MULTISOURCE PERFORMANCE RATINGS: MEASUREMENT EQUIVALENCE ACROSS GENDER

by

Jacqueline Brooke Elkins

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Arts in Industrial-Organizational Psychology

Middle Tennessee State University
May 2015

Thesis Committee:

Dr. Mark C. Frame

Dr. Michael Hein

Dr. Ying Jin

ACKNOWLEDGMENTS

I would like to thank Dr. Mark Frame for his continuous encouragement, guidance, and understanding throughout this thesis process. His support was essential to the completion of this thesis within the desired time-frame, and I greatly appreciate the personal sacrifice of his time on my behalf. I will always remember and appreciate his eagerness to answer my questions and to provide humor and reassurance during stressful moments. I attribute much of my learning throughout this process to him, and I owe him my sincere gratitude for making the thesis experience rewarding.

I would also like to thank my committee members, Dr. Michael Hein and Dr. Ying Jin, for their contributions and time commitment to further develop my thesis. Their feedback and statistical wisdom were essential to its completion. I greatly valued their additional input and feedback that ensured the statistical rigor of this project.

I especially want to thank Dr. Dale Rose, President of Data Driven Decisions Inc. (3D Group) for providing the data used in this study. I am also thankful to John Diller of 3D Group for his invaluable support and assistance with the data provided.

In addition, I also owe a special thank you to Rachael Brooke Brashears as well as Dr. Mark Frame for assisting in coding the gender for the data. This was a tedious task and a significant time commitment, and I am thankful for their dedication to helping me.

Lastly, I want to thank my family and friends who have supported me not only throughout the completion of my thesis but throughout my graduate school career. I specifically want to thank the members of my cohort for being supportive and encouraging me along the way. I am eternally grateful for the support I have received that has enabled the success of this project.

ABSTRACT

Organizations often use 360 degree feedback to provide employees insight into their performance from multiple perspectives. However, for the feedback to be effective at modifying job behaviors, the feedback must be based on true differences in the individual's performance and not based on differences in raters' conceptualizations of the behavior constructs. To determine if the comparison of ratings across gender and rating source dyads is even appropriate, the purpose of this study was to determine to what degree there is measurement equivalence across gender (female, male) and rating source (self, direct report) dyads in 360 degree ratings of corporate leaders. The findings of this study reveal that the 360 degree rating instrument is not directly comparable across rating groups (gender and rating source) because measurement variance indicated that the instrument is not measuring the same underlying construct.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER I: LITERATURE REVIEW 1

Introduction 1

Literature Review 4

Agreement. 4

Convergence. 5

Measurement Equivalence. 6

Biases. 9

Gender. 10

Gender and Performance Appraisal. 11

Research Questions. 12

CHAPTER II: METHODS 14

Measure 14

Background. 14

Performance Evaluation. 14

Survey Structure. 15

Validity and Reliability. 15

Participants 16
Procedure 19
Data Collection Procedure. 19
Data Analysis Model. 19
Data Analysis Procedure. 21
How the Data was Analyzed. 22
CHAPTER III: RESULTS 23
Research Questions 23
Research Question 1. 23
Research Question 2. 25
Research Question 3. 28
Research Question 4. 29
Research Question 5. 29
Research Question 6. 30
CHAPTER IV: DISCUSSION 31
Limitations and Future Research 33
Conclusion 35
Practical Implications. 35
Research Implications. 35

REFERENCES 37

APPENDICES 42

Appendix A: IRB Approval 43

Appendix B: Leadership Navigator for Corporate Leaders Items 44

LIST OF TABLES

Table 1. Gender Dyad Count 18

Table 2. Measurement Equivalence for Male and Female Corporate Leaders Across Male and Female Sources (Direct Reports and Self) 24

Table 3. Average Competency Ratings for Rating Source and Rater Gender 26

Table 4. Pairwise Comparisons for Competency Rating Differences 27

LIST OF FIGURES

Figure 1. Modified Corporate Leader Model 20

Figure 2. Rater Gender by Rating Source Interaction for Inclusiveness Competency. ... 28

CHAPTER I: LITERATURE REVIEW

Introduction

Multisource, or 360-degree, feedback is commonly used in organizations for assessing performance (for individual development, performance management, and sometimes for administrative decisions). In the multisource feedback process, information is provided to an individual regarding their job performance based on ratings gathered from various sources themselves, their peers, their direct reports (those subordinates that report directly to them), and their supervisors. In some instances the sources that provide this information may also include customers and subordinates that do not report directly to them (often called indirect reports). Raters from each source are asked questions related to the employee's work behaviors using a standardized instrument. In traditional approaches to measuring performance, organizations provide employees with feedback from only their supervisors. However, multisource feedback seems to be more effective than the traditional supervisor-only feedback for an employee's performance (McGarvey & Smith, 1993). The purpose of this multiple perspective feedback is to provide the individual being rated with a range of insight into their performance. This additional insight is useful for measuring employee performance and developing performance based on suggestions from multiple sources that observe different components of the employee's job behavior (Vukotich, 2014).

Most commonly in organizations, the feedback provided in performance appraisals is used for either developmental or administrative purposes (Tornow, 1993; Hannum, 2007). When used for developmental purposes, the employee being rated uses

the feedback from others to improve their work behaviors. On the other hand, when used for administrative purposes, the performance ratings are used for pay raises, bonuses, internal selection, and promotion decisions. While we acknowledged that 360-degree feedback is used by some organizations for administrative decisions, the practice is not recommended (DeNisi & Kluger, 2000). Thus, the primary focus of the current study is the use of 360-degree feedback for developmental purposes.

The introduction of 360-degree feedback into organizations has allowed for the unique examination of ratings across sources. Research in the area focuses on how the performance ratings compare across multiple rating groups. The need for research in the area of 360-degree feedback stems from several assumptions made concerning multisource perspectives: multiple sources provide different and relevant perspectives about the individual's performance (Tornow, 1993), and raters conceptually define performance dimensions similarly. It is assumed that multiple sources likely have differing perspectives on an employee's performance, and therefore do not always rate an employee's performance similarly. As well, it is assumed that when raters rate an individual on a particular performance dimension or item, that all raters conceptually define that dimension similarly. Research aims to investigate whether there are significant performance rating differences among groups and to determine why differences may exist (Frame, 1999). Additionally, it also seeks to establish similarity among raters' conceptualizations of underlying performance constructs (Frame, 1999).

Much of the literature has examined the psychometric aspects of multisource performance ratings. These examinations have provided insight into the differences or similarities present when comparing ratings across sources. Measurement equivalence,

also called measurement invariance, is of particular concern to researchers when comparing ratings from different rating sources. Essentially, measurement equivalence in the 360-degree feedback realm ascertains the degree of similarity in the conceptualization of measurement constructs across raters when rating performance. When directly comparing raters, or ratings from different sources, it is assumed that the ratings reflect the same construct being measured (Hannum, 2007; Maurer, Raju, & Collins, 1998). According to this assumption, performance dimensions and items are conceptually defined the same by the self, peers, supervisors, and direct reports when rating an individual's performance. However, if sources do not perceive constructs similarly, the ratings for an individual's performance are not directly comparable across sources. Differences in performance feedback should reflect differences in performance rather than measurement variance (Frame, 1999). In order for the feedback to be effective at modifying the individual's job behaviors, the differences in performance feedback must be due to differences in the perception of his or her performance rather than measurement variance. Therefore, examining whether ratings have measurement equivalence is essential in order to determine if the comparison of ratings is even appropriate.

Previous research on measurement equivalence of multisource performance ratings has focused mostly on measurement equivalence among rating sources but also gender of the individual being rated (Frame, 1999; Woehr, Sheehan, & Bennett, 1999; Facteau & Craig, 2001; Hannum, 2007; Bynum, Hoffman, Meade, & Gentry, 2013). The proposed study sought to replicate and to extend previous findings on measurement equivalence among raters, specifically self and direct reports, and gender of the rater. Specifically, the proposed study replicated research conducted by Frame (1999). Frame

(1999) examined gender differences in self ratings and direct report ratings, comparing gender differences for each source in isolation of the other rating source for executive performance ratings. The current study replicated and extended this research by comparing gender self ratings to gender direct report ratings. In other words, male and female self ratings were compared to male and female direct report ratings. This was done for corporate leaders.

Literature Review

Previous research comparing ratings in 360-degree assessments has concentrated primarily on three areas: the agreement between raters and ratings, the convergence of raters and ratings, and the degree to which ratings have measurement equivalence across sources. Agreement between raters refers to the consistency of similar rating scores across multiple sources for an individual. It answers the question: do different sources evaluate an individual's performance with similar ratings? Convergence among raters refers to the degree to which ratings are similar among raters (Viswesvaran, Schmidt, & Ones, 2002). A lack of convergence among raters justifies the use of 360-degree evaluations because it supports the idea that utilizing different sources provides the employee with multiple differing perspectives on their performance. Measurement equivalence indicates the ratings can be compared across sources and justifies the use of multisource ratings, whereas measurement inequivalence indicates they cannot be compared (Bynum, Hoffman, Meade, & Gentry, 2013).

Agreement. Previous research has found that self and other rating agreement is related to performance and outcomes (Atwater, Ostroff, Yammarino, & Fleenor, 1998;

Frame, 2003). Atwater, Ostroff, Yammarino, and Fleenor (1998) found agreement between self ratings and subordinate ratings for managerial performance. However, their research suggests that assessing self ratings and other ratings based on the degree of agreement can have an impact on performance effectiveness. Specifically, individuals with high agreement on self and other ratings or who have self ratings that are significantly lower than other ratings, tend to have the highest effectiveness outcome scores.

Convergence. As previously mentioned, convergence is the degree of fit between ratings on a performance dimension. Research in this area examines mean differences and correlations among different rating sources (Frame, 1999). The degree of convergence among raters can be impacted by the level of convergence at the construct level.

Three major meta-analyses have examined the convergence among rating sources in multisource ratings. Harris and Schaubroeck (1988) compared average correlations of self ratings to peer and supervisor ratings as well as peer ratings to supervisor ratings. They found more agreement between peer-supervisor ratings than self-peer and self-supervisor ratings. Mabe and West (1982) examined average correlations between self, peer, and supervisor ratings and found that self ratings have a low correlation with other rating sources. Conway and Huffcutt (1997) examined average correlations between subordinate, supervisor, peer, and self-evaluations and found low correlations between rating sources. The findings from each of these three meta-analyses indicates that rating sources do not all converge on their rating of an individual.

Measurement Equivalence. Measurement equivalence differs from agreement and convergence because measurement equivalence determines whether ratings are comparable while agreement and convergence assess mean differences between raters. As previously mentioned, measurement equivalence is the similar conceptualization of measurement constructs across raters when rating an individual's performance. It is important to determine the measurement equivalence when examining 360 degree performance measures because only comparing scores across rating sources does not account for the potential differences in interpretation and conceptualization behind performance dimensions measured. Measurement equivalence must first be established in order to know whether scores should be compared across rating sources because if constructs are not perceived similarly by rating sources, the ratings cannot be compared to each other. Therefore, establishing measurement equivalence of a measure is important because it determines if the comparison of ratings is even appropriate. If measurement equivalence is established, it means the 360-degree instrument is defined similarly across sources (Greguras, 2005; Cheung, 1999; Vandenberg & Lance, 2000). In order to establish measurement equivalence of ratings across groups, factor loadings, error variances, and the variance covariance matrix for underlying constructs must be invariant, or the same. Measurement variance is a potential indication of rating biases (Greguras, 2005).

Much of the previous research conducted on measurement equivalence suggests that multi-source performance ratings are equivalent across sources (Hannum, 2007; Greguras, 2005; Diefendorff, Silverman, & Greguras, 2005; Facteau & Craig, 2001; Woehr, Sheehan, & Bennett, 1999; Cheung, 1999). Facteau & Craig (2001) found similar

conceptualizations regarding manager performance across self, peer, supervisor, and subordinate rating groups. Hannum (2007) found measurement equivalence for boss, peer, and direct report ratings when controlling for organizational level. Diefendorff, Silverman, and Greguras (2005) found that self, peer, and supervisor ratings were equivalent on performance dimensions for non-managers. Frame (1999) found measurement equivalence for self and direct report ratings as well as measurement equivalence across gender.

However, there is conflicting literature as well. Bynum, Hoffman, Meade, and Gentry (2013) found measurement variance across sources. The study also found that raters from different sources did not rate performance similarly. However, raters from the same source did provide similar ratings. Therefore, same source raters appeared to conceptualize performance similarly while different sources did not.

Differences in findings may be due to two different methods commonly used when assessing 360-degree measurement equivalence: confirmatory factor analysis (CFA) and item response theory (IRT). Each method offers unique information regarding the measure's conceptual similarity across different sources. The method employed depends on whether the researcher wants to examine ratings at the scale level or the item level. Differences in the findings on 360-degree measurement equivalence can possibly be attributed to the use of this different statistical methodology to assess measurement equivalence. To address this issue, Maurer, Raju, and Collins (1998) analyzed peer and subordinate ratings using both CFA and IRT in order to compare the results of the two statistical methods. They found that CFA and IRT produced similar results. As well, Facteau & Craig (2001) employed both methods in their comparison of groups of raters

and found somewhat similar results between the two methods. Although the research suggests CFA and IRT produce similar conclusions, the different methodology could lead to somewhat different conclusions about the results.

CFA assesses the measurement equivalence of an instrument through examining multiple factors (Greguras, 2005). Vadenberg and Lance (2000) suggest using the CFA method within hierarchically nested models. IRT assesses the measurement equivalence of an instrument at the scale and item levels through examining a single factor at a time (Greguras, 2005). Differential functioning is determined from either item or scale scores, which is the difference in expected scores on the items or scales in relation to others with similar standings on the construct (Greguras, 2005).

Although CFA and IRT procedures provide somewhat different information, this study will only use CFA to assess the measurement equivalence of the 360-degree instrument. One advantage to using CFA over IRT is that CFA procedures accommodate for measurement error (Greguras, 2005; Diefendorff, Silverman, & Greguras, 2005; Bollen, 1989). In addition, there are a few other reasons that CFA is being used for this particular study. In general, IRT is commonly used to analyze measurement equivalence at the item level while CFA is used for examining measurement equivalence across multiple factors or scales. While limited research has been conducted using graded responses in IRT for measurement equivalence, the majority of IRT research in the measurement equivalence realm has focused on dichotomously scored measures while CFA is more commonly performed with polytomously scored measures such as 360-degree ratings. Vandenberg and Lance (2000) suggest in their synthesis and review of the measurement equivalence literature that CFA identifies differences between groups better

than other procedures because it is the most comprehensive (Diefendorff, Silverman, Greguras, 2005). Therefore, the current study will be using the CFA method to assess measurement equivalence.

Biases. Incongruence among rating sources and measurement variance of the rating instrument can possibly be attributed to rating biases. The multisource feedback instrument is a subjective evaluation of an individual's performance from the perspective of multiple raters with varying evaluations. Given the nature of the 360-degree feedback instrument, biases in ratings of an individual's performance can arise.

Self-Rating Biases: Harris and Schaubroeck (1988) compared self ratings to peer and supervisor ratings as well as peer ratings to supervisor ratings. They found more agreement between peer-supervisor ratings than self-peer and self-supervisor ratings. Their meta-analysis concluded that disagreement between self ratings and peer-supervisor ratings occurred due to an egocentric bias because of attributions and moderated defensiveness.

Another similar explanation for these findings is commonly referred to as leniency bias. Leniency bias occurs when an individual overestimates or inflates performance ratings. Much of the research conducted has found inflated self ratings compared to others' ratings (Fleenor, McCauley, & Brutus, 1996; Atwater & Yammarino, 1992). In these studies, individuals tended to rate themselves more highly than their supervisors and peers.

Other's Rating Biases: Biases also occur when an individual is rating the performance of another individual. Halo bias refers to the tendency to rate an individual's performance based on an overall impression or evaluation of the individual. This prevents

variability in performance dimension ratings for that rater (Jones & Fletcher, 2002). Therefore, if an individual really likes the person they are rating, they may let this influence their ratings on each question favorably instead of really considering their behaviors. Thornton (1980) found that others' ratings typically exhibit higher rates of halo bias than do self ratings.

Gender. Previous research has heavily examined gender differences in performance ratings and potential biases impacting 360 performance ratings. Research on rating biases between men and women, however, has produced mixed findings.

Gender Self Ratings. Previous research on multisource ratings in field settings has found no significant gender differences in performance ratings between sources. Shore & Thornton (1986) did not find gender differences in ratings for self and supervisory ratings. Frame (1999) examined the relationship between executive level multi-source performance ratings and target gender and found that self ratings between men and women at the executive level demonstrate measurement equivalence. Additionally, men and women direct reports demonstrated measurement equivalence when rating the target. Therefore, women and men interpreted items the same when rating themselves and the target.

Moshavi, Brown, and Dodd (2003) examined self-other agreement in transformational leadership performance ratings in overestimators, underestimators, and those in-agreement. Their results showed that self ratings for male leaders were higher than female leaders' self ratings. However, ratings of others did not reflect any difference between the ratings for the male and female leaders. In terms of feedback, Roberts and Nolen-Hoeksema (1989) found that women view performance feedback from others as

more helpful than men. As well, Roberts and Nolen-Hoeksema (1994) found that women integrate the feedback of others' into their self-evaluations.

Gender of the Rater. Shore and Thorton (1986) summarized the findings of previous research on gender differences in self ratings and rating others. Previous research has found that men tend to rate their own performance higher than women do, and women tend to rate the performance of others higher than men do. Previous research has also found that other ratings tend to be higher for men than women (Eagly, Karrau, & Makjijani, 1992).

Gender and Performance Appraisal. Previous research on rater and ratee gender effects in performance ratings has produced mixed and inconclusive results (Cochran, 1994). Due to the inconclusive previous research, Cochran (1994) conducted a study examining rater and ratee gender effects in performance appraisals. She found that the interaction of gender of the rater and ratee did not influence performance ratings. However, she did find interesting results regarding differences in male and female perceptions of performance dimensions. Women tended to think of self-management and interpersonal skills as one skill whereas men had a tendency to separate the skill sets. Additionally, her results indicated that males distinguished influencing from thinking and decision-making, whereas females perceived thinking, influencing, and decision-making as similar. Cochran connects these perception differences in performance dimensions to gender stereotypes.

Very little research has examined measurement equivalence across gender in 360-degree assessments. Etchegaray (2007) found measurement equivalence across male and female executive direct report ratings. The finding is noteworthy because it indicates that

ratings within the direct report rating source are comparable for male and female raters. To date, this is the only research that has been conducted on gender dyads which has examined measurement equivalence. The current study differs from this research by examining both self and direct reports across gender for corporate leaders, or mid-level managers. Unlike Etchegaray (2007), the present study makes use of an entire 360 degree feedback measure and analyzes all of the sub-factors (dimensions) as part of one model rather than using selected sub-scales assessed independently.

Research Questions. The following research questions were explored using a 360 degree feedback tool which assesses eight corporate leader specific competencies:

Question 1: Is there measurement equivalence for the corporate leader model measured in 360 degree ratings among male and female self ratings and direct report ratings?

Question 2: If there is measurement equivalence in 360 degree ratings among male and female self ratings and direct report ratings, are there mean differences in 360 degree ratings for male and female ratees?

Question 3: Is there measurement equivalence for the corporate leader model measured in 360 degree ratings among same gender dyads for self ratings and direct report ratings?

Question 4: If there is measurement equivalence in 360 degree ratings among same gender dyads for self ratings and direct report ratings, are there mean differences in 360 degree ratings between same gender dyads for self ratings and direct report ratings?

Question 5: Is there measurement equivalence for the corporate leader model measured in 360 degree ratings among mixed gender dyads for self ratings and direct report ratings?

Question 6: If there is measurement equivalence in 360 degree ratings among mixed gender dyads for self ratings and direct report ratings, are there mean differences in 360 degree ratings between mixed gender dyads for self ratings and direct report ratings?

CHAPTER II: METHODS

Measure

Background. The data used in this study was archival data provided by Data Driven Decisions, Inc. (3D Group) from their 360 degree assessment tool, Leadership Navigator. Leadership Navigator surveys were developed by Industrial-Organizational Psychologists through job analysis techniques and a careful focus on item development (Healy & Rose, 2003). 3D Group offers seven custom-tailored 360 degree feedback surveys for senior executives, executives, corporate leaders, individual contributors, executive directors, organizational leaders, and retail managers. Results are delivered to participants in the form of a report for individual or group feedback. Reports contain detailed information regarding their ratings on each dimension and provide leaders with overall ratings from each group of raters. Reports provide participants with strengths and opportunities for development by highlighting the top 10 behaviors and the bottom 10 behaviors rated by raters. This section is followed by all open-ended comments provided by raters relating to the leader's strengths and opportunities for development. In addition to the report, participants are provided with an interpretation guide to help them understand how to read their strengths and developmental areas within the report. Leadership feedback coaches are also offered for deeper analysis of results, for determining how to respond to results, and for guidance in action planning steps based on aggregate results.

Performance Evaluation. Leadership Navigator is comprised of different factors and competencies for each of the seven custom surveys. This study analyzed data for

corporate leaders. Two overarching factors are contained within the survey for corporate leaders: *Interpersonal Effectiveness* and *Work Process*. The Interpersonal Effectiveness factor focuses on meeting the needs of employees through managing and interacting. The Work Process factor focuses on the duties required of a mid-level manager. These two factors consist of eight competencies, or leader workplace behaviors. The Interpersonal Effectiveness factor consists of the following competencies/behaviors: Developing Talent; Inclusiveness; Team Leadership; Integrity; and Communication. The Work Process factor consists of the following competencies/behaviors: Business Focus; Results Orientation; and Customer Focus. This study analyzed the two factors and each of their underlying competencies.

Survey Structure. The corporate leader 360-degree Leadership Navigator survey consists of 50 closed-ended questions in addition to the opportunity for open-ended feedback. Raters indicated how often the corporate leader engaged in the behavior on a 6-point Likert scale: 1=*Almost Never*, 2=*Sometimes*, 3=*Frequently*, 4=*Almost Always*, 5=*Always*, and "not applicable/do not know." Respondents rate questions such as "Sets challenging, yet appropriate, goals" and "Asks clarifying questions to confirm understanding."

Validity and Reliability. Validation and reliability studies on all of the custom Leadership Navigator 360 assessments have suggested them to be valid and reliable 360-degree instruments (Healy & Rose, 2003; Robinson & Rose, 2004; Robinson & Rose, 2006). 3D Group conducted a study specifically evaluating the reliability, validity, and norms of their custom corporate leader 360-degree instrument (English & Rose, 2010). Reliability of each competency was calculated using Cronbach's alpha: Communication

(α = .91), Integrity (α = .86), Business Focus (α = .89), Results Orientation (α = .92), Customer Focus (α = .88), Team Leadership (α = .94), Developing Talent (α = .92), and Inclusiveness (α = .91). Each of the eight executive competencies demonstrates acceptable and more than acceptable reliability values.

Evidence for construct validity was initially supported through positive correlations among all eight competencies, which are commonly related to performance at the corporate leader level. Additionally, the structure of each of the two factors, Interpersonal Effectiveness and Work Process, and their corresponding competencies was analyzed using correlations. Analyses revealed that competencies were more strongly related to their corresponding factor than the other factor. Interpersonal Effectiveness competencies were correlated with their factor composite at .90 and correlated with the Work Process factor at .75. Work Process competencies were correlated with their factor composite at .83 and correlated with the Interpersonal Effectiveness factor composite at .74. A follow-up factor analysis confirmed the corporate leader two factor conceptual model.

Participants

The current study used data consisting of 360-degree performance ratings for corporate leaders. Corporate leaders being rated included mid-level managers. These employees were rated by six different sources: self, peer, boss, direct report, board member, and other. This study analyzed only self and direct report ratings. The original data file contained 487 corporate leaders and 2146 ratings from the direct reports.

Participants and raters were not required to indicate their gender in the survey. Therefore, three researchers determined the gender of each participant (self ratings) and rater (direct reports) based on the participant's or rater's first name by coding each as male or female. Researchers left the field blank for names that were either gender neutral or not easily identified as male or female. Inter-rater reliability estimates across the three researchers were calculated and ranged from .927 to .985. Any participant or rater with at least one blank field across researchers was removed from the data. The rest of the participants and raters that were not agreed upon by all researchers were also removed. Lastly, any participant or rater with missing data was deleted from the data prior to running analyses in AMOS. This resulted in 2418 total ratings (447 corporate leaders self ratings and 1971 direct reports. Of the 447 corporate leaders, 135 were females and 312 were males. Of the 1971 direct reports, 721 were females and 1250 were males.

In order to create self-direct report dyads, self ratings were paired with each direct report rating. In some instances, there were self ratings that did not have any corresponding direct report ratings and in other instances there were direct report ratings and no self ratings available (due to the data cleaning procedures outlined earlier). In those cases the data associated were deleted. In most cases, each self rating was matched with more than one direct report rating. Thus the number of dyads was determined by the number of direct report ratings and self ratings were repeated for each associated direct report. For example the self ratings of manager X were paired with the direct report ratings of person A, person B, person C, and Person D, and the self ratings of manager Y were paired with the direct report ratings of person E, person F, and Person G. So the corresponding dyads would be X-A, X-B, X-C, X-D, Y-E, Y-F, and Y-G.

For analysis in AMOS, the dyads were divided into four separate data sets: female corporate leaders rated by female direct reports (female self, female direct report dyad), male corporate leaders rated by male direct reports (male self, male direct report dyad), female corporate leaders rated by male direct reports (female self, male direct report dyad), and male corporate leaders rated by female direct reports (male self, female direct report dyad). Finally, the four data files were divided into a total of eight data sets because each file had to be divided into two files: one for self ratings and one for direct report ratings. Separating the data sets allowed for model estimates to be calculated in AMOS since both direct report ratings for each question and self ratings for each question cannot be contained within the same file for AMOS to calculate analyses. Table 1 presents the amount of corporate leaders being rated by direct reports within the gender dyads.

Table 1.

Gender Dyad Count

	Corporate Leader Gender		
Direct Report Gender	Female	Male	Total
Female	265	373	638
Male	209	866	1075
Total	474	1239	

N = 1713.

Procedure

Data Collection Procedure. Results from this 360-degree survey were used for developmental purposes and not personnel decisions. Participants complete a custom content 360-degree survey based on their role in the organization through a paper-based or online version of the survey. For online surveys, a 3D Group coach distributes the survey to the participant, supervisors, direct reports, and peers via e-mail. Individuals will receive reminder e-mails if they have not completed the survey within a certain timeframe after the initial e-mail was sent. Once 3D Group receives all necessary responses, the data is analyzed and compiled into reports to be delivered to the organization and leaders.

Data Analysis Model. Confirmatory Factor Analysis (CFA) was used to analyze measurement equivalence. AMOS 22.0, or Analysis of Moment Structures, was used to conduct CFA on the raw data provided by 3D Group. Mean differences were analyzed using a variety of analysis of variance procedures in SPSS 22.0.0.

The corporate leader model used to analyze measurement equivalence in this study is a working model that was established in a previous study (Elkins, Frame, Hein, & Rose, 2015). That study improved the model fit of the initial model using CFA to determine measurement equivalence across self and direct reports. The resulting model can be seen in Figure 1.

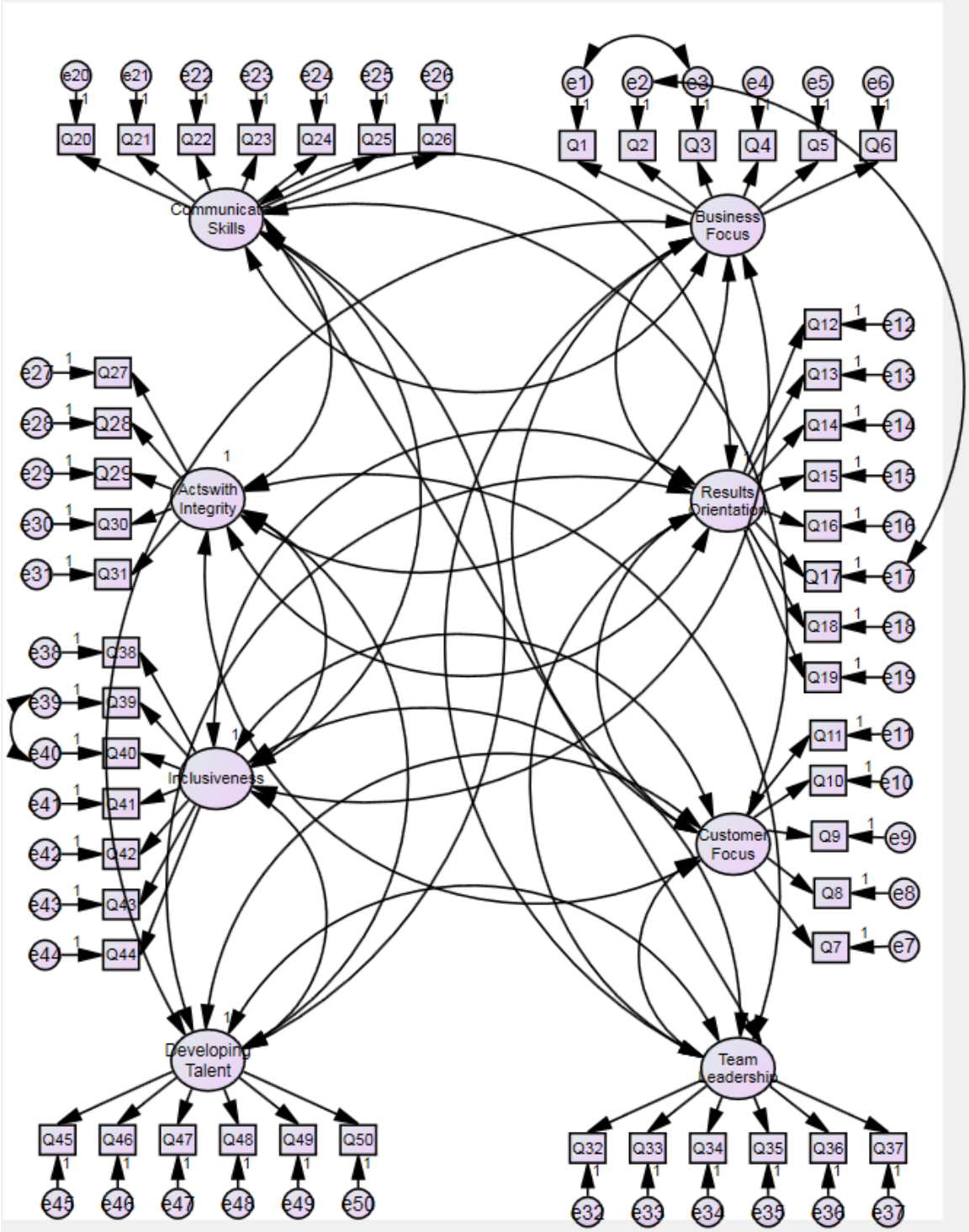

Figure 1. Modified Corporate Leader Model.

Data Analysis Procedure. Research questions guided data analyses for each of the conditions: gender comparisons for source and gender dyad comparisons. Data analyses provided results for the measurement equivalence for each of these conditions. Mean differences were analyzed among the groups if measurement equivalence was found but were not conducted if measurement variance was found. The degree to which measurement equivalence is established aids in our ability to interpret mean differences. If measurement variance was found and mean differences were analyzed, then the reason behind any differences found could not be established. It would be unknown if real differences in ratings or item interpretation caused mean difference ratings. Analyzing mean differences if measurement equivalence was not established was outside the scope of this study.

Measurement equivalence was demonstrated through several statistics derived from CFA procedures including Goodness of Fit indices: Chi-Square, Goodness of Fit Index (GFI), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), Root Mean Square Residual (RMR), and Adjusted Goodness of Fit Index (AGFI). Goodness of Fit indices establish how well the model fits the data compared to another model. Goodness of Fit indices determined whether follow-up mean differences will be analyzed.

First, analyses were run to determine whether there was measurement equivalence for the corporate leader model among male and female self ratings and direct report ratings. Then, if there was measurement equivalence, overall mean differences in

competency ratings between male and female self and direct report ratings were analyzed. Next, analyses were run to determine whether there was measurement equivalence for the corporate leader model among self ratings and direct report ratings for same gender dyads as well as mixed gender dyads. Then, if there was measurement equivalence, overall mean differences in competency ratings between same gender dyads and mixed gender dyads for self and direct report ratings were analyzed.

How the Data was Analyzed. The data was analyzed for measurement equivalence using Confirmatory Factor Analysis (CFA). Simultaneous multiple group CFA was used to determine the degree of measurement equivalence between rating sources (self, direct report as well as between genders (males, females)). Separate CFA analyses was used to determine the degree of measurement equivalence between same gender dyads and mixed gender dyads. AMOS 22.0 was used to conduct CFA on the raw data provided by 3D Group. In addition, a variety of analysis of variance procedures in SPSS were used to determine mean differences in ratings for each of these conditions if measurement equivalence was found.

CHAPTER III: RESULTS

Research Questions

Simultaneous Confirmatory Factor Analysis (SCFA) was conducted for research questions 1, 3, and 5 to determine the measurement equivalence across gender (female, male) and rating source (self, direct reports) for corporate leader 360 degree ratings. For research question 1, two SCFAs were conducted to determine measurement equivalence first across female self and male self ratings, and then female direct report and male direct report ratings. For research question 3, two SCFAs were conducted to determine measurement equivalence across same gender dyads, one for male self and direct reports and one for female self and direct reports. Finally, for research question 5, two SCFAs were conducted to determine measurement equivalence across mixed gender dyads, one for female self and male direct reports and one for male self and female direct reports. Research questions 2, 4, and 6 required conducting analysis of variance procedures only if measurement equivalence was found. Although measurement equivalence was not found, analysis of variance procedures were conducted for research question 2.

Research Question 1. The first research question asked whether there is measurement equivalence among male and female self ratings and direct report ratings. Two separate SCFAs were performed to examine measurement equivalence across gender for self and direct reports. The results of the two SCFAs are presented in Table 2.

Table 2.

Measurement Equivalence for Male and Female Corporate Leaders Across Male and Female Sources (Direct Reports and Self)

Research Question	Comparison	Chi Square	GFI	CFI	RMSEA	RMR	AGFI
1	Female Self vs. Male Self	4812.06	0.725	0.815	0.047	0.086	0.693
	Female DR vs. Male DR	10673.83	0.809	0.873	0.041	0.150	0.787
3	Female Self vs. Female DR	9314.03	0.594	0.671	0.076	0.162	0.548
	Male Self vs. Male DR	13289.61	0.763	0.786	0.053	0.101	0.736
5	Female Self vs. Male DR	9317.92	0.568	0.607	0.086	0.118	0.519
	Male Self vs. Female DR	7896.80	0.696	0.764	0.057	0.126	0.661

df = 2288.

First, the self rating SCFA for research question 1 was performed to determine measurement equivalence across 135 female self ratings and 312 male self ratings. Although the GFI and CFI values did not meet acceptable criteria indicating good fit at values greater than 0.90, the RMSEA value of .047 did meet the acceptable criteria of less than .05 (Browne & Cudeck, 1993).

Then, the direct report rating SCFA for research question 1 was performed to determine measurement equivalence across 721 female direct report ratings and 1250 male direct report ratings. Although the GFI and CFI values were higher than the female self and male self ratings fit, they still did not meet the acceptable criteria level of greater than .90. Similar to the previous SCFA, however, the RMSEA value of .041 did meet the acceptable criteria of less than .05.

Overall, taking into consideration all of the goodness of fit indices, the results from both SCFAs indicate measurement variance across female self and male self ratings as well as female direct report and male direct report ratings in this sample.

Research Question 2. The second research question indicated that mean differences would only be analyzed if measurement equivalence was found among male and female self ratings and direct report ratings. However, analyses were run to determine if male and female ratings were similar even though the items were not conceptually defined similarly. Therefore, a multivariate analysis of variance (MANOVA) was conducted to assess rating differences on rating source and gender.

Before conducting a two-way MANOVA, average ratings for each of the eight competencies were calculated into separate variables so that an average score for each of the eight competencies was calculated for each rater and participant. Table 3 presents the competency averages by rating source and gender.

Overall, the results from the two-way MANOVA revealed a significant interaction between rating source and gender, Wilks' Lambda $F(8,2407) = 2.53$, $p < .05$. There was also a main effect found for rating source, Wilks' Lambda F (8, 2407) = 0.91, $p < .05$, but there was not a significant main effect found for rater gender.

To determine if individual competencies were significantly different among gender or rating source, univariate tests were analyzed. Univariate tests for rater gender indicated significant differences for Business Focus ($F(1, 3) = 8.80$, $p < .01$) and Customer Focus ($F(1, 3) = 7.70$, $p < .01$). As shown in Table 4, pairwise comparisons indicated that male raters tended to rate Business Focus and Customer Focus higher than female raters. The descriptive statistics show that both male self and direct reports rated themselves more highly on Business Focus and Customer focus than both female self and direct reports, as can be seen in Table 3.

Table 3.

Average Competency Ratings for Rating Source and Rater Gender

Competency	Rating Source	Rater Gender	*M*	*SD*	*N*
Business Focus	Self	Female	3.97	0.71	135
		Male	4.11	0.57	312
	Direct Report	Female	3.86	1.13	721
		Male	4.04	0.99	1250
Customer Focus	Self	Female	3.71	1.19	135
		Male	3.92	0.93	312
	Direct Report	Female	3.56	1.42	721
		Male	3.74	1.26	1250
Results Orientation	Self	Female	3.94	0.61	135
		Male	3.94	0.58	312
	Direct Report	Female	3.73	1.07	721
		Male	3.90	0.93	1250
Communication Skills	Self	Female	3.89	0.59	135
		Male	3.99	0.56	312
	Direct Report	Female	4.02	0.92	721
		Male	4.07	0.80	1250
Acts with Integrity	Self	Female	4.38	0.55	135
		Male	4.37	0.50	312
	Direct Report	Female	3.93	1.07	721
		Male	4.12	0.93	1250
Team Leadership	Self	Female	3.84	0.96	135
		Male	3.85	0.81	312
	Direct Report	Female	3.65	1.24	721
		Male	3.83	1.09	1250
Inclusiveness	Self	Female	4.28	0.52	135
		Male	4.22	0.55	312
	Direct Report	Female	3.70	1.14	721
		Male	3.87	1.03	1250
Developing Talent	Self	Female	3.58	1.09	135
		Male	3.65	0.95	312
	Direct Report	Female	3.31	1.37	721
		Male	3.48	1.27	1250

Table 4.

Pairwise Comparisons for Competency Rating Differences

Competency	(I)	(J)	Mean Difference (I -J)	95% CI Lower Bound	95% CI Upper Bound
Rater Gender					
Business Focus	Male	Female	0.16*	0.06	0.27
Customer Focus	Male	Female	0.20*	0.06	0.34
Rating Source					
Customer Focus	Self	Direct Report	.16*	0.02	0.31
Results Orientation	Self	Direct Report	.13*	0.03	0.23
Communication Skills	Direct Report	Self	.11*	0.02	0.20
Acts with Integrity	Self	Direct Report	.36*	0.25	0.46
Inclusiveness	Self	Direct Report	.46*	0.35	0.57
Developing Talent	Self	Direct Report	.22*	0.08	0.36

* The mean difference is significant at .05.

Univariate tests for rating source indicated significant differences for Customer Focus ($F(1, 3) = 5.25$, $p < .05$), Results Orientation ($F(1, 3) = 6.06$, $p < .05$), Communication Skills ($F(1, 3) = 5.60$, $p < .05$), Acts with Integrity ($F(1, 3) = 46.71$, $p < .001$), Inclusiveness ($F(1, 3) = 66.76$, $p < .001$), and Developing Talent ($F(1, 3) = 9.55$, $p < .01$). As can be seen in Table 4, Pairwise comparisons indicated that self reports rated themselves higher for Customer Focus, Results Orientation, Acts with Integrity, Inclusiveness, and Developing Talent compared to their direct reports. Direct reports rated their corporate leader higher on Communication Skills than the corporate leader did. Table 3 presents means for each competency by rating source.

Univariate tests for the interaction between rating source and gender indicated a significant difference for Inclusiveness, $F(1, 3) = 3.96$, $p < .05$. Since Inclusiveness was

not a significant main effect for rater gender but a significant main effect for rating source, it is likely that rating source was driving this interaction effect. Figure 2 illustrates the relationship between the mean differences in ratings for Inclusiveness by rater gender and rating source.

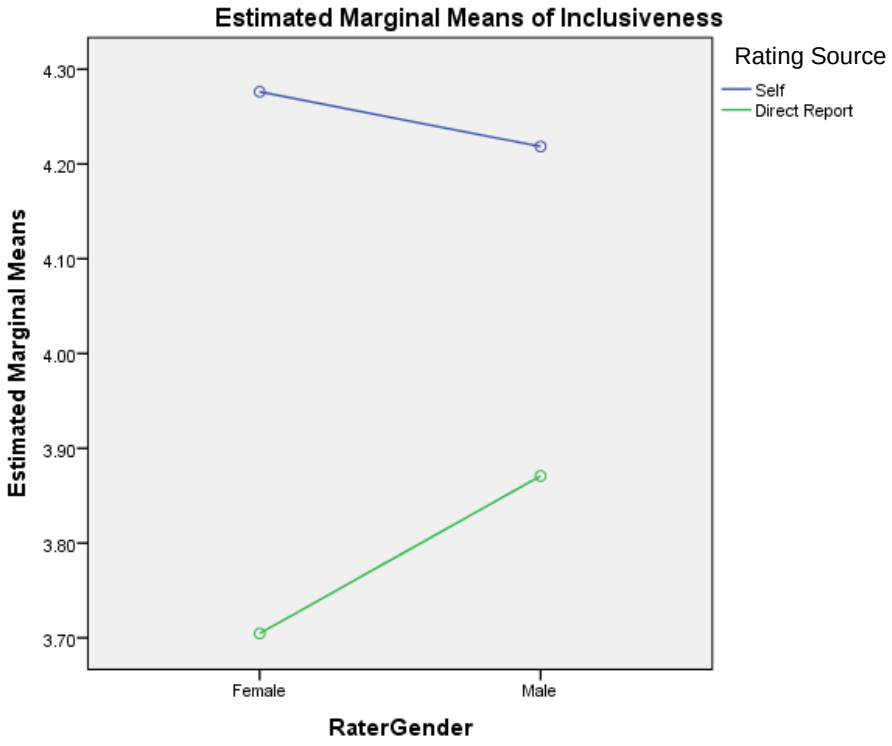

Figure 2. Rater Gender by Rating Source Interaction for Inclusiveness Competency.

Research Question 3. The third research question asked whether there is measurement equivalence among same gender dyads for self and direct report ratings. Two separate SCFAs were conducted to examine measurement equivalence across

female self and female direct report ratings versus male self and male direct report ratings. The results of the SCFAs can be seen in Table 2.

First, the female same gender dyad rating SCFA for research question 3 was conducted to determine measurement equivalence across 265 female self ratings and female direct report ratings. While the chi-square statistic indicated a good model fit, additional goodness of fit indices were evaluated given the sensitivity of chi-square to large sample sizes (Facteau & Craig, 2001). The GFI and CFI were both well below the acceptable criteria of above 0.9. As well, the RMSEA value of .076 was near the 0.08 value, indicating poor fit (Browne & Cudeck, 1993).

Then, the male same gender dyad rating SCFA for research question 3 was conducted to determine the measurement equivalence across 866 male self ratings and male direct report ratings. The chi-square statistics did not indicate a good model fit. Similar to the female same gender dyad SCFA, the GFI and CFI values were not considered acceptable. In addition, the RMSEA value was above the acceptable level of 0.05. Although the male self ratings and male direct report ratings exhibit goodness of fit indices closer to acceptable levels than the female self and direct report ratings, they are still not considered acceptable.

Overall, the goodness of fit indices as a whole indicate that across same gender dyads, there is measurement variance.

Research Question 4. Mean differences across same gender dyads were not examined since results indicated measurement variance.

Research Question 5. The fifth research question asked whether there is measurement equivalence among mixed gender dyads for self and direct report ratings.

Two separate SCFAs were conducted to examine measurement equivalence across female self and male direct report ratings versus male self and female direct report ratings. The results of the SCFAs can be seen in Table 2.

The female-male mixed gender dyad rating SCFA for research question 5 examined the measurement equivalence across female self ratings and male direct report ratings. While the chi-square statistic indicated a good model fit, additional goodness of fit indices were evaluated. The GFI and CFI were both well below the acceptable criteria of above .9, and the RMSEA value of .086 indicates poor fit (Browne & Cudeck, 1993).

The male-female mixed gender dyad rating SCFA for research question 5 examined the measurement equivalence across male self ratings and female direct report ratings. The chi-square statistic indicated a good fit, but the GFI and CFI were well below 0.90, and the RMSEA did not meet the acceptable criteria of below 0.05.

Overall, the SCFAs indicated measurement variance among mixed gender dyads for self and direct report ratings.

Research Question 6. Mean differences across mixed gender dyads were not examined since results indicated measurement variance.

CHAPTER IV: DISCUSSION

The purpose of this study was to determine the degree of measurement equivalence in 360 degree ratings across gender and rating source dyads. It is important to establish whether raters across gender and rating source use the instrument similarly. Overall, results indicated measurement variance across male and female raters for both self and direct report ratings and for both same and mixed gender dyads. After reviewing relevant literature on 360 degree ratings and measurement equivalence across gender, it was determined there was very little research conducted on the comparison of male and female ratings in relation to each other and in combination with their rating source.

The first research question examined whether there was measurement equivalence across male and female self ratings and male and female direct report ratings of corporate leaders. This research question was replicated from a previous study conducted on executive self and direct report 360 degree ratings by Frame (1999). This study differed from Frame's study by examining corporate leaders rather than executives, and this study examined the whole 360 degree model rather than the individual factors that make up the model. The previous study found measurement equivalence among male and female executive self ratings and measurement equivalence among male and female direct report ratings of executives. In contrast to those findings, the current study found measurement variance across male and female corporate leader self ratings as well as measurement variance across male and female direct report ratings of corporate leaders. The findings of this study do not support previous research on measurement equivalence and indicate that males and females do demonstrate measurement variance in performance ratings by

different rating sources. This means that ratings across male and female self corporate leader ratings are not comparable and that male and female direct report ratings of corporate leaders are not comparable.

Based on these findings, mean differences across male and female self and direct reports were analyzed to assess rating agreement across males and females for self and direct reports. The findings revealed differences in male and female ratings of Business Focus and Customer Focus, and self and direct report differences for Customer Focus, Results Orientation, Communication Skills, Acts with Integrity, Inclusiveness, and Developing Talent. The majority of significant differences were seen in rating source with self ratings rating themselves higher than direct report ratings. The only dimension on which an interaction between gender and rating source was found was for the dimension of Inclusiveness.

The third research question examined whether there was measurement equivalence across same gender dyads for self and direct report ratings of corporate leaders. In addition, the fifth research question in this study examined whether there was measurement equivalence in corporate leader ratings among mixed gender dyads. These research questions are similar to research conducted by Etchegaray (2007) who examined measurement equivalence across gender dyads in executive direct report ratings and found measurement equivalence across all gender dyads. The current study differs from this research in a few ways. First, the current study examined both self and direct reports across gender for corporate leaders. In addition, the present study analyzed the entire 360 degree measure and analyzed the corporate leader model as a whole rather than analyzing selected sub-scales. As well, there was enough variability in rating responses within this

study that collapsing response options was not needed (as was needed in Etchegaray, 2007). Therefore, this study was able to utilize the full range of the response scale. In contrast to Etchegaray (2007), the present findings indicate that participants' use of the measure varied across same gender and mixed gender dyads. This indicates that no matter which gender is rating either perspective, the rating scale is being used differently across gender and rating sources.

Findings from this study are significant because they reflect measurement variance across the entire model. Unlike Etchegaray (2007), this study examined measurement equivalence across the entire model rather than assessing selected sub-scales. The advantage to using the current study's approach is it provides information on differences in the use of the overall 360 degree instrument rather than just components of the instrument. Based on the findings from this study, the same underlying factor structure is not being assessed by each gender and rating source across the 360 degree instrument as a whole. Unfortunately, this means that ratings are not directly comparable across sources or gender and that differences in performance feedback may not reflect actual differences in performance. Therefore, feedback given to the individual based on their 360 ratings across self and direct reports may not be effective at modifying job behaviors.

Limitations and Future Research

Limitations of this study include the shrinking of the sample size due to coding rater gender. Although inter-rater reliability among researchers coding males and females was above .90, much of the sample size was reduced by deleting participants with gender

neutral names. While the reduced sample size was not ideal, it was necessary to delete these participants from the data to ensure all of the rater genders were agreed upon across the three researchers.

Also, the use of only Confirmatory Factor Analysis to analyze the data could be considered a limitation. For the purposes of the study, it was decided to examine ratings at the scale level rather than the item level. Item response theory (IRT) offers a different lens to look at the data through at the item level, and it was not utilized in this study. However, Maurer, Raju, and Collins (1998) found CFA and IRT produced similar results when they analyzed peer and subordinate ratings using both CFA and IRT in order to compare the results of the two statistical methods. Despite this research, the use of only SCFA methodology without IRT methodology can be considered a limitation.

Another limitation of this study was only analyzing corporate leaders. Examining the measurement equivalence across other organizational levels, such as the executive level or organizational leader level (i.e., non-corporate leaders, such as mid-level managers at non-profit organizations, government agencies and community group), would be an area for future study. The examination of organizational leader self and direct report ratings would likely provide a more balanced number of male and female self ratings since there are more females at the organizational leader level.

Future research should examine measurement equivalence across gender dyads that are more female dominated or equally held by both males and females. The current research also makes a compelling argument that future measurement equivalence research in the 360 degree feedback realm should focus on analyzing the entire factor

structure of the measures and not limiting analyses to individual sub factors, dimensions, or competencies.

Conclusion

The purpose of this study was to determine the degree of measurement equivalence across gender and rating source dyads in corporate leader 360 degree ratings. Previous research has investigated measurement equivalence of executive 360 degree ratings across gender within self and direct reports (Frame, 1999) and gender dyads within direct reports (Etchegaray, 2007). This study differs from previous research by examining measurement equivalence across both self and direct reports and gender dyads on corporate leader 360 degree ratings.

Practical Implications. This research can help inform practices for developing improved 360 degree measures. This study found that 360 degree ratings are not directly comparable across gender or rating sources. Therefore, the differing perspectives of raters across sources and gender may not be indicative of actual differences in performance on the job. This means that feedback reports developed based on 360 degree ratings across gender and sources that are given to leaders for developmental planning may not be effective at modifying job behaviors. Studies such as this one can be used by practitioners to continually improve the psychometric properties of their feedback measures to ensure that leaders receive meaningful feedback for creating their development plan.

Research Implications. The findings of this study reveal that the 360 degree rating instrument is not directly comparable across rating groups (gender and rating source) because measurement variance indicated that the instrument is not measuring the

same underlying construct. The findings of this study highlight the need for further investigation of measurement equivalence in gender dyads and across the entire 360 degree measure factor structure.

In conclusion, the current study addresses implications for both practitioners and researchers concerning measurement equivalence across gender dyads and the entire 360 degree factor structure. This research and future research can help inform practitioners on practices for developing better 360 degree measures. As well, it highlights a future need for investigation of measurement equivalence across gender dyads and the entire factor model in 360 degree ratings for researchers.

REFERENCES

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology*, *51*, 577 – 598.

Atwater, L. E. & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology*, *45*, 141-164.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303 – 316.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage Publications, Inc.

Bynum, B. H., Hoffman, B. J., Meade, A. W., & Gentry, W. A. (2013). Reconsidering the equivalence of multisource performance ratings: Evidence for the importance and meaning of rater factors. *Journal of Business Psychology*, *28*, 203-219.

Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology, 52,* 1 – 36.

Cochran, C. (1994, April). Gender interaction effects on the level and structure of performance ratings. In *9th annual conference of the Society for Industrial and Organizational Psychology, Inc, Nashville, TN.*

Conway, J., & Huffcutt, A. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*(4), 331 – 360.

DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *Academy of Management Executive, 14*(1), 129 – 139.

Diefendorff, J. M., Silverman, S. B., & Greguras, G. J. (2005). Measurement equivalence and multisource performance ratings for non-managerial positions: Recommendations for research and practice. *Journal of Business and Psychology, 19*, 399 – 425.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111*, 3 – 22.

Elkins, J. B., Frame, M., Hein, M., & Rose, D. (2015, May). *Measurement equivalence across self and direct report 360-degree ratings.* Poster presentation at 27th annual Association for Psychological Science convention in New York City, NY.

English, A., & Rose, D. S. (2010). *2010 Normative comparison, reliability analysis, validity revisions report for the leadership navigator for corporate leaders* (White Paper). Berkeley, CA: Data Driven Decisions, Inc.

Etchegaray, J. M. (2007). Measurement equivalence of executives' performance ratings for same- and opposite-gender dyads. *Journal of Leadership Studies, 1*, 21 – 32.

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

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly, 7*, 487 – 506.

Frame, M. C. (1999). Executive level multi-rater performance ratings: Measurement equivalence across source and gender. (Master's thesis) Available from ProQuest Dissertations and Theses database.

Frame, M. C. (2003). Executive level multi-rater performance ratings: A study of agreement, gender, and outcomes. (Dissertation) Available from ProQuest Dissertations and Theses database.

Greguras, G. J. (2005). Managerial experience and the measurement equivalence of performance ratings. *Journal of Business and Psychology, 19,* 383-397.

Hannum, K. M. (2007). Measurement equivalence of 360-assessment data: Are different raters rating the same constructs? *International Journal of Selection and Assessment, 15,* 293 – 301.

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

Healy, M. C., & Rose, D. S. (2003, April). *Validation of a 360-degree feedback instrument against retail sales performance: Content matters.* Interactive poster session presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, Florida.

Jones, L. & Fletcher, C. (2002). Self-asessment in a selection situation: An evaluation of different measurement approaches. *Journal of Occupational and Organizational Psychology, 75,* 145-161.

Mabe, P. & West. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 37,* 687 – 702.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*(5), 693-702.

McGarvey, R. & Smith, S. (1993). When workers rate the boss. *Training, 30*(3), 31-34.

Moshavi, D., Brown, F. W., & Dodd, N. G. (2003). Leader self-awareness and its relationship to subordinate attitudes and performance. *Leadership & Organization Development Journal, 24*, 407 – 418.

Roberts, T., & Nolen-Hoeksema, S. (1994). Gender comparisons in responsiveness to others' evaluations in achievement settings. *Psychology of Women Quarterly, 18*, 221 – 240.

Roberts, T., & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles, 21*, 725 – 747.

Robinson, G.N. & Rose, D.S. (2006, May). *Reliability and construct validity of a 360° assessment survey for executives*. Interactive poster session presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, Texas.

Robinson, G. N. & Rose, D. S. (2004). *Development and Content Validation of the Leadership Navigator for Executives* (White paper). Berkeley, CA: Data Driven Decisions, Inc.

Shore, L. M. & Thornton III, G. C. (1986). Effects of gender on self- and supervisory ratings. *Academy of Management Journal, 29*(1), 115-129.

Thornton, III, G. (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology, 33*, 263 – 271.

Tornow, W. W. (Ed.). (1993). Special issue on 360-degree feedback. *Human Resource Management, 32*, 211 – 219.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, *87*, 345 – 354.

Vukotich, G. (2014). 360° feedback: Ready, fire, aim- issues with improper implementation. *Performance Improvement, 53*(1), 30-35.

Woehr, D. J., Sheehan, M. K., & Bennett, W. Jr. (1999). *Understanding disagreement across rating sources: An assessment of the measurement equivalence of raters in 360 degree feedback systems.* Poster presented at the Fourteenth Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

APPENDICES

Appendix A: IRB Approval

MIDDLE TENNESSEE STATE UNIVERSITY

3/18/2015

Investigator(s): Brooke Elkins, Mark Frame
Department: Psychology
Investigator(s) Email Address: jbe3i@mtmail.mtsu.edu; Mark.Frame@mtsu.edu

Protocol Title: Measurement Equivalence across Self and Direct Report Ratings of Corporate Leaders

Protocol Number: #15-206

Dear Investigator(s),

Your study has been designated to be exempt. The exemption is pursuant to 45 CFR 46.101(b)(4) Collection or Study of Existing Data.

We will contact you annually on the status of your project. If it is completed, we will close it out of our system. You do not need to complete a progress report and you will not need to complete a final report. It is important to note that your study is approved for the life of the project and does not have an expiration date.

The following changes must be reported to the Office of Compliance before they are initiated:

- Adding new subject population
- Adding a new investigator
- Adding new procedures (e.g., new survey; new questions to your survey)
- A change in funding source
- Any change that makes the study no longer eligible for exemption.

The following changes do not need to be reported to the Office of Compliance:

- Editorial or administrative revisions to the consent or other study documents
- Increasing or decreasing the number of subjects from your proposed population

If you encounter any serious unanticipated problems to participants, or if you have any questions as you conduct your research, please do not hesitate to contact us.

Sincerely,

Lauren K. Qualls, Graduate Assistant
Office of Compliance
615-494-8918

Appendix B: Leadership Navigator for Corporate Leaders Items

Business Focus:

1. Understands our company's industry.
2. Exercises fiscal responsibility and manages budgets appropriately.
3. Understands current market issues and market drivers.
4. Makes decisions based on company goals/strategy.
5. Advocates our company's strategic vision.
6. Faces the key challenges for the company's future.

Customer Focus:

7. Understands the needs of our company's most important customers.
8. Manages customer expectations.
9. Champions initiatives that expand customer base, sales, or market share.
10. Makes customers a top priority.
11. Understands the impact of his/her decisions on our customers.

Results Orientation:

12. Proactively addresses issues before they become problems.
13. Conveys a sense of urgency when necessary.
14. Uses company resources effectively (including staff, time, budget).
15. Effectively prioritizes initiatives, projects, and tasks.
16. Delegates initiatives, projects, and tasks appropriately.
17. Considers the financial impact of his/her decisions.
18. Sets challenging, yet appropriate, goals.
19. Stays abreast of progress on key projects, initiatives and goals.

Communication Skills:

20. Listens to others attentively.
21. Adjusts message according to the audience.
22. Expresses ideas clearly and concisely.
23. Speaks with confidence and credibility.
24. Shares information as needed by others.
25. Asks clarifying questions to confirm understanding.
26. Uses appropriate grammar and avoids jargon.

Acts with Integrity:

27. Says what he/she means.
28. Admits mistakes.
29. Follows through on commitments.
30. Is honest and forthcoming.

31. Does not take credit for others' work.

Team Leadership:

32. Makes sure his/her team has adequate resources to succeed.
33. Establishes clear expectations for his/her team.
34. Ensures that his/her team is working well together.
35. Leads by example.
36. Gets his/her team working toward shared goals.
37. Selects, develops, and retains high quality talent.

Inclusiveness:

38. Treats people with different backgrounds as equals.
39. Encourages others to express diverse opinions.
40. Values diversity.
41. Does not "play favorites."
42. Confronts inappropriate behavior in others.
43. Shows respect for others, regardless of position or background.
44. Considers alternative ideas and opinions when making decisions.

Developing Talent:

45. Understands strengths and weaknesses of his/her direct reports.
46. Sets appropriate development goals with direct reports.
47. Coaches direct reports and others when necessary.
48. Holds direct reports accountable for improving their skills.
49. Mentors others within our company.
50. Provides both positive and negative feedback in a constructive way.