should have produced "expert ratings", and similarly had a higher level of agreement.

Chirico et al. (2004) believed that rating accuracy is improved through FOR training by creating a shared understanding amongst raters of the performance dimensions and standards for evaluating behavior(s) relevant to those standards. This same process that increases rater accuracy may also increase rater agreement. The authors stated that FOR training creates common expectations of performance amongst raters. As for why FOR training is more effective than other measures of training at producing more accurate ratings, the authors suggested that those trained using FOR training can better remember the content presented during the training than those trained using other techniques. Essentially, this allows FOR trained raters to form more accurate impressions of the ratee within different performance dimensions. Chirico et al. ultimately concluded that FOR training produced more effective/accurate raters because the FOR training teaches the raters how to better categorize information about the ratee's behaviors. This better categorization ultimately leads to better information retention and recall on the part of the rater when actually giving the performance ratings.

In explaining the effects of FOR training on raters, Gorman and Rentsch (2009) claimed that FOR training has a positive influence on the rater's processing and representation of information, as well as the amount of information retained by the rater. The more accurate the retention and the higher the amount of information retained should ultimately allow raters to have a high level of agreement, assuming they have interpreted behaviors or items the same way. Ultimately, raters will also have to accurately recall the information retained in order to reach high levels of agreement. Roch and O'Sullivan (2003) investigated rater training issues of recall and time. The authors noted that multitudes of research have shown FOR training to be highly effective at increasing

rating accuracy. Since specific behavioral feedback is necessary to improve performance, FOR training is well worth the time, cost, and effort for an organization. Likewise, this feedback could potentially lead to a better understanding of how to rate behaviors, and thus increase rater agreement.

However, Roch and O'Sullivan (2003) also noted that even with FOR training, raters could provide accurate ratings, but not actually be correctly identifying behaviors. They point to the fact that raters usually do not have behavioral statements about a ratee's performance available, yet the rater is required to recall specific behaviors in order to provide accurate feedback. If the raters are not recalling correct behaviors, this could be a source for lack of agreement.

Roch and O'Sullivan (2003) claimed that one major benefit of FOR training is that it allows raters to develop prototypes representing differing levels of performance, which the rater then translates into categories of performance. This can be both a positive and a negative. While this allows for higher recognition of behaviors, it also allows an opportunity for behaviors that were not actually performed to be "observed" and recalled by the rater simply because that behavior fit into a category that a behavior actually performed was grouped into. In fact, Sulsky and Day (1992) found that FOR trained raters recognized and recorded behaviors that did not actually occur at a higher rate than those not receiving FOR training, which could cause a problem with agreement. Another problem that raters may face is that while ratees will undoubtedly perform prototypical behaviors, not every ratee will exhibit only prototypical behaviors. If a category for an exhibited behavior is not established in the FOR training, that behavior could go unrated

or be improperly rated by the rater. Roch and O'Sullivan suggested that by adding behavioral observation training to FOR training, this problem could be reduced.

Roch and O'Sullivan (2003) found that FOR trained raters provided more accurate ratings than the control group which received no training. They also found that FOR trained raters recalled more behaviors and more behaviors that actually occurred than the untrained raters. However, Roch and O'Sullivan noted that although the FOR trained raters did recall more correct behaviors (those that actually happened); they were not necessarily increasing the quality of the recalled behaviors. The authors posited that this may be due to FOR training increasing the use of categories which may lead to raters recalling behaviors prototypical of a performance category, but that were not necessarily shown by the ratee. These findings led the authors to believe that FOR training does not improve observation or memory. Their findings indicated that raters may be recalling behaviors learned in training rather than behaviors actually portrayed by the ratee. Should this be true, it would certainly effect the level of rater agreement as raters might be missing behaviors that actually occurred which could, in turn, result in lower scores than the ratee deserved.

Research such as Woehr and Huffcutt (1994) and Chirico et al. (2004) have shown that frame of reference training is the best rater training available in terms of improving rater accuracy. Chirico et al. also showed that qualitative scores work just as well as quantitative scores. Roch and O'Sullivan (2003) provided evidence that FOR training also leads to raters recalling more behaviors. Jackson, Atkins, Fletcher, and Stillman (2005) provided a real-world setting and application of FOR training that showed the benefits of FOR training seen in experimental settings can be obtained in

field settings as well. FOR training has a wide array of benefits, most of which are derived directly from the process of FOR training. By establishing a shared perception among raters of what constitutes good versus bad performance based on expert ratings, FOR training increases rating accuracy and will likely increase rater agreement as well.

Current Study

The purpose of the current study was to investigate the effect of FOR training on the relationship between ratings of varying levels of behavioral observability and rater agreement. Based on a review of the literature, it was hypothesized that FOR training would provide raters with a common perception of what constitutes good and bad performance and, as such, was expected to reverse the findings of Roch et al. (2009). In other words, trained raters would be less likely to default to a general impression on less observable items and thus would demonstrate higher rater agreement on more observable items than untrained raters.

Hypothesis 1: There will be a difference in the relationship between rater agreement and behavioral observability between trained and untrained raters.

Hypothesis 1a: Ratings from untrained raters will have a negative relationship between rater agreement and behavioral observability.

Hypothesis 1b: Ratings from FOR trained raters will have a positive relationship between rater agreement and behavioral observability.

Method

Participants

Sixty-six undergraduate students attending a southeastern university participated in the study. Thirty-four participants were in the control group receiving no rater training, while thirty-two participants were in the treatment group that received frame of reference training. The control group consisted of a mean participant age of 19.09 years. Of the participants, there were 19 females and 15 males in the control group. The participants consisted of 23 Caucasian, 7 African-American, 1 Hispanic, 1 Asian, and 2 Other ethnicities. The treatment group had a mean age of 19.88 years of age and included 18 females and 14 males. There were 25 Caucasians, 6 African-Americans, no Hispanics or Asians, and one participant did not provide their ethnicity. The participants were all enrolled in a psychology course that either required participation in a research project or allowed students to earn extra credit for their course.

Stimulus Performance

A twenty-five minute videotape of a leaderless group discussion was used to present the performance information. The videotape was originally used as a practice videotape for training raters for an assessment center. The videotape depicted four people role playing in a leaderless group discussion. Of the four people in the videotape, two were male and two were female. One of the male participants in the videotape was used as the target individual for both groups. He was chosen for observation due to his dominance of the group, which provided more observable behaviors.

Rating Form

The rating form consisted of 85 items. The rating form is the same rating form used in the original study conducted by Roch et al. (2009) and may be found in Appendix A. Each of the items belonged to one of the following dimensions: team skills, oral communication, professionalism, or problem solving. Often associated with leadership, these dimensions are commonly rated in assessment centers. Each dimension contained approximately the same amount of items and levels of behavioral observability. The rating form consisted of two five-point Likert scales: *Performance Ratings* which ranged from 1 “not at all” to 5 “to a very great extent”, and *Difficulty of Rating* which ranged from 1 “very easy to rate” to 5 “very difficult to rate.”

Item Observability

Item observability ratings were originally established by Roch et al. (2009) by using four expert raters (upper level Ph.D. students enrolled in an upper level performance appraisal seminar). The expert raters were given a paragraph with examples of items that varied in levels of observability. Once they were familiar with how to rate observability, the expert raters worked with Dr. Roch to determine the level of observability of each item based on the extent to which the behavior could be directly observed and the extent to which judgment was needed to answer each item. These ratings may be found in Appendix B.

Procedure

Three to ten participants took part in each experimental session. The sessions were randomly assigned to be either the control or treatment condition. Participants were unaware of the assigned condition of the session when they registered for a session.

Participants in the treatment groups, upon signing the informed consent form, received a brief, thirty minute frame of reference training. This included instruction on what behaviors should receive a high or low rating, as well as viewing a video, providing performance ratings for a target individual in the video, and receiving feedback on their ratings. The practice videotape, also a leaderless group discussion involving two males and two females, was the stimulus performance videotape used in Roch et al. (2009). The control groups, upon signing the informed consent form, received a thirty minute presentation about assessment centers. This allowed for an equal amount of time spent by the participants in each trial. At the end of thirty minutes, both the treatment and control groups viewed the videotaped leaderless group discussion and then rated the performance of the target male (and the difficulty associated with rating each item as part of a larger study). At the end of the ratings, the participants completed the demographics questionnaire. Each experimental session lasted between an hour and an hour and a half.

Results

Hypothesis 1, which posed a difference in agreement and behavioral observability between trained and untrained raters, was analyzed by correlating r_{wg} and item observability (Roch et al., 2009) and the testing the difference between the two correlations. Specifically, control group ($r = -.299, p < .01$) and the treatment group ($r = -.304, p < .01$) correlations were both found to be significant and were then analyzed using a two correlation samples z-test ($z = .0213$) to test the difference between the correlations. No significant differences were found between the correlations. Hypotheses 1a, that ratings from untrained raters would have a negative relationship between rater agreement and behavioral observability, and 1b, that ratings from FOR trained raters will have a positive relationship between rater agreement and behavioral observability, were measured by testing for significant correlations. Both correlations were significant at alpha .01, however, only Hypothesis 1a was correct in regard to the predicted direction of a negative relationship between rater agreement and behavioral observability. The treatment group was also found to have a negative correlation between rater agreement and behavioral observability. These results are consistent with the findings in the previous study conducted by Roch et al. (2009).

Discussion

This study was designed to extend and help provide insight to the study conducted by Roch et al. (2009) that investigated rater agreement in relation to item observability. While the current study used the same rating form and item observability scores as the Roch et al. study, the current study differed in that it added FOR training, and used a different videotaped leaderless group discussion for the rating process. The FOR training stemmed from the future research recommendations in the Roch et al. study. It was believed that the inclusion of FOR training would increase the differences between groups in regards to the correlations between rater agreement and behavioral observability.

Despite the change in the videotape, and the inclusion of FOR training, the results of the current study replicated the results of the original study conducted by Roch et al. (2009) in that rater agreement for untrained raters would be negatively correlated with the behavioral observability of items. The hypothesis that rater agreement between trained raters would be positively correlated with the behavioral observability of items was not supported. In fact, the two correlations were nearly identical, indicating that the FOR training had little to no effect on the level of rater agreement in relation to the behavioral observability of the items in the rating form. The first hypothesis that there would be a stronger relationship between rater agreement and behavioral observability for the trained raters also failed to be supported.

These results could be a consequence of raters in both groups reverting to a default score or a general impression of the target individual's performance. This possibility, however, was not assessed during the course of this study. Ultimately, this

could lead to a conclusion that the levels of agreement are more related to a general evaluation of the target individual rather than the specific behaviors targeted by the items on the rating form. Another potential explanation is that the participants could have been exhibiting halo error. Should this be the case, the halo error could have influenced the dimension ratings (Balzer & Sulsky, 1992).

Limitations

There were several limitations to this study. The first and most obvious limitation was the small number participants. This resulted in low power. Faul, Erdfelder, Lang, and Buchner (2007) noted that significance tests that lack statistical power are of limited use because they cannot reliably discriminate between the null hypothesis and the hypothesis of interest. As such, these results must be taken with a grain of salt. However, the differences in the correlations across conditions are so small (and in the same direction) that while the power was low, it is unlikely that larger samples would have changed the results.

A more likely explanation of the results of this study was the time constraint put on the length of the FOR training in the treatment condition. As students were participating in the experiment for course credit, the sessions had to be kept to an hour and a half in duration. This only allowed for an extremely brief 30 minute FOR training. Typically, FOR training will take several hours and include at least half an hour of practice time. In this study, the students received only 5 to 10 minutes of actual practice and feedback on the scores they provided during the practice session, limited introduction to the performance dimensions, and only minimal explanation of how to provide ratings. Likewise, there was no time available to evaluate the success of the training, and as such,

there was not an indication of whether or not the participants acquired the necessary knowledge to provide quality ratings. It is believed that the results of this study might have been as originally hypothesized had there been an adequate amount of training time for the FOR training.

Another limitation to this study is that the participants used were students, and as such, the participants may have lacked the motivation to provide their best efforts. The students knew that they would receive full credit for participating in the trial whether they provided accurate ratings or not. Likewise, trials were conducted during the day around their class times, so participants likely came into the study feeling mentally fatigued and then participated in either a 30 minute lecture or a 30 minute training session before watching a roughly 30 minute videotape and providing 170 ratings (performance and item difficulty ratings for 85 ratings). Thus, rater fatigue and a lack of rater motivation were likely major confounds in the study.

Future Research

As noted previously, FOR training is the best rater training technique for improving rater accuracy (Woehr & Huffcutt, 1994) and accurate ratings are expected to improve rater agreement. Future research should allow for a more thorough and complete frame of reference training. Specifically, the training should include a more thorough explanation of the performance dimensions, as well as more time to view the practice video, the opportunity to practice on a wider array of items, and additional time to explain any discrepancies between the trainee ratings and the true scores in order to allow for a complete understanding of what is expected when providing ratings.

In accordance with the suggestions presented by Roch et al. (2009), future research should also investigate the effects of a default or general impression response to items that are less observable. As has been previously mentioned, this type of responding could lead to higher rater agreement on a general set of behaviors rather than on the specific behaviors of each item. This could be done by replicating this study and including a general impression question. Furthermore, this question should be divided amongst the participant rating forms in a manner that allows half of the participants to provide the general impression rating prior to providing the performance ratings and the other half of participants to rate the general impression of the target individual after having provided performance ratings. This will allow the researchers to detect potential biases resulting from both the general impression of the target individual's performance as well as from the placement of the general impression question itself.

Another recommendation for future research is to conduct a study with a shorter rating form. It is possible that having an 85-item rating form resulted in rater fatigue and possibly effected the motivation of the raters in a negative manner. Finally, Brutus and Facticeau (2003) suggested that supervisors would be less likely to be influenced by individual item characteristics since they have experience in providing ratings. As such, it is recommended that the participants in future replication studies be supervisors, or at minimum, individuals with supervisory experience. This could potentially eliminate poor levels of motivation that result from having student participants, as well as allow for more experienced raters to provide ratings since the participants would likely have had some experience with providing performance ratings of some type in their work history.

In conclusion, it is still valuable to investigate the relationship between rater agreement and the observability of items as performance ratings are constantly used in the business world as a matter of record keeping and in selection processes. It is still believed by the author that frame of reference training, if adequately provided, will shed light on the impact of varying levels of item observability and rater agreement on performance ratings. Once a firm understanding of how the observability of items impacts performance ratings is understood, managers will be able to provide more consistent and accurate performance ratings.

References

- Balzer, W. K., & Sulsky, L. M. (1992) Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975-985.
- Brown, R. (2006). *Fundamentals of Psychological Testing*. Author.
- Brutus, S., & Fecteau, J. (2003). Short, simple, and specific: The influence of item Design characteristics in multi-source assessment contexts. *International Journal of Selection and Assessment, 11*, 313-325.
- Chirico, K. E., Buckley, M. R., Wheeler, A. R., Fecteau, J. D., Bernardin, H. J., & Beu, D. S. (2004). A note on the need for true scores in frame-of-reference (FOR) training research. *Journal of Managerial Issues, 26*, 382-395.
- Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.
- Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*, 1336-1344.
- Jackson, D. J., Atkins, S. G., Fletcher, R. B., & Stillman, J. A. (2005). Frame of reference training for assessment centers: Effects on interrater reliability when rating behaviors and traits. *Public Personnel Management, 34*, 17-25.
- Littlefield, J. H., & Troendle, G. R. (1986, April). *Rating format effects on rater agreement and reliability*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training and Development, 7*, 93-107.
- Roch, S. G., Paquin, A.R., & Littlejohn, T. W. (2009). Do raters agree more on observable items? *Human Performance, 22*(5), 391-409.

- Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*, 501-510.
- Van Hooft, E. A., Van Der Flier, H., & Minne, M. R. (2006). Construct validity of multi-source performance ratings: An examination of the relationship of self-, supervisor-, and peer-ratings with cognitive and personality measures. *International Journal of Selection and Assessment, 14*, 67-82.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 64*, 189-205.
- Wohlers, A. J., & London, M. L., (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self awareness. *Personnel Psychology, 42*, 235-261.
- Wu, S. M., Whiteside, U., & Neighbors, C. (2007). Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy, 36*, 230-239.

Appendix A:

Rating Form

Please use the following rating scales to rate the performance of the target person and the difficulty of the item. Remember when rating performance you are rating the target persons performance in the leaderless group discussion. Be sure to rate their performance on every item. When you rate difficulty you are rating the difficulty of each item. Be sure that in rating performance you place your answer in the column labeled "Performance" and the column labeled "Difficulty" for rating difficulty.

Please use the following scales to rate the corresponding items.

Performance Rating

1 _____ 2 _____ 3 _____ 4 _____ 5 _____

Not at all

To a very great extent

Difficulty Rating

1 _____ 2 _____ 3 _____ 4 _____ 5 _____

Very easy

Very Difficult to rate

Item	Performance	Difficulty
1. Accepted other's ideas	① ② ③ ④ ⑤	① ② ③ ④ ⑤
2. Acted appropriately	① ② ③ ④ ⑤	① ② ③ ④ ⑤
3. Acted judiciously	① ② ③ ④ ⑤	① ② ③ ④ ⑤
4. Acted professionally	① ② ③ ④ ⑤	① ② ③ ④ ⑤
5. Acted with poise and maturity	① ② ③ ④ ⑤	① ② ③ ④ ⑤
6. Allowed another group member to speak by saying such things like "Mary has something to say" or "Let's hear what Joe has to say."	① ② ③ ④ ⑤	① ② ③ ④ ⑤
7. Analyzed problems well	① ② ③ ④ ⑤	① ② ③ ④ ⑤
8. Asked fellow group members if they all agreed either with own opinion or someone else's opinion.	① ② ③ ④ ⑤	① ② ③ ④ ⑤
9. Asked other team members for their opinions by saying such things as "What do you think?"	① ② ③ ④ ⑤	① ② ③ ④ ⑤
10. Asked others regarding the details of their plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
11. Asked the group how the group should proceed by saying such things as "what is our next step" or "what do you think we should do next."	① ② ③ ④ ⑤	① ② ③ ④ ⑤
12. Avoided use of speech crutches (such as "umm," "ah," and "err")	① ② ③ ④ ⑤	① ② ③ ④ ⑤
13. Behaved conscientiously	① ② ③ ④ ⑤	① ② ③ ④ ⑤
14. Behaved suitably	① ② ③ ④ ⑤	① ② ③ ④ ⑤

Performance Rating

1 2 3 4 5

Not at all

To a very great extent

Difficulty Rating

1 2 3 4 5

Very easy

Very Difficult to rate

Item	Performance	Difficulty
15. Blamed others or made excuses	① ② ③ ④ ⑤	① ② ③ ④ ⑤
16. Communicated effectively	① ② ③ ④ ⑤	① ② ③ ④ ⑤
17. Comprehended group functioning	① ② ③ ④ ⑤	① ② ③ ④ ⑤
18. Constructed clear sentences	① ② ③ ④ ⑤	① ② ③ ④ ⑤
19. Delivered message in a manner appropriate to audience	① ② ③ ④ ⑤	① ② ③ ④ ⑤
20. Delivered message in an effective manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
21. Delivered message in an enthusiastic manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
22. Delivered the message competently	① ② ③ ④ ⑤	① ② ③ ④ ⑤
23. Demonstrated an inappropriate sense of humor	① ② ③ ④ ⑤	① ② ③ ④ ⑤
24. Demonstrated appropriate body language	① ② ③ ④ ⑤	① ② ③ ④ ⑤
25. Dressed professionally	① ② ③ ④ ⑤	① ② ③ ④ ⑤
26. Gave consideration to others' plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
27. Had a good grasp of the problem	① ② ③ ④ ⑤	① ② ③ ④ ⑤
28. Had short hair	① ② ③ ④ ⑤	① ② ③ ④ ⑤
29. Helped to clarify group goals	① ② ③ ④ ⑤	① ② ③ ④ ⑤
30. Highlighted group functioning	① ② ③ ④ ⑤	① ② ③ ④ ⑤
31. Identified trade-offs	① ② ③ ④ ⑤	① ② ③ ④ ⑤
32. Included other team member's ideas in the solution	① ② ③ ④ ⑤	① ② ③ ④ ⑤
33. Integrated proposals from several team members	① ② ③ ④ ⑤	① ② ③ ④ ⑤
34. Knew how to resolve conflicts	① ② ③ ④ ⑤	① ② ③ ④ ⑤
35. Knew how to solve problems	① ② ③ ④ ⑤	① ② ③ ④ ⑤
36. Lost temper or appeared frustrated	① ② ③ ④ ⑤	① ② ③ ④ ⑤
37. Made eye contact with other people	① ② ③ ④ ⑤	① ② ③ ④ ⑤
38. Made inappropriate comments	① ② ③ ④ ⑤	① ② ③ ④ ⑤
39. Made logical arguments or statements	① ② ③ ④ ⑤	① ② ③ ④ ⑤
40. Mentioned possible solutions to the problem	① ② ③ ④ ⑤	① ② ③ ④ ⑤
41. Paid attention to others' plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
42. Perceived relationships among the plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
43. Pointed out problems with the plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
44. Praised other team members by saying such things as "good", "good idea", or "I like that" in response to their ideas.	① ② ③ ④ ⑤	① ② ③ ④ ⑤
45. Presented message in an organized manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
46. Processed information	① ② ③ ④ ⑤	① ② ③ ④ ⑤

Performance Rating

1 2 3 4 5

Not at all

To a very great extent

Difficulty Rating

1 2 3 4 5

Very easy

Very Difficult to rate

Item	Performance	Difficulty
47. Processed information effectively	① ② ③ ④ ⑤	① ② ③ ④ ⑤
48. Proposed an answer to the problem	① ② ③ ④ ⑤	① ② ③ ④ ⑤
49. Proposed priorities for the plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
50. Proposed solutions	① ② ③ ④ ⑤	① ② ③ ④ ⑤
51. Protected minority point of view	① ② ③ ④ ⑤	① ② ③ ④ ⑤
52. Provided clarification of the problem	① ② ③ ④ ⑤	① ② ③ ④ ⑤
53. Raised voice in response to others' comments	① ② ③ ④ ⑤	① ② ③ ④ ⑤
54. Rambled	① ② ③ ④ ⑤	① ② ③ ④ ⑤
55. Recognized strategic opportunities for success	① ② ③ ④ ⑤	① ② ③ ④ ⑤
56. Remained quiet while other people were speaking	① ② ③ ④ ⑤	① ② ③ ④ ⑤
57. Sat erect in his/her chair	① ② ③ ④ ⑤	① ② ③ ④ ⑤
58. Saw connections between plans	① ② ③ ④ ⑤	① ② ③ ④ ⑤
59. Saw how the plans fit together	① ② ③ ④ ⑤	① ② ③ ④ ⑤
60. Sifted irrelevant data	① ② ③ ④ ⑤	① ② ③ ④ ⑤
61. Sought consensus	① ② ③ ④ ⑤	① ② ③ ④ ⑤
62. Spoke in a concise manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
63. Spoke in a loud manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
64. Spoke well	① ② ③ ④ ⑤	① ② ③ ④ ⑤
65. Spoke with adequate volume and enunciation	① ② ③ ④ ⑤	① ② ③ ④ ⑤
66. Successfully involved others in group process	① ② ③ ④ ⑤	① ② ③ ④ ⑤
67. Summarized other people's views and questions	① ② ③ ④ ⑤	① ② ③ ④ ⑤
68. Supports others' viewpoints	① ② ③ ④ ⑤	① ② ③ ④ ⑤
69. The individual was an effective oral communicator	① ② ③ ④ ⑤	① ② ③ ④ ⑤
70. The person had effective team skills	① ② ③ ④ ⑤	① ② ③ ④ ⑤
71. The person was an effective problem solver.	① ② ③ ④ ⑤	① ② ③ ④ ⑤
72. Treated others in a professional manner	① ② ③ ④ ⑤	① ② ③ ④ ⑤
73. Tried to satisfy group goals	① ② ③ ④ ⑤	① ② ③ ④ ⑤
74. Twisted hair around fingers	① ② ③ ④ ⑤	① ② ③ ④ ⑤
75. Understood group functioning	① ② ③ ④ ⑤	① ② ③ ④ ⑤
76. Used a constructive approach to resolve conflicts	① ② ③ ④ ⑤	① ② ③ ④ ⑤
77. Used coarse or vulgar language	① ② ③ ④ ⑤	① ② ③ ④ ⑤
78. Used gestures fittingly	① ② ③ ④ ⑤	① ② ③ ④ ⑤

Performance Rating

1 _____ 2 _____ 3 _____ 4 _____ 5

Not at all

To a very great extent

Difficulty Rating

1 _____ 2 _____ 3 _____ 4 _____ 5

Very easy

Very Difficult to rate

Item	Performance	Difficulty
79. Used information from multiple sources	① ② ③ ④ ⑤	① ② ③ ④ ⑤
80. Used sound criteria for selecting options	① ② ③ ④ ⑤	① ② ③ ④ ⑤
81. Used suitable language	① ② ③ ④ ⑤	① ② ③ ④ ⑤
82. Used visual aids	① ② ③ ④ ⑤	① ② ③ ④ ⑤
83. Varied pitch of voice	① ② ③ ④ ⑤	① ② ③ ④ ⑤
84. Welcomed diverging viewpoints	① ② ③ ④ ⑤	① ② ③ ④ ⑤
85. Wore a vest	① ② ③ ④ ⑤	① ② ③ ④ ⑤

Appendix B:
Items, R_{wg} , & Item Observability

Item	Rwg - Control	Rwg - Treatment	Observability
1. Accepted other's ideas	0.67	0.80	2.25
2. Acted appropriately	0.62	0.71	1.75
3. Acted judiciously	0.69	0.68	2.00
4. Acted professionally	0.30	0.57	2.50
5. Acted with poise and maturity	0.29	0.57	2.75
6. Allowed another group member to speak by saying such things like "Mary has something to say" or "Let's hear what Joe has to say."	0.29	0.26	4.75
7. Analyzed problems well	0.53	0.79	1.50
8. Asked fellow group members if they all agreed either with own opinion or someone else's opinion.	0.66	0.32	5.00
9. Asked other team members for their opinions by saying such things as "What do you think?"	0.39	0.36	5.00
10. Asked others regarding the details of their plans	0.64	0.73	4.75
11. Asked the group how the group should proceed by saying such things as "what is our next step" or "what do you think we should do next."	0.60	0.52	4.75
12. Avoided use of speech crutches (such as "umm," "ah," and "err")	0.40	0.54	4.50
13. Behaved conscientiously	0.37	0.49	1.75
14. Behaved suitably	0.37	0.44	1.75
15. Blamed others or made excuses	0.63	0.33	3.50
16. Communicated effectively	0.70	0.85	1.75
17. Comprehended group functioning	0.74	0.71	1.25
18. Constructed clear sentences	0.42	0.71	3.00
19. Delivered message in a manner appropriate to audience	0.49	0.73	2.75
20. Delivered message in an effective manner	0.63	0.71	2.00
21. Delivered message in an enthusiastic manner	0.36	0.66	3.25

22. Delivered the message competently	0.62	0.58	2.25
23. Demonstrated an inappropriate sense of humor	0.19	0.32	2.5
24. Demonstrated appropriate body language	0.45	0.28	3.25
25. Dressed professionally	0.27	0.49	4.00
26. Gave consideration to others' plans	0.41	0.64	2.50
27. Had a good grasp of the problem	0.62	0.75	1.50
28. Had short hair	0.63	0.71	5.00
29. Helped to clarify group goals	0.64	0.69	3.00
30. Highlighted group functioning	0.63	0.71	3.25
31. Identified trade-offs	0.37	0.55	3.25
32. Included other team member's ideas in the solution	0.75	0.48	3.75
33. Integrated proposals from several team members	0.67	0.45	3.50
34. Knew how to resolve conflicts	0.60	0.77	2.75
35. Knew how to solve problems	0.55	0.73	2.75
36. Lost temper or appeared frustrated	0.30	0.75	4.75
37. Made eye contact with other people	0.20	0.47	5.00
38. Made inappropriate comments	0.07	0.73	3.25
39. Made logical arguments or statements	0.58	0.39	2.50
40. Mentioned possible solutions to the problem	0.56	0.68	4.50
41. Paid attention to others' plans	0.27	0.52	2.75
42. Perceived relationships among the plans	0.70	0.61	1.25
43. Pointed out problems with the plans	0.56	0.62	4.25
44. Praised other team members by saying such things as "good", "good idea", or "I like that" in response to their ideas.	0.27	0.22	5.00
45. Presented message in an organized manner	0.52	0.67	3.25
46. Processed information	0.66	0.60	1.00

47. Processed information effectively	0.73	0.58	1.25
48. Proposed an answer to the problem	0.60	0.78	4.00
49. Proposed priorities for the plans	0.62	0.71	3.75
50. Proposed solutions	0.68	0.82	4.00
51. Protected minority point of view	0.40	0.49	2.25
52. Provided clarification of the problem	0.53	0.57	2.75
53. Raised voice in response to others' comments	0.58	0.28	4.25
54. Rambled	0.14	0.61	4.00
55. Recognized strategic opportunities for success	0.57	0.71	1.50
56. Remained quiet while other people were speaking	0.50	0.50	4.75
57. Sat erect in his/her chair	0.39	0.46	4.75
58. Saw connections between plans	0.67	0.49	1.25
59. Saw how the plans fit together	0.48	0.44	1.25
60. Sifted irrelevant data	0.55	0.55	1.75
61. Sought consensus	0.61	0.55	3.50
62. Spoke in a concise manner	0.58	0.69	3.50
63. Spoke in a loud manner	0.51	0.32	4.00
64. Spoke well	0.43	0.65	2.50
65. Spoke with adequate volume and enunciation	0.46	0.65	3.25
66. Successfully involved others in group process	0.52	0.39	3.25
67. Summarized other people's views and questions	0.40	0.52	4.00
68. Supports others' viewpoints	0.54	0.54	2.25
69. The individual was an effective oral communicator	0.65	0.71	2.50
70. The person had effective team skills	0.51	0.58	2.25
71. The person was an effective problem solver.	0.62	0.70	2.25

72. Treated others in a professional manner	0.53	0.52	2.50
73. Tried to satisfy group goals	0.66	0.58	2.50
74. Twisted hair around fingers	0.72	0.70	5.00
75. Understood group functioning	0.67	0.50	1.25
76. Used a constructive approach to resolve conflicts	0.72	0.62	2.50
77. Used coarse or vulgar language	0.79	0.626	1.75
78. Used gestures fittingly	0.44	0.42	4.25
79. Used information from multiple sources	0.55	0.36	3.75
80. Used sound criteria for selecting options	0.55	0.47	3.75
81. Used suitable language	0.55	0.47	2.25
82. Used visual aids	0.58	0.23	3.50
83. Varied pitch of voice	0.34	0.37	4.75
84. Welcomed diverging viewpoints	0.55	0.19	4.50
85. Wore a vest	0.59	0.36	3.25