ABSTRACT

MULLEN, TORREY RIESER. How Long Should we Follow the Leader? Using Latent Growth Models of Longitudinal Leadership Performance Change to Predict Leader Outcomes. (Under the direction of Samuel B. Pond and S. Bartholomew Craig.)

A paucity of research has examined longitudinal performance and the predictive ability of performance change on important outcomes. In addition, few studies have investigated the effects of rater variables on performance over time and the effect of rater group composition or rater perspective on longitudinal performance ratings. The purpose of this research was to investigate consequences related to rater characteristics including rater context, perspective and composition in the measurement and prediction of longitudinal performance. Results suggested that longitudinal self-ratings, boss ratings, and direct report ratings were equivalent. The results of this study also concur with earlier findings about the dynamic nature of performance (Thoreson, Bradley, Bliese & Thoreson, 2004). Longitudinal change in performance was found for every leadership performance factor in ratings from every rater group. Latent growth curves for all rater groups were remarkably similar although boss ratings showed the most consistent longitudinal change. Adding sector and/or subdivision covariates to the models improved model fit for each rating source group. Using growth mixture modeling with the rater context covariates allowed the estimation of latent classes that clarified the direction of leadership performance growth. Results also indicated the importance of rater composition. Direct reports who consistently rated the same leader tended to rate those leaders more highly than the direct reports who rated different leaders. Adding the composition moderator variable to the boss rating models improved model fit for four of the five leadership performance models. The composition covariate also significantly

predicted the intercept and slope for boss ratings of Ethics and Character, suggesting that obtaining leadership performance ratings from consistent bosses plays an important role in detecting linear change in leader performance, especially for ratings of Ethics and Character. Leaders with positive development on Ethics and Character had higher consensus performance scores, confirming past research showing that integrity and ethical behavior are important characteristics in successful managers (Posner & Schmidt, 1984; Mortensen, Smith, & Cavanagh, 1989).

HOW LONG SHOULD WE FOLLOW THE LEADER? USING LATENT GROWTH MODELS OF LONGITUDINAL LEADERSHIP PERFORMANCE CHANGE TO PREDICT LEADER OUTCOMES

by
TORREY RIESER MULLEN

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

PSYCHOLOGY

Raleigh, North Carolina

2007

APPROVED BY:

_____	_____
Dr. Samuel B. Pond	Dr. S. Bartholomew Craig
Co-Chair of Advisory Committee	Co-Chair of Advisory Committee
_____	_____
Dr. Mark A. Wilson	Dr. Lori Foster Thompson

Dedication

This dissertation is dedicated to my parents, Debby Messick and Geoff Rieser, who instilled the value of higher education in me at a very young age. With a lawyer and a doctor as parents, what else was I to do but earn my PhD? Ryan, you're next! In all sincerity, my family has provided a great deal of support and encouragement throughout my graduate career. I could not have done this without them.

Biography

The author was born Torrey Elizabeth Rieser on May 18, 1979 in San Diego, California. She graduated from West Forsyth High School in Clemmons, North Carolina in 1997 and received Bachelor of Arts degrees, with honors, in Psychology and Spanish from the University of North Carolina at Chapel Hill in 2001. After graduating from college, she moved to Raleigh, North Carolina to attend North Carolina State University where she earned her Master of Science degree in industrial/organizational psychology in 2003. Torrey is currently working at the National Center for O*NET Development where she leads teams dedicated to occupational research.

Acknowledgements

I would like to thank my friends and family for their encouragement and support throughout this process. I want to thank my advisor, Dr. Bob Pond, not only for his guidance throughout graduate school, but for his advice before I even applied to school at NC State. Thank you for helping me choose industrial/organizational psychology as a program of study and a career. I would like to thank my co-chair, Dr. Bart Craig, for his continuous encouragement and assistance in finding this dissertation project. I would also like to thank the other members of my doctoral committee: Dr. Mark Wilson and Dr. Lori Foster Thompson.

I would like to thank several members of my family. Thanks to my dad for his excellent advice and steadfast belief in me. Thanks to my mom for always looking on the bright side even when I did not want to. Thanks to my brother Ryan for supporting me in life and recognizing my need for independence. Thanks to Shannon for being the only person outside of school to read my thesis (sorry this dissertation is so long). Thanks to Bill for always having a kind word. Thanks to Graeme for always being awed by the number of years I've been in school. Finally, thanks to my friend and colleague, April Cantwell, for listening, commiserating, and mentoring me throughout this experience.

Table of Contents

Table of Tables

Table of Figures

CHAPTER ONE: INTRODUCTION

Overview of the Study

Until recently, most researchers investigating job performance have treated it as a stable construct, measuring employee performance with a cross-sectional design at only one point in time. However, job performance is something that evolves with time, necessitating a longitudinal research design (Thoreson, Bradley, Bliese & Thoreson, 2004). Past research has found that job performance can change idiosyncratically and systematically from one performance rating to another (Hofmann, Jacobs, & Gerras, 1992; Ployhart & Hakel, 1998). Despite the fact that past research has shown evidence of the dynamic nature of performance (Bass, 1962; Deadrick, Bennett, & Russell, 1997; Ghiselli, 1956; Ghiselli & Haire, 1960; Hofmann, Jacobs, & Gerras, 1992; Ployhart & Hakel, 1998), a paucity of research has examined longitudinal performance and the predictive ability of performance change on important outcomes.

Organizations have become increasingly reliant on multisource, or 360-degree, performance appraisals and feedback programs. Given the growing dependence on multiple raters and the importance of the job performance construct, factors that influence such performance ratings should be identified and quantified (Scullen, Mount, & Goff, 2000). Little research has investigated the effects of rater variables on performance over time, especially concerning a leadership-rating dimension, and prior research has not studied the effect of rater group composition or rater perspective on longitudinal performance ratings. In a recent study, Scullen and colleagues stated that "models seeking to explain performance ratings should include factors associated with the perspective of the rater (p. 967)." Past research has shown that raters from different

organizational perspectives are attuned to different behaviors, and thus provide unique information in their ratings (Scullen et al., 2000). These researchers have made an urgent call for additional research investigating the antecedents and consequences of job performance and rater perspective effects in multisource ratings.

Past studies of cross-sectional job performance capture a still photograph of an ever changing and evolving relationship, one that might lead to costly assumptions about performance appraisal and management systems. The present research will answer calls for the investigation of consequences related to rater characteristics including rater context, perspective and composition in the measurement and prediction of longitudinal performance. This research study will examine longitudinal executive-level leadership performance, its ability to predict objective performance outcomes, and the moderating effects of multiple raters in a large *Fortune* 100 global communications firm. Essentially, the study is designed to examine the validity of multisource repeated measure leadership performance appraisals in predicting promotion and job performance levels.

In addition to extending the research literature, the findings of this study add significantly to practice. For example, by examining the longitudinal effects of rater group composition, this research will help an organization determine if it should require that the same rater provide ratings to the same target year after year in their performance management system. It was expected that leadership performance change over time would predict an executive's promotion and job performance outcomes. It was also expected that the rate of performance change would differ according to both rater perspective and composition, such that utilizing consistent raters across all measurement

intervals would prove critical to understanding the true change in leadership performance and its resulting predictive ability.

General Research Questions

There are 7 research questions addressed by this study. 1) Is there measurement equivalence between rating source groups on the leadership performance construct? 2) Is there measurement equivalence across measurement occasions on the leadership performance construct? 3) What form does the latent growth curve take for leadership performance over a multi-year period? 4) Are longitudinal leadership performance trajectories similar for raters in different organizational perspectives (i.e., boss, self, direct report)? 5) Is the rate of leadership performance development contingent upon rater composition (i.e., consistent or inconsistent raters at each time interval)? 6) Are the effects of the exogenous variables (i.e., rater composition, sector and subdivision) partially or completely mediated by initial status and/or change in leadership performance ratings? 7) Does change in leadership performance predict objective performance outcomes, such as promotion and consensus performance level?

Leadership performance appraisal data obtained over a 3-year period combined with second-order factor latent growth modeling (SOF LGM) will be used to test the research questions listed above. SOF LGM is a process that places a confirmatory factor analytic structure on variables that are measured longitudinally. SOF LGM allows for the incorporation of hypothesized predictors and outcomes of individual differences in true initial status and change in longitudinally measured variables. It allows a test of the hypothesis that longitudinal leadership performance mediates relationships between rater characteristics and objective performance variables. Essentially, individual-level growth

trajectories will be fit to measures of leadership performance over 3 measurement occasions, and between-individual differences will be modeled where the exogenous rater variables are introduced as predictors of individual differences in the leadership performance growth trajectories.

Chapter 2 provides a background and rationale for pursuing each question, and presents the research models to be tested. Chapters 3 and 4 present the methodology and analyses used to test the proposed models and the results of those analyses. Finally, Chapter 5 presents a discussion of the findings, implications of the results, and suggestions for future research.

CHAPTER TWO: RESEARCH LITERATURE REVIEWED

The following chapter will review the research literature on multisource feedback systems, dynamic performance, rating source measurement equivalence, and latent growth modeling. Multisource feedback systems have become an important piece of many organizations' performance management systems. Feedback is often given and received annually, yet little research has examined the longitudinal effects of providing multisource feedback despite evidence that performance is dynamic. When job performance is considered a dynamic construct, evaluating performance trends and their resulting outcomes becomes essential. Research has also found that there may be inequivalence among rating sources. While proponents of multisource feedback systems point to the additional information gained by raters from different organizational perspectives, rating source inequivalence might obscure a target's longitudinal performance trend if data from all rating sources is combined. Before evaluating individual performance trends and their resulting outcomes, measurement equivalence of the leadership performance construct across time intervals must first be established to ensure that the same construct is measured and can be compared from year to year. Upon finding measurement equivalence, the most appropriate way of analyzing performance trends with exogenous and outcome variables is second-order factor latent growth modeling. The following pages will elaborate upon the direction provided above beginning with a review of multisource feedback systems.

Multisource Feedback Systems

Job performance is central to most organizational personnel decisions, including compensation, promotion, and retention practices and is an important source of developmental feedback (Borman, 1997). Multisource feedback systems, often called 360-degree feedback, have become increasingly popular, especially for leadership development purposes (Avolio, Sosik, Jung, & Berson, 2003). Supervisors, peers, customers, direct reports, and the target each play important roles in job performance assessment (Borman, 1997; Scullen, Mount, & Goff, 2000). Waldman and Atwater (1998) estimated that 20 to 25 percent of organizations use multisource performance ratings for developmental, learning, and/or performance evaluation purposes, and every year Fortune 500 companies spend hundreds of millions of dollars to implement feedback programs.

According to Waldman and Atwater (1998), the multisource feedback process involves several steps in order to reach the goal of improving leadership. First, the organization must identify observable behaviors that are important to the targets' success in the job. Next, the target is rated by any number of individuals, including self-ratings and ratings made by supervisors, peers, direct reports, and/or internal and external customers on an anonymous survey. Finally, the target receives a summary of the responses in a report, often with averaged scores from each rater group. In contrast to other organizational development (OD) techniques, multisource feedback is used to change the behavior of the leader, not organizational policies or job design.

Much of the increase in multisource feedback popularity is due to the idea that evaluating a target's performance from different perspectives can play a central role in

building a comprehensive picture of the target's affect on others (Day & Lance, 2004). Multisource feedback is used to increase the accuracy of a target's self-perception and to provide information about how others perceive the target's behavior (Waldman & Atwater, 1998). Research has shown improvements in overall performance following multisource feedback (Atwater, Roush, & Fischthal, 1995).

To improve performance, the target must first understand his or her job performance strengths and weaknesses. Evaluations from employees who work closely with the target are effective at giving insight to strengths and weaknesses (Borman, 1997). Due to the changing nature of jobs and increased teamwork, individuals might work more closely with their peers, direct reports or customers rather than their supervisors; thus, a greater number of vantage points with multisource feedback provide a broader picture of the target's performance than traditional supervisor ratings (Waldman & Atwater, 1998).

Information obtained from a multisource feedback procedure provides the opportunity for greater self-awareness and guides and motivates the target to change their behavior. Unfortunately, individuals do not tend to evaluate themselves accurately or similarly to how others evaluate them. Inaccurate self-perceptions are common because coworkers, supervisors, peers, and others tend to withhold negative feedback because they feel uncomfortable providing it (Waldman & Atwater, 1998). Even when others provide negative feedback, it is unlikely to be delivered in a way that allows the target to accept and efficiently use the information (Uggerslev & Sulsky, 2002). Therefore, the target suffers from a lack of information about how others perceive their behaviors, and often develops an overly positive impression of their own behavior. Without proper

performance feedback, a leader's self-perceptions might remain inaccurate and inappropriate or unsuccessful behaviors might never change.

Unfortunately, others' perceptions of a target's behavior might be more important than the target's self-perceptions. Others' ratings of a target's performance tend to relate more closely to objective criteria than do the target's self-ratings (Waldman & Atwater, 1998). Recent studies have also suggested that targets who inflate their self-ratings relative to others' ratings have poorer performance and are less effective than targets who rate themselves in greater agreement with others (Atwater & Yammarino, 1992). For a leader who must interact with others in order to accomplish organizational goals, others' perceptions of his or her behavior are more important than the leader's self-perceptions because others' perceptions are what influence how followers will react to the leader's behavior (Kaplan & Kaiser, 2003).

Acknowledging the problems with obtaining accurate feedback as noted above, Waldman and Atwater (1998) espouse four different types of feedback that a target may receive from a rating source (p. 83):

1. Confirmatory positive feedback;
2. Disconfirmatory negative feedback;
3. Confirmatory negative feedback; and
4. Disconfirmatory positive feedback.

Confirmatory positive and negative feedback occurs when the target receives high or low ratings respectively from a rating source, which confirms the target's high or low self-ratings. Disconfirmatory negative feedback occurs when the target receives ratings that are lower than his or her self-ratings. Finally, disconfirmatory positive feedback occurs when the target's self-ratings are lower than the ratings he or she receives from others. Targets will most readily accept confirmatory positive feedback. This type of

feedback can positively reinforce already desirable behaviors. However, the most common type of feedback is disconfirmatory negative feedback. This type of feedback, if accepted by the target as valid, is the most likely to motivate changes in behavior. Presumably, the discrepancy between self- and other ratings arouses dissonance and causes the target discomfort. To alleviate this discomfort, the target will likely become motivated to make behavioral corrections to minimize the discrepancy (Waldman & Atwater, 1998).

Within the context of these four feedback orientations, Waldman and Atwater (1998) further list 7 potential benefits of multisource feedback (p. 7). Feedback promotes:

1. Enhanced organizational involvement of those asked to provide the feedback;
2. Positive reinforcement for leaders' good performances;
3. Greater interest in feedback on the part of leaders;
4. Better communication between leaders and their followers, peers, customers, and superiors;
5. Improvements in leader behaviors;
6. Change in organizational culture toward more participation and openness; and
7. Additional sources of input into the formal performance appraisal process.

Despite the recent emphasis on employee participation, empowerment, and satisfaction surveys, many employees do not truly believe that their input is valued by the organization. Anonymous multisource feedback, however, provides clear signals to employees that their opinions are important. Consequently, most employees are eager to provide ratings (Waldman & Atwater, 1998). Positive feedback can be very reinforcing and motivating, especially if it is unexpected. While positive feedback might not indicate needed behavioral change, it can motivate a leader to sustain such positive performance so that he or she lives up to future expectations. Positive feedback may also signal to a leader that he or she no longer needs to spend as much attention or exert as much effort on a behavior that he or she worried was being underperformed. This feedback might

allow the leader to free cognitive resources and concentrate on other areas of weakness (Kaplan & Kaiser, 2003).

Receiving constructive feedback often results in greater interest in subsequent feedback. If a leader is motivated to make changes after an initial round of feedback, that leader will often desire subsequent feedback after making those changes to see if others have perceived and appreciated the changes (Waldman & Atwater, 1998). Multisource feedback tends to stimulate greater communication between the target and the raters as well as the organization and its employees. Multisource feedback programs can create opportunities for clarifying and strengthening an organization's values and vision by stimulating conversations and discussions around those topics. The feedback report can also be a valuable tool for participative goal setting between the target and his or her supervisor (Waldman & Atwater, 1998).

Despite all of the benefits of multisource feedback, Waldman and Atwater (1998) also highlight several potential hazards of such feedback. If rater surveys are not kept anonymous or very few raters are included in the rater group, employees who have contributed to the process may worry about retribution for poor ratings. Targets receiving negative feedback may become defensive, may try to deny the ratings, or may suffer a blow to their self-esteem. Without proper counseling or training, targets may misinterpret ratings upon receiving feedback reports, or they may have difficulty interpreting conflicting ratings from different groups. Targets may also try to exert pressure on employees to provide overly positive ratings just before ratings are requested in order to receive higher scores. Such hazards negatively affect the quality of feedback and the integrity of the feedback process.

The integrity of multisource feedback systems relies on the accuracy of the ratings provided. Views on multisource ratings differ between science and practice; however both perspectives rely on rating accuracy. Scientific interest places emphasis on rating accuracy and minimizing error so that an accurate description of the target's performance is described. Practical views focus on the usefulness of the ratings in improving and maintaining individual and organizational performance. Having good science that supports practice is a foundational idea in industrial/organizational psychology. Borman (1997) suggests that one important place that science can aid practice is by assessing whether additional rating sources provide incremental validity beyond ratings of a single source. Each rating source should provide unique data on a target's performance; thus, multiple perspectives of a target's behavior can lead to greater learning and appreciation for different viewpoints. This assumption implies a desire for low to moderate interrater agreement across sources and high interrater agreement within sources.

It is also generally assumed that the variance in observed performance ratings is accounted for by the target's actual job performance. However, a meta-analysis conducted by Bommer and colleagues indicated that corrected correlations between ratings and objective job performance measures are moderate, with a mean correlation of only .389 (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995). Objective performance measures are assumed to be free of systematic bias and random error because they are not contaminated by an individual's biases or opportunities to observe behavior. Bommer and colleagues concluded that ratings of performance cannot be substituted for objective measures of performance, but that both measures might contribute unique information about the performance construct. Given the growing

dependence on multiple raters and the importance of the job performance construct, factors that influence such performance ratings need further research (Scullen, Mount, & Goff, 2000).

Wherry and Bartlett (1982) supported the idea that three factors influence performance ratings. These factors include the target's actual job performance, rater biases in the observation and recall of that job performance, and measurement error. The target's performance is a function of true ability or aptitude and the influence of the environment. Raters observe a function of the target's performance as well as the bias of observation. What the rater later recalls and reports on a performance appraisal is a result of both their observations and their recall biases. Scullen and colleagues (2000) examined a model that measured the rating variance associated with five factors similar to the three factors described by Wherry and Bartlett (1982) in two separate samples of managers from a wide variety of industries. These five factors included the target's general level of performance, the target's performance on a specific dimension, the rater's idiosyncratic rating tendencies (i.e., halo and leniency error), the rater's organizational perspective (i.e., self, direct report, peer and boss), and random measurement error.

Results from the Scullen and colleagues study revealed that idiosyncratic rater effects accounted for the largest percentage of variance in observed ratings. Of the five factors measured, idiosyncratic rater effects accounted for 62% and 53% of the rating variance of target performance in the two data sets, perspective effects accounted for 9% and 7%, general performance accounted for 13% and 14%, dimensional performance accounted for 8% and 11%, and measurement error accounted for 11% and 18%. Idiosyncratic

effects in self-ratings accounting for the greatest (71%) amount of variance and boss ratings accounted for the lowest (51%) among rater groups. Peer ratings accounted for 62% variance and direct report ratings accounted for 64%. Among rating groups in organizational perspective, peer ratings accounted for 0% variance, bosses accounted for 11%, and direct reports accounted for 17%. These effects represent unique systematic rating variance from within each rater perspective, and not across rater perspectives. It seems that bosses and direct reports have relatively unique perspectives when it comes to rating the targets' performance, meaning that each contributed unique and possibly valuable information to understand a target's performance.

The combined effects of the general and dimensional performance factors accounted for approximately 25% of the rating variance, less than half that of the idiosyncratic rater effects. Boss ratings of actual performance accounted for approximately 27% of the rating variance, which was the largest amount for any rater perspective, but was followed closely by peers. However, if perspective-related effects are considered specific aspects of the criterion space, and do not represent rater biases, performance-related variance increases to 38% for bosses and 31% for direct reports. There is no gain for peers because their perspective effect was zero. Boss ratings captured the greatest amount of performance-related variance while also being the least idiosyncratic, meaning boss ratings may be considered more valid than ratings from other rater sources. Specific dimensional performance components contributed little unique variance beyond what was accounted for by the general performance ratings. Random error effects accounted for 11% and 18% of rating variance in each sample, and the effects of random measurement error were similar across the different rater perspectives. Unfortunately, these results

show that who is conducting performance ratings accounts for more variance than what is actually being rated. The ratings in this study were a stronger reflection of rater biases than true performance.

Similar to the main finding in Scullen et al. (2000) that idiosyncratic rater effects accounted for a very large proportion of performance rating variance, a meta-analysis conducted by Viswesvaran, Ones, and Schmidt (1996) found that 29% of observed rating variance was method related. While these conclusions highlight a potential pitfall in performance appraisal systems, they also embrace the use of multirater performance feedback systems. Because of the large error component, averaging across multiple raters serves to significantly reduce the effects of bias and random error (Scullen, et al., 2000).

In part, multisource feedback systems have become popular among different organizations due to their apparent ability to capture unique performance perspectives from different raters. For example, individuals might behave differently with peers than with supervisors, thus peers might be better able to rate certain aspects of performance (Woehr, Sheehan, & Bennett, 2005). In addition, it is generally thought that feedback will increase the target's future performance by highlighting performance discrepancies. Improvements in performance might come from an awareness of self-other rating discrepancies that help a target focus on areas for improvement and increased attention (Atwater, Waldman, Atwater, & Cartier, 2000). Constructive feedback results in greater interest in subsequent feedback by motivating a leader to make changes after an initial round of feedback and seek further feedback after changes are made to see if the discrepancies remain (Waldman & Atwater, 1998). Self-regulation theory (Higgins,

1997) suggests that individuals are motivated to change behavior when they perceive discrepancies between their desired goals and their current performance. Striving to reduce these performance-goal discrepancies by enhancing existing knowledge or skills (promotional regulatory focus) is likely to produce positive change over time (Klein & Ziegert, 2004).

As stated by Avolio and colleagues (2003), after nearly six decades of leadership development research, the field is finally beginning to answer one of leadership's most important questions: can leadership be developed over time? However, before we can determine whether leadership can be developed, we must know if leadership performance changes over time and if these changes can be reliably measured. Klein and Ziegert (2004) caution that leader development is a hypothesis. A leader is thought to, over time, improve existing skills and knowledge, gain new skills and knowledge, and forget earlier skills and knowledge. However, not all leaders become more effective with time. Some leaders get worse and others do not change at all. Therefore, Klein and Ziegert (2004) define leader change as the slope of multiple dimensions of leader skills or knowledge over time, not necessarily that the leader must improve in effectiveness.

Dynamic Performance

Over time, a leader's knowledge, skills and abilities must develop if the leader is to become more effective. Performance feedback is often given and received annually in acknowledgment that the target's performance changes from year to year. Otherwise, a one-time performance review would be adequate. However, no research has examined the longitudinal effects of providing multisource feedback despite evidence that performance is dynamic, nor has research explored how changes in performance over

time are predictive of outcomes that are valuable to the leader and the organization such as promotion and performance level. Motowidlo's (2003) definition of performance as expected behavioral value to an organization over a standard period of time underlines the dynamic nature of performance. An individual's performance might vary over time due to changes in motivational factors and situational constraints or due to training and development efforts. In any case, the value of an individual's behavior to an organization does not depend on that individual's behavior at one point in time, but rather the expected outcomes of that behavior repeated over many occasions (Motowidlo, 2003).

Performance, including leadership performance, is best considered a dynamic construct (Thoreson, et al., 2004). Individual performance patterns have been investigated in several research studies, but little attention has been focused on describing or explaining the leadership growth process from this perspective (Day & Lance, 2004). Unfortunately, we know very little about how leadership changes individuals and how these individuals influence their organizations. We also know very little about the leadership of senior executives (Avolio, Sosik, Jung, & Berson, 2003) despite some research that suggests between 20% and 40% of an organization's effectiveness is due to executive leadership (Ireland & Hitt, 1999).

Leadership interventions using personality assessments and multisource feedback have been shown to result in decreases in performance in some studies, including a meta-analysis of the effects of performance feedback (Kluger & DeNisi, 1996). Feedback might result in an immediate decline in performance, followed by a subsequent increase in that performance once the target's new knowledge and skills have become engrained

(Klein & Ziegert, 2004). Evaluation efforts investigating the effects of feedback in leadership development; however, are almost nonexistent (Day & Lance, 2004; Klein & Ziegert, 2004).

Numerous studies have shown that rank-ordered performance changes systematically over time (Hofmann, Jacobs, & Baratta, 1993; Hofmann, Jacobs, & Gerras, 1992). Personnel selection research has had to grapple with the common problem of simplex patterns of correlations where predictor-criterion relationships are highest at Time 1 and steadily deteriorate as the time between the predictor and criterion increases (Schmitt, Cortina, Ingerick, & Wiechman, 2003). Because inferences derived from personnel selection are always longitudinal, researchers need to further investigate the role that time plays in selection models (Schmitt, et al., 2003). In addition, further research is needed to investigate the nature of dynamic criteria to differentiate between systematic change and random within-person performance change (Deadrick et al., 1997).

When job performance is considered a dynamic construct, evaluating performance trends and their resulting outcomes becomes essential. Several research attempts have been made to understand the reasons behind performance change, including moving through different job or skill acquisition stages, individual differences, and environmental factors.

In examining the effects of job stages, Murphy (1989) used a two-stage model of job performance that included transition and maintenance stages of performance. Transition stages occur when a new employee enters a job or when any of the major duties or responsibilities of a job change. Even when the job title and description remain constant, changes in job demands or the job environment can trigger further transition stages.

Maintenance stages occur when employees are no longer confronted with novel job demands such that their job tasks are well understood and learned (Deadrick, et al., 1997). Performance during transition stages should be predictable by cognitive ability while performance during maintenance stages should be predictable by dispositional characteristics. Individual performance trends will depend on individual differences in performance stage as well as cognitive ability and dispositional characteristics (Murphy, 1989).

Kanfer and Ackerman (1989) developed a theory very similar to Murphy's (1989) such that individuals are expected to have different rates of performance during early and later job stages. However, Kanfer and Ackerman attribute these performance changes to rates of skill acquisition rather than changes in job demands or the job environment. These researchers argued that performance changes during different stages of skill acquisition: declarative knowledge, knowledge compilation, and procedural knowledge. Their model suggests that individual performance trends are determined by an employee's stage of skill acquisition on the job as well as by individual differences in ability and motivation. During the declarative knowledge stage, employees are still learning the job, thus, performance is slow and an employee might commit more frequent errors. Cognitive ability and motivation are important for better performance at this stage. Later, in the procedural knowledge stage, employees have become accustomed to their job tasks and the organization's rules and procedures such that fast and accurate job performance requires less attention, cognitive ability and motivation are less important, and job skill becomes a primary determinant of job performance.

More recent research conducted by Ployhart and Hakel (1998) found support for a classic learning curve in a study investigating insurance sales performance over eight consecutive business quarters in a sample of 303 securities analysts. Results using linear, quadratic, and cubic models of performance showed sharp performance increases in the initial quarters followed by decelerated increases in performance in subsequent quarters. Similarly, Thoreson and colleagues (2004) investigated the relationship between personality and sales performance growth in early and late job stages in a sample of 137 pharmaceutical sales representatives. Results indicated that agreeableness and openness to experience predicted performance growth in early job stages while conscientiousness predicted performance growth in later job stages even while the effects of job tenure were statistically controlled. The relationship between job tenure and performance has been shown to be initially positive and to then plateau (Avolio, Waldman, & McDaniel, 1990; Schmidt, Hunter & Outerbridge, 1986). The results of these studies suggest that changes in performance occur more dramatically in early job stages and then decelerate in later job stages.

Several research efforts have tested the argument that performance changes over time are due to individual differences in performance stages and individual differences in abilities and dispositions. A group of researchers in the early 1990s examined individual growth curves in major-league baseball players and life insurance salespeople to increase their understanding of dynamic performance (Hofmann, Jacobs, & Baratta, 1993; Hofmann, Jacobs, & Gerras, 1992). Results from these studies indicated that intra-individual performance changes were systematic and that subgroups of participants with different change patterns could be identified. In a follow-up to the Hofmann et al. studies

(1992; 1993), Deadrick and colleagues (1997) used hierarchical linear modeling (HLM) to investigate the effects of general mental ability and job tenure to forecast performance change. These researchers found that job experience and psychomotor ability were positively related to initial levels of job performance in a sample of 408 sewing machine operators. General mental ability was positively correlated with performance increases while previous job experience was negatively correlated with performance increases. The authors concluded that the abilities that determine an individual's performance over time change. Despite the findings that individual differences in ability and experience were significantly correlated with performance trends, these factors only accounted for 5% of the variance in the rate of performance change. Unmeasured variables that moderate performance change need to be further investigated (Deadrick, et al., 1997). This research study will investigate rater and contextual variables as possible moderators of individual leadership performance development.

Klein and Ziegert (2004) propose that individual differences and organizational climate have indirect and moderating effects on leader change. An indirect effect of personal characteristics such as self-monitoring, learning orientation, self-efficacy, and proactive behavior might influence whether a leader seeks challenging work assignments and receives feedback and training. Individual differences might also have a moderating effect on leader change such that they condition a leader's response to challenging work assignments, feedback, and training. Organizational climate might have indirect effects on leader change by influencing how leaders experience work challenges, feedback, and training while also having a moderating effect by influencing leader acquisition of new skills and knowledge as a result of work challenges, feedback, and training.

An organization with a strong climate for leader development should be expected to reward leader change. However, organizations present a continuum of rewards for leader development. Organizations might reward and promote leaders due to seniority, political connections, or factors that have little relationship with leader development. Organizations might also reward and promote leaders due to the financial performance of their units. Finally, an organization might truly reward and promote leaders due to their participation in leadership development programs and their positive change in leadership skills and knowledge (Klien & Ziegert, 2004). Following Klein and Ziegert's (2004) suggestion to use leadership measures derived from broad and basic leadership typologies, this study used a five-factor model of leadership performance that closely coincides with well-researched leadership theories. This approach is advantageous such that it allows the comparison of a diverse group of leaders' performances over time. Unfortunately, this approach is not highly sensitive to any context-specific changes that may occur.

Past research has investigated the performance changing effects of job stages, skill acquisition stages, individual differences, and environmental factors on a variety of populations. These studies have found that changes in performance over time might be due to changes in job demands, the job environment, rate of skill acquisition, challenge and feedback seeking behaviors, and a leadership development climate. However, these studies did not focus on leader development. The term "leader development" suggests that leaders improve over time; however, this remains a largely untested assertion (Klein & Ziegert, 2004). One suggested reason for the lack of leadership development research is the degree of difficulty it takes to model and measure such development. Day and

Lance (2004) discussed nine issues concerning the conceptualization and measurement of leadership development. These nine issues are briefly discussed here.

The first issue discussed by Day and Lance (2004) concerns random versus systematic change. Developing models and measures of leadership performance is hampered because leadership performance might fluctuate randomly, or might be so prone to situational influences that no systematic changes are possible to detect. On the other hand, leadership performance might follow an organized, systematic developmental trajectory. Day and Lance note that change in leadership performance must be modeled on the basis of true leader performance scores or estimates rather than on measures containing substantial measurement error. Unfortunately, change in leadership performance is often operationalized using observed measures that contain measurement error and response biases in addition to true score components. Use of unreliable or biased measures of leadership performance might consequently hide true change.

Second, measuring lasting change in leadership performance is difficult because the change might be reversible or irreversible. The need for refresher training is an acknowledgement that leaders might revert or relapse to stages prior to training occurrence. Third, changes in leadership performance may be unitary or multipath. The measurement of change must allow for the identification of subgroups or clusters of individuals who share similar patterns of change. Fourth is the matter of continuous versus discontinuous change. Most changes on psychological characteristics are assumed to be continuous; however, measurement models should be able to detect discontinuous change where appropriate.

Fifth, change may be conceptualized as quantitative or qualitative. Three types of change can lead to differences in observed scores over time (Golembiewski, Billingsley, & Yeager, 1976). Alpha change is a true change in an underlying construct and requires scale measurement non-invariance to interpret. Beta change is a respondent's reconceptualization of the measurement scale (i.e., the lengthening or shortening of intervals between scale points) from one measurement interval to another despite maintaining a constant conceptualization of the construct. Beta change may occur due to experiences between the two time periods and results in a respondent choosing different item response categories from one measurement interval to the next. Gamma change is a fundamental change and qualitative redefinition in how a respondent understands and defines a construct over time such as when an employee perceives that an organization is more supportive after it provides job-related training (Mullen, Kroustalis, Meade, & Surface, 2006). Tests for beta and gamma change must be conducted prior to assessing longitudinal alpha change (e.g., true change in leadership performance) in order to allow accurate comparisons between time intervals (Day & Lance, 2004).

Sixth, change should be assessed at the individual and group level of analysis because leadership concepts may be more theoretically meaningful at either the individual or group level. It is important, therefore, that the measurement model should allow for either possibility. Seventh is the idea that there are individual differences in leadership performance change. Individuals may differ in both their initial status and their rate of change, and exogenous variables may have effects on differential growth patterns. Eighth, concomitant, or tandem, changes in constructs should be part of a strategy for

change assessment. Finally, models assessing change should investigate differential change across groups.

In sum, Day and Lance (2004) note that the vast majority of studies assessing longitudinal change use measurement techniques (i.e., descriptive statistics, change scores, *t* tests, ANOVA, lagged regression, and MANOVA) that do not address the nine criteria for an effective approach to longitudinal measurement change described above. Second-order factor latent growth modeling, however, fulfills all of these criteria (Day & Lance, 2004). For this reason, this research study will use SOF LGM when assessing longitudinal leadership performance ratings. SOF LGM and how it addresses Day and Lance's (2004) nine criteria for the effective measurement of change will be explained in detail in a following section. In addition, the use of consistent raters over rating time periods has not been studied. This study investigated rater perspective and composition (i.e., important idiosyncratic effects) using multisource longitudinal performance ratings and SOF LGM.

Rating Source Agreement

Proponents of multisource feedback systems believe that additional information about a target's performance can be gained by using raters from different organizational perspectives. Job performance ratings are a function of both the rating source and the dimensions of performance being measured (Woehr, Sheehan, & Bennett, 2005). Multisource agreement in performance-rating systems has been well researched, and most of this research indicates that individuals from different organizational perspectives view the performance of the target differently (Woehr, Sheehan, & Bennett, 2005). In a meta-analysis conducted by Harris and Schaubroeck (1988), self-ratings of job

performance moderately correlated with supervisor ratings (r = .35) and peer ratings (r = .36), but peer and supervisor ratings of the target's performance were correlated much higher (r = .62). Several researchers have discussed reasons for rating source disagreement.

Waldman and Atwater (1998) present 8 factors that contribute to discrepancies between rater groups. First, they note that individuals being rated tend to view themselves positively, which might lead them to ignore, distort, rationalize, or attach less importance to negative information they might receive. Second, raters' reluctance to give negative feedback perpetuates the target's belief that he or she is performing better than expected. Third, different rater sources also have unique perspectives on a target's behavior. Fourth, if raters believe that providing ratings will affect important job-related outcomes, raters might distort ratings toward the more positive or negative end of the rating scale depending on their motivation. Fifth, raters might hold stereotypes or biases about the target or their performance. Implicit theories about leaders and their performance might contribute to bias in their ratings. Sixth, raters might experience different emotions while rating, such as fear of retaliation from the target. Seventh, attitudes, such as liking the target, might influence ratings. Finally, cognitive processing variables, such as selective attention, primacy, and recency can affect information recalled and, thus, rated. (Waldman & Atwater, 1998).

Further research exists and has found, for instance, that there are often higher levels of agreement between different supervisors than between different peers rating the same target (Viswesvaran, Ones, & Schmidt, 1996). Several studies reported in Borman (1997) show that agreement within rating sources tends to be greater than across rating

sources. In a somewhat surprising finding, Salam, Cox, and Sims (1997) found that leaders who empowered direct reports were rated positively by direct reports but negatively by supervisors. Interestingly, performance ratings from direct reports had the highest correlations to actual leader performance among all rater groups. Despite the lack of consensus on these findings, organizations continue to use multisource feedback systems (Woehr, Sheehan, & Bennett, 2005).

Advocates of multisource feedback systems generally agree that cross-source rating differences represent true differences in perspective and opportunities to observe the target's behavior (Woehr, Sheehan, & Bennett, 2005). According to Lance and Woehr's (1989) "ecological perspective," source effects may actually represent valid, systematic sources of performance information. Multisource feedback systems take several organizational perspectives into account. Supervisors rarely witness a leader engaging in day-to-day behaviors, which comprise a large proportion of that leader's overall performance. Direct reports are able to see many more of these day-to-day behaviors, yet their personal feelings might create an idiosyncratic bias in their specific performance ratings. Finally, while a leader may have the clearest view of his or her behaviors and characteristics, there may be a self-serving bias in his or her self-ratings (Salam, Cox, & Sims, 1997).

Borman (1997) sums up the arguments for true rater differences by presenting three reasons to expect perspective-related biases in multisource performance ratings. First, raters in different organizational positions might focus their attention on different aspects of the target's performance. Second, raters might attach different weights to these performance aspects. Finally, raters from different perspectives might have varying

opportunities to observe the target's performance. In such cases, perspective-related effects are not biases but reflect true performance. In order to effectively use performance information for organizational purposes, differences among raters should be correctly identified and understood (Cheung, 1999).

This study includes the rater characteristics of organizational perspective, composition, organizational sector, and organizational subdivision. If these rater characteristics influence judgments of leadership performance, the rater evaluations might indicate more about the rater than the leader. Leader performance typically plays a dominant role in leadership performance ratings, but research has found that rater characteristics also have effects on performance ratings (Cardy & Dobbins, 1994). Examining rater characteristics such as those included in this study may help an organization identify groups of raters with similar characteristics that provide more accurate and predictive leadership performance ratings.

Rating Source Measurement Invariance

A string of recent research has begun to examine the measurement invariance of performance ratings across multiple rater sources through the use of confirmatory factor analysis (CFA), the most common way to evaluate measurement invariance (Vandenberg & Lance, 2000). In the present study, rating source noninvariance is an important hypothesis to test because it might obscure changes in a target's longitudinal performance when data from all rating sources are combined. Rating source inequivalence also makes comparisons across ratings sources inaccurate (Bollen, 1989).

In the past, measurement quality was rooted in the partitioning of observed scores into true and error score components. Classical test theory provides a solid foundation

for the study of manifest variables and measurement properties such as reliability and validity. However, additional questions extend beyond the traditional use of classical test theory and the study of manifest variables (Vandenberg & Lance, 2000). These questions include concerns about measurement invariance, or alternately, measurement equivalence. Such questions can include whether respondents from different cultures interpret a measure in the same conceptual manner, whether different rating sources define and then rate performance in a similar manner, and whether there are gender, ethnic, or other individual characteristics that affect how respondents answer surveys (Vandenberg & Lance, 2000).

CFA allows researchers to answer these questions. CFA allows tests of hypotheses relating to measurement equivalence such as 1) that all items on a scale adhere to the same conceptual framework in defining the construct in each comparison group, 2) the regression slopes and intercepts from the manifest variables to the latent construct are invariant across groups, 3) the unique variances are invariant across groups, 4) variances and covariances among the latent variables are invariant across groups, and 5) the CFA model is equivalent and holds a common structure across groups (Vandenberg & Lance, 2000). All of these hypotheses are testable within a CFA framework.

Several studies have used a CFA approach to test the measurement equivalence of performance ratings provided by different rater groups. Cheung (1999) examined the equivalence of self- and supervisor ratings of a sample of mid-level managers by using a CFA approach. Maurer, Raju, and Collins (1998) used CFA to examine the measurement equivalence of peer and direct report ratings of team building skills for a group of managers. Similarly, Facteau and Craig (2001) also used CFA to examine self-,

supervisor, direct report, and peer performance ratings across seven performance dimensions. In all studies, results indicated that ratings from each source were equivalent. In sum, these studies provide important evidence for the measurement equivalence of different rating sources.

Utilizing a multitrait-multirater approach, Woehr and colleagues (2005) tested cross-source measurement equivalence in a sample of 1,028 Air Force airmen. Similar to previous research, results indicated low levels of cross-source agreement; however, ratings reflected similar underlying performance dimensions across rating sources. They concluded that it is, therefore, possible to make meaningful comparisons across rating sources. These researchers found that 34% of the variance in performance ratings was accounted for by the performance dimension component, and 21% was accounted for by the rating source component. Approximately 45% of the rating variance across rating source and performance dimensions was accounted for by a uniqueness component. These researchers concluded that rating source effects are substantial but that the largest influence on ratings are due to parameters not attributable to performance or rating source. Apparently, lack of agreement across rating sources is more likely due to unmeasured effects associated with the performance construct being rated than from rating source-specific effects.

Woehr and colleagues suggest that future research further investigate the unique effects associated with the performance construct by utilizing CFA techniques to examine the relationship of the unique effects with other criterion or predictor measures. If the unique effects reflect systematic performance-related variance, such as that due to

perspective or opportunity to observe the target, they should correlate with such criterion and predictor measures.

Latent Growth Modeling

LGM is a two-stage process that places a confirmatory factor analytic structure on variables that are measured longitudinally. In general, LGM is concerned with longitudinal change. It is used when modeling an individual's change or development on a given latent variable over three or more time intervals. LGM may also be augmented to allow for tests of measurement invariance, the prediction of outcome variables, and the influences of exogenous variables. Such an augmentation would take the form of second-order factor latent growth modeling (SOF LGM), which is the most appropriate way of analyzing measurement equivalence and dynamic performance models with exogenous and outcome variables (Day & Lance, 2004). In this study, SOF LGM was used to evaluate the development of leadership performance in a group of executives over a three-year time period. The use of LGM techniques allows researchers to ask and evaluate important questions in longitudinal data, such as this study's question regarding the ability of longitudinal leadership performance ratings to predict future promotion and job performance and whether or not ratings obtained from different rating sources affect this prediction.

Longitudinal data are multilevel and nested, and researchers need to treat them as such (Deadrick et al., 1997). The development of LGM occurred almost 50 years ago, yet it received sparse attention until the 1980s and is now only beginning to appear in the applied psychology literature (Day & Lance, 2004). The measurement of longitudinal change has seen its fair amount of controversy. There is little consensus on the best

method to assess longitudinal change in theory and in practice (Lance, Vandenberg, & Self, 2000). However, advances in software and computing availability have allowed LGM to gain acceptance as a method for assessing longitudinal change, and has been successfully used to study a number of variables in industrial/organizational psychology.

Lance and colleagues (2000) successfully modeled longitudinal change in work adjustment among a sample of 104 new hires at a banking institution. Using SOF LGM, these researchers directly addressed newcomers' commitment as a key component of employee adjustment to the work environment. Results indicated that an attachment dimension called "Internalization" declined over time, that it was predictable from antecedent variables, and that it predicted turnover intentions. This study addresses similar issues in the evaluation of longitudinal leadership performance, the effects of exogenous rater variables, and the prediction of promotion and performance outcomes.

To interpret change on a given variable requires measurement equivalence, or measurement invariance, across measurement occasions (Chan, 1998). That is, the nature of the construct and how respondents interpret it must remain constant over time. Unfortunately, this assumption is rarely tested and is, perhaps, often violated thus producing misleading interpretations. Measurement non-invariance, or measurement inequivalence, indicates that the observed measures of a construct actually represent assessments of different constructs at different points in time, thus, prohibiting the use of LGM.

LGM is a two-stage process used to evaluate variables that are measured longitudinally. In the first stage, individual-level growth trajectories are fit to measures of the same construct over 3 or more measurement occasions. For example, Figure 1

shows individual growth trajectories for a group of hypothetical leaders. Each individual has a unique initial status, or starting point, on the leadership performance measure at Time 1 (T1) as well as a unique rate of change on the measurement over the four time periods (indicated by line slopes). This example shows that there is a substantial amount of variability in individual leaders' initial status. While most leaders exhibit a positive rate of change, there is also substantial variability in their slopes.

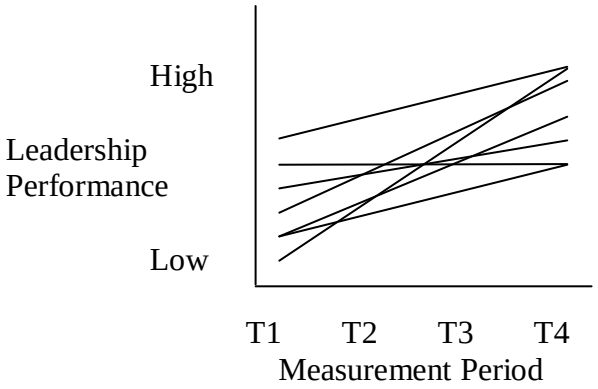

Figure 1. Plot of linear growth trajectories for a group of seven hypothetical leaders' development in leadership performance.

Individual-level growth trajectories, estimated mean initial status and change for the sample, and estimated variance of the initial status and change parameters for the sample are the basic units of analysis in LGM (Day & Lance, 2004). In other words, the first stage of LGM models aspects of intraindividual change, including changes in the trajectory of sample means and within-sample variability. The second stage models between-individual differences. In this stage, additional variables are introduced as predictors of individual differences in the growth trajectories.

LGM satisfies three of Day and Lance's (2004) criteria for longitudinal measurement. First, LGM controls for random measurement error because change is modeled at the level of latent variables, which are, in theory, perfectly reliable (Criterion 1). Second,

individual differences in change can be estimated as the variance on the initial status and change variables (Criterion 7). Finally, change may be assessed at the individual and group levels of analysis because mean growth parameters can be estimated for the sample (Criterion 6).

Extending LGM to SOF LGM satisfies Day and Lance's (2004) remaining criteria for longitudinal change measurement. Reversible or irreversible change (Criterion 2) may be modeled because some factor loadings for the latent variable of change are freely estimated parameters, allowing one to identify situations in which individuals return to their initial status after some growth. Latent class analysis and growth mixture modeling (GMM) with SOF LGM allow the investigation of unitary and multipath change (Criterion 3) by identifying subgroups of individuals who share similar change patterns. SOF LGM satisfies the modeling of both continuous and discontinuous change (Criterion 4) by allowing the specification of higher order curves through the addition of another latent variable of change and the appropriate factor loading coefficients for modeling the desired function (e.g., quadratic or cubic functions). Qualitative and quantitative change (Criterion 5) may be examined by including multiple manifest indicators for each measurement interval. Using tests for measurement equivalence, SOF LGM examines whether the relationships between observed measures and their underlying constructs are constant over time or whether beta or gamma change has occurred. Finally, SOF LGM allows the investigation of concomitant change (Criterion 8) by allowing the examination of change in multiple domains simultaneously, and allows the analysis of between-group differences in change (Criterion 9) by allowing the simultaneous estimation of multiple groups.

One extension of LGM to SOF LGM is to include multiple manifest indicators for each measurement interval. These indicators may be different measures, individual items from a single measure, or subsections from a single measure. In this study, composite scores on the five leadership performance factors from the leadership performance instrument serve as the manifest indicators for LGM analyses and items from the factors serve as the manifest indicators for SOF LGM analyses. Using multiple indicators accomplishes three important goals. First, it controls for random measurement error by operationalizing the measurement of the focal variable (i.e., leadership performance) at each time period as a latent variable. Second, it separates nonsystematic measurement error from systematic time-specific effects. Finally, it allows for tests of longitudinal measurement invariance (Day & Lance, 2004). The SOF LGM model used in this study is shown in Figures 2a -2c.

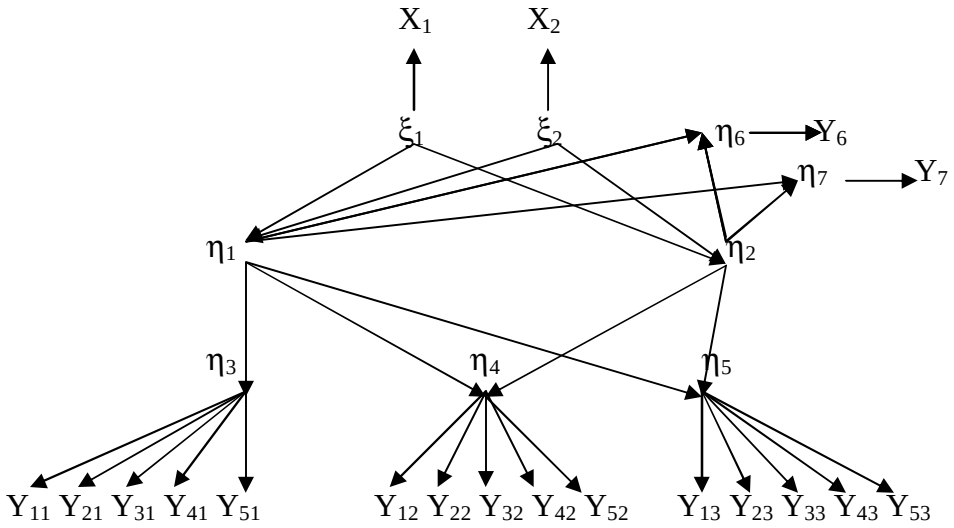

Figure 2a. Structural model for individual and combined rater groups including context predictors and outcomes of true initial status and change on each leadership performance factor. *Note:* Occasion-specific variances are omitted for visual ease.

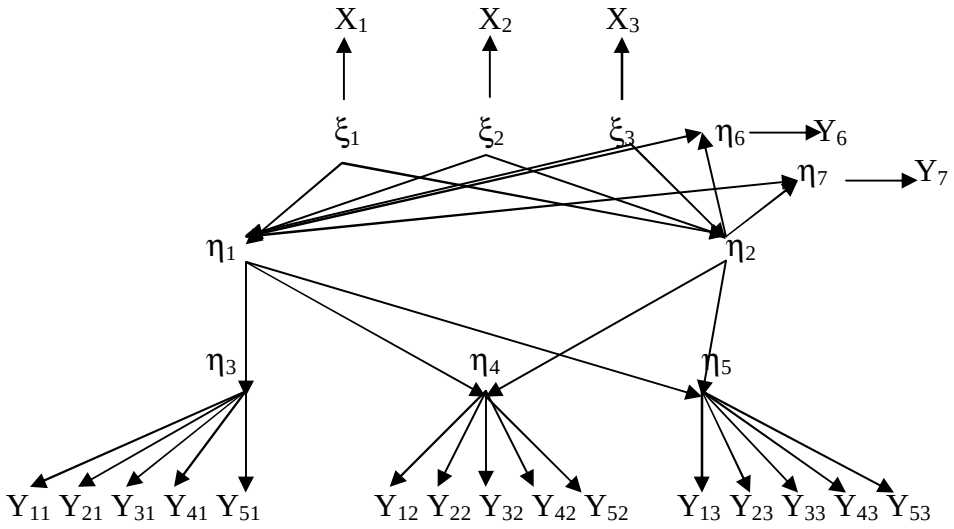

Figure 2b. Structural model for boss ratings including context and composition predictors and outcomes of true initial status and change on each leadership performance factor. *Note:* Occasion-specific variances are omitted for visual ease.

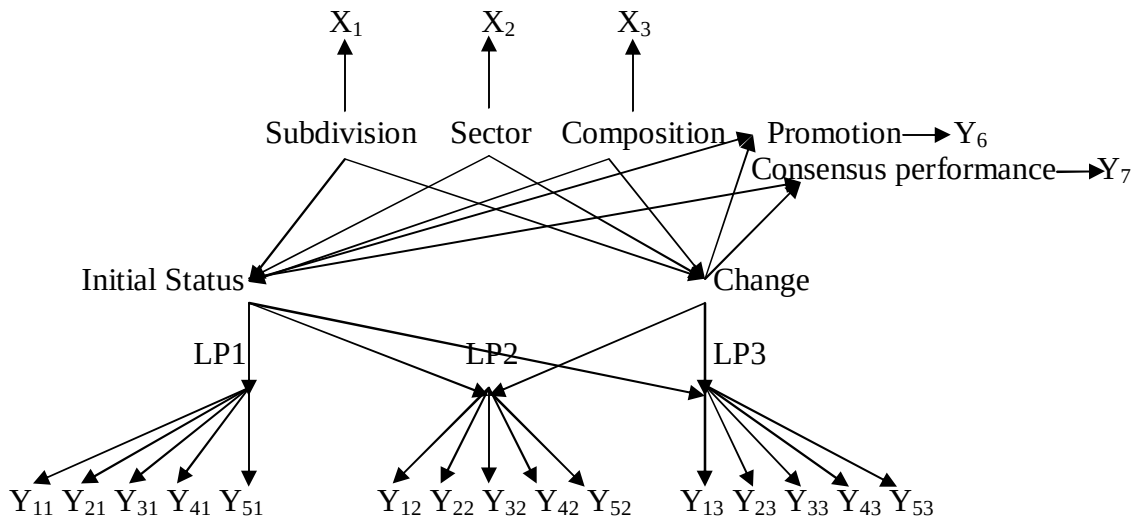

Figure 2c. Structural model for boss ratings including predictors and outcomes of true initial status and change on each leadership performance factor (LP). *Note:* Occasion-specific variances are omitted for visual ease.

In the figures above, each of *N* individuals is observed on the focal construct, η_i, at 3 measurement occasions. Multiple manifest variables, Y_i, are used to measure the focal construct at each occasion. Here, the multiple indicators consist of 5 composite survey scores used to operationalize the focal construct of leadership performance for LGM analyses and individiual item scores for each leadership performance factor for SOF LGM analyses. The latent variable η_1 represents individual executives' true initial status on η_i at Time 1. The latent variable η_2 represents individual executives' true change on leadership performance over time. The ξ_i are exogenous variables that serve to predict individual differences in latent initial status, η_1, and change, η_2. The hypothesized relationship between the ξ_i and η_1 represent a static relationship similar to those found often in cross-sectional studies. However, the hypothesized relationship between the ξ_i and η_2 represents a moderator effect because η_2 represents executives' change in the focal construct, leadership performance. Therefore, as shown in the figures above, the relationships between values of leadership performance and time in boss ratings of leadership performance are contingent upon the exogenous rater composition and context variables. Finally, the latent variables η_6 and η_7 are hypothesized outcomes of initial status and change. This allows for the possibility that individual differences in true leadership performance change, and that individual differences in initial status can predict important outcome variables. This model tests for mediational hypotheses in that the effects of the antecedent (i.e., subdivision, sector, and rater composition) and outcome variables (i.e., promotion, consensus performance) are partially or completely mediated by initial status and/or change in leadership performance.

Research Questions

The research questions addressed by this study include: 1) Is there measurement equivalence between rating source groups on the leadership performance construct? 2) Is there measurement equivalence across measurement occasions on the leadership performance construct? 3) What form does the latent growth curve take for leadership performance over a multi-year period? 4) Are longitudinal leadership performance trajectories similar for raters in different organizational perspectives (i.e., boss, self, direct report)? 5) Is the rate of leadership performance development contingent upon rater composition (i.e., consistent or inconsistent raters at each time interval)? 6) Are the effects of the exogenous variables (i.e., rater composition, rater sector, and rater subdivision) partially or completely mediated by initial status and/or change in leadership performance? 7) Does change in leadership performance predict objective performance outcomes, such as promotion and consensus performance level?

The answers to these questions will have significant implications for executive succession planning and performance appraisal and management systems. Examining leadership performance growth will allow management to estimate how the rate of leadership development can have implications for executive succession planning, including how long it will take a new hire to reach a departing executive's level of leadership performance. This issue becomes especially important in light of the impending retirement of the baby boomer generation. In addition, examining leadership performance growth and its relationship to objective measures of job performance allows for validation of the performance appraisal system. Exploring possible rater moderators will inform management teams of the most appropriate use of raters in terms of rater

organizational perspective and the importance of using consistent groups of raters at multiple time intervals.

CHAPTER THREE: METHOD

Participants

The data in this study were taken from a large *Fortune* 100 global communications company recently involved in developing a leadership performance instrument. The sample consisted of over one thousand senior executives and senior managers working in 49 countries within 4 regions of the world during the years 2001-2003. Leaders worked within 11 different organizational sectors and 108 different organizational subdivisions. Organizational sectors are functional business units, such as Personal Communications or Semiconductor Products. Organizational subdivisions are distinct business units within the sectors, such as Semiconductor Products Marketing and Global Supply. Bosses and direct reports provided ratings on the leadership performance instrument for each target executive.

All ratings obtained from an unknown source were eliminated. Cases with greater than 25% (7 items) missing data were also eliminated. Mean replacement for each year and source group was used to replace the remaining missing values. Due to the variable number of direct reports providing ratings at each year, direct report ratings were compared and then averaged within each target and year as explained in greater detail later. For measurement equivalence testing, data were retained for 588 targets who had ratings from at least one rater from each rating source group for each of the three years. For longitudinal growth modeling, further data cleaning resulted in 331 targets with a self-rating, a boss rating, and an averaged direct report rating for each year of measurement. The final sample represented 11 organizational sectors and 108 subdivisions. No demographic data were available.

Rating Instrument

Measurements of leadership performance were obtained on three occasions. Leadership performance was assessed using a multisource instrument, the Leadership Standards Assessment (LSA). The instrument was developed based on the company's leadership model, which measures 5 key areas of leadership, labeled Envision, Energize, Edge, Execute, and Ethics and Character. Ratings were assessed using a single 5-point rating scale (0 = *ineffective*; 4 = *exceptionally effective*). Confirmatory factor analyses (CFA) indicated that the instrument fits the 5-factor leadership model well (CFI = .966, GFI = .951, TLI = .955, RMSR = .027, RMSEA = .066). Internal consistency reliability (coefficient alpha) for each of the five scales is above .86, and the overall reliability of the instrument is equal to .884 (Kaiser, Craig, & Kaplan, 2002). Brief descriptions of each of the leadership performance factors are presented in Table 1.

Table 1

Leadership Performance Factors

Factor	Number of Items	Description
Envision	7	Being innovative. Having the strategy and vision needed to accomplish goals.
Energize	7	Motivating others for change and transition.
Edge	4	Making difficult decisions quickly with incomplete or imperfect data.
Execute	5	Achieving results and meeting commitments.
Ethics and Character	2	Honesty. Placing the goals of the organization first.

The factors Energize, Edge, and Execute were first defined in Cohen and Tichy's (1997) work with *Fortune* 500 companies such as General Electric. Edge is defined as

making tough decisions in a timely manner with incomplete or imperfect data. Energize is defined as motivating others regarding change and transition. Execute is defined as achieving results and meeting commitments. Envision involves innovation and change as well as the strategy and vision needed to accomplish goals. Finally, the Ethics and Character factor involves the ethical conduct of business and placing the organization's goals ahead of personal ambitions.

The leadership areas measured by the LSA reflect many current leadership theories. Edge, Execute, and Energize describe the behaviors of a leader. These facets are concerned with how a leader motivates followers, or his or her style of leadership. Researchers have found that leader behaviors cluster into two types: initiating structure and consideration (Stogdill, 1974). Initiating structure includes behaviors such as organizing work, giving structure to work context, defining role responsibilities, and scheduling work activities. Consideration includes behaviors such as building camaraderie, respect, trust, and liking between leaders and followers (Northouse, 2004). Edge and Execute may be considered facets of task orientation, or initiating structure, and Energize may be considered a facet of relationship orientation, or consideration.

Envision is a facet of leadership that captures the actions of a leader, and is concerned with what the leader does to accomplish a goal. Envision is similar to work conducted by Marshall-Mies and colleagues on cognitive and metacognitive constructs relevant to executive leadership. These researchers believe that executive leadership involves complex social problem solving. Leaders are expected to identify key issues that are important for organizational goal attainment and then generate solutions or plans that

address these issues (Marshall-Mies, Fleishman, Martin, Zaccaro, Baughman, & McGee, 2000).

Ethics and Character is a measure of both leader integrity and emotional stability. Integrity and trust have been shown to be very important characteristics in highly effective and transformational leaders (Bass, 1985). Such characteristics in leaders have been shown to improve direct report performance and satisfaction (Podsakoff, Mackenzie, Moorman, & Fetter, 1990). The big five factor of neuroticism, or emotional stability, measures such emotions as fear, guilt, anxiety, depression, embarrassment, insecurity, and frustration. Individuals scoring high in neuroticism tend to have irrational ideas, difficulty controlling their impulses, and trouble coping with stress. Individuals scoring low on neuroticism are calm, relaxed, even-tempered, and face stress easily (Costa & McCrae, 1992). Researchers have recommended including measures of ethical behavior in performance appraisals so that ethical behavior becomes relevant and reinforceable (Buckley, 2001; Weaver, 2001).

Outcome Variables

Promotion. The promotion variable was calculated for each target by subtracting his or her self-reported organizational level (E grade) in 2001 from their self-reported E grade in 2003. E grades in this dataset range from E13 to E17, with E17 being the highest organizational level possible. Executives fill grades E15, E16, and E17; senior managers fill grades E13 and E14.

Consensus Performance. Consensus performance refers to the meeting between a leader and his or her supervisor where they mutually agree on the leader's final administrative performance rating. During that meeting, the supervisor takes into

account the LSA ratings along with other performance data and arrives at a final rating (consensus performance score) that affects tangible outcomes such as bonuses and raises. Consensus performance scores are correlated with LSA ratings because the LSA is one input into the consensus performance rating; however, the correlation is not 1.0 because other factors are also taken into account (e.g., objective unit financial performance). There are three possible consensus performance scores. A score of 1 signifies that the target is among the least effective managers in the company. A score of 2 signifies that the target is solidly effective, and a score of 3 signifies that the target is among the most effective managers in the company.

CHAPTER FOUR: RESULTS

Previous analyses

Previous analyses on similar data collected between the years 2000 and 2002 used CFA and item response theory (IRT) to evaluate measurement equivalence (Kaiser, Craig, & Kaplan, 2002). CFA analyses comparing the 2000 data to the 2002 data showed that the same model fit about equally well at both time points for all rating sources combined. The items and scales on the LSA also showed no differential functioning from the 2000 data to the 2002 data. Data from 2001 were not entered into the analyses following Craig and Kaiser's (2003) targeted methodology of first testing for measurement inequivalence in the most likely place (i.e., a comparison of 2000 data to 2002 data). Tests of measurement equivalence were also conducted across rater cultural groups (United States, China/Hong Kong/Taiwan, U.K./Ireland, Singapore/ Malaysia, and Israel). The data for each group fit the leadership performance model adequately.

A one-way analysis of variance (ANOVA) with year of administration as the grouping variable was used to assess trends in the data over time. ANOVA suggested that several potential trends were evident in the ratings data. Boss ratings showed an upward trend in observed means from 2000 to 2002 while direct report ratings showed a downward trend from 2001 to 2002 (ratings from direct reports were not collected in 2000). Self-ratings increased from 2000 to 2001, but decreased from 2001 to 2002. In all, the greatest gains were made in the same year as the company's increased attention to leadership, 2000 to 2001. This might have improved leadership motivation, developmental innovations or administrative interventions for that year (Kaiser, Craig, &

Kaplan, 2002). To avoid this honeymoon period in ratings, the current study begins with ratings obtained in 2001.

These previous analyses used ANOVA to evaluate trends in the data. ANOVA is a comparison of variances with an underlying concern with mean differences, not variability. Differences in cross-group means create between-group variability. In ANOVA, within-group differences are considered to be error variance, and when between-group differences exceed that of within-group differences, the result is that the groups are said to be significantly different. In LGM analyses, the focus is on the systemic relationship between variables as indicated by their covariance. Means are inconsequential and are often set equal to zero. For this reason, LGM analyses become the preferred method of analyzing longitudinal data for growth trends over time at the latent level (Muthen, 2002). Kaiser and colleagues (2002) suggested that future research examine the effects of rater perspectives and rater composition to account for alternative explanations for the performance trends over time.

The LGM analyses performed in this study proceeded in four phases: 1) tests of measurement invariance between rating source groups, 2) longitudinal tests of measurement invariance for each rater group, 3) development and evaluation of the LGM model of leadership performance for each rating source group and for all rating source groups combined, and 4) augmentation to the SOF LGM measurement models with moderator and outcome variables. All analyses were conducted with maximum likelihood estimates using Mplus software (Muthen & Muthen, 2000).

Tests of Measurement Invariance

The first two research questions address the existence of measurement invariance between rating source groups and measurement occasions on the leadership performance construct. Comparisons such as these require equality of the measurement model and of the factor covariance matrix. Tests for measurement invariance involve generating estimates for the model presented in Figure 1, and then with subsequent model runs, placing increasingly stringent constraints on the data and comparing model fit with each additional level of constraint. A series of CFA model comparisons were conducted to confirm measurement equivalence among rating source groups and measurement occasions. First, a comparison of rating source groups was made at each measurement occasion. Second, a comparison was made for each rater group across measurement occasions. Third, a comparison of all ratings combined was conducted simultaneously with a SOF LGM analysis in Mplus.

The first model (Model 1) in invariance testing holds the pattern of fixed and free factor loadings to be equal. Model 1 fit a five-factor leadership performance model to each of the rater groups and each of the measurement occasions. In other words, nine CFA models were tested; one for each of three rater groups at each of three measurement occasions. An acceptable fit to Model 1 indicates that the five-factor leadership performance model generalized across rating source groups and across measurement occasions. An inadequate fit to Model 1 would indicate some fundamental difference in the underlying leadership performance dimensions across rater groups or measurement occasions; thus, testing for measurement equivalence would halt.

If an adequate fit of Model 1 is found, the sequential testing for equivalence continues to a more restrictive level in Model 2. Model 2 allows the measurement models to vary, but requires that the factor covariance matrices be equal across groups or measurement occasions. An adequate fit to Model 2 indicates that the correlations among the five factors do not differ significantly by rating source group at all three measurement occasions.

Finally, Model 3 is identical to Model 2 except that the same items' factor loadings are constrained as invariant across rating source groups or measurement occasions. Model 3 involves freeing the factor correlations to be estimated separately by rating source group or measurement occasion, while constraining the factor loadings to be equal across groups. If the fit of Model 3 is acceptable, the slopes of the items' regressions on leadership performance factors can be said to be equivalent. A fourth model testing for error variance equivalence could be tested; however, such equivalence is not necessary for meaningful cross-source comparisons, but would rather suggest that the ratings contain the same amount of error variance across sources (Woehr, Sheehan, & Bennett, 2005).

The tests of these models help to ascertain whether the leadership performance measures are conceptually equivalent across measurement occasions and across rater groups. This equivalence is necessary in order to compare the measurement results over time and across raters. If equivalence across measurement occasions is not supported, no longitudinal comparisons are warranted because the underlying performance dimensions represent different constructs at each time point. Evidence of measurement equivalence

across cultural groups was found in Kaiser, Craig, and Kaplan (2002); thus, those tests were not performed in this study.

Rating Source Measurement Invariance

Model 1 of the test for rating source measurement invariance fit the hypothesized five-factor leadership performance model to each of the three rater groups. For each group, there were 588 observations, 27 items, and 5 hypothesized factors. A summary of the fit statistics obtained for all rating source groups for each of the three years are presented in Tables 2 - 4.

The most commonly used fit statistic for covariance structure analysis is the χ^2 statistic. A small, nonsignificant value for χ^2 is usually indicative of adequate model fit. However, this statistic is functionally dependent on sample size; thus, for large sample sizes such as those used in this study, χ^2 is not a practical test of model fit (Cheung & Rensvold, 2002). When using large samples, other indices are commonly considered when evaluating model fit. It is common to use and report multiple fit indices when evaluating structural equation models (Cheung & Rensvold, 2002).

The root mean square error of approximation (RMSEA) provides an indication of the relationship between the predicted factor pattern of each scale and the observed factor pattern derived from the data. Unlike other fit indices, RMSEA accommodates for the effects of model complexity (Cheung & Rensvold, 2002). RMSEA values are equal to 0.00 when the model provides a perfect fit to the data. Values less than .05 are generally considered an indication of good fit, and values between .05 and .10 indicate moderate fit (Byrne, 1998; Reise et al., 1993).

In addition to examining RMSEA values, multiple indicators of fit should be used to compare models because each of the standard indices of fit is imperfect in some way (Byrne, 1998; Cheung & Rensvold, 2002; Reise et al., 1993). The Comparative Fit Index (CFI) begins with a baseline model in which all items are presumed to be uncorrelated, and makes incremental comparisons until the hypothesized fitted model is achieved. CFI statistics closer to 1.0 indicate better model fit (Byrne, 1998; Cheung & Rensvold, 2002; Reise et al., 1993). However, models with more items and factors, such as the model used in this study, typically reveal smaller CFI and TLI values because these statistics omit small, theoretically significant factor loadings and error terms (Cheung & Rensvold, 2002).

Although not a perfect fit to the data, the fit for Model 1 across all three measurement occasions was acceptable for all rater groups with no RMSEA values obtained above .10. The other fit statistics provided similar indication of a modest but acceptable fit. Typically, measurement invariance is evaluated with a likelihood ratio test (difference of χ^2); however, as noted by Cheung and Rensvold (2002), it does not make sense to argue against using the χ^2 statistic when evaluating model fit with large sample sizes and to then advocate using a likelihood ratio test to evaluate measurement invariance.

Table 2

Fit Statistics for Model 1 for Year 2001

			Rating Source Groups		
Fit Statistic	df	p	Self	Boss	Direct Report
Chi Square	314	<.0000	889.606	1034.951	1741.665
CFI			.906	.894	.860
TLI			.895	.882	.844
RMSEA			.056	.062	.088
SRMR			.048	.055	.059

Table 3

Fit Statistics for Model 1 for Year 2002

			Rating Source Groups		
Fit Statistic	df	p	Self	Boss	Direct Report
Chi Square	314	<.0000	981.906	1057.918	1866.697
CFI			.892	.887	.865
TLI			.879	.874	.849
RMSEA			.060	.063	.092
SRMR			.047	.053	.061

Table 4

Fit Statistics for Model 1 for Year 2003

			Rating Source Groups		
Fit Statistic	df	p	Self	Boss	Direct Report
Chi Square	314	<.0000	923.449	912.308	1835.802
CFI			.897	.923	.876
TLI			.885	.913	.862
RMSEA			.057	.057	.091
SRMR			.047	.044	.060

For the second level of invariance testing, Model 2, the data provided evidence of moderate, but acceptable fit. See Table 5 for Chi Square statistics and fit indices and Table 6 for factor correlations. The factors are intercorrelated at relatively high levels, with the exception of Factor 5, which demonstrated lower correlations with the other four factors across all rater groups. Acceptable fit for Model 2 indicates that the correlations among the five factors do not differ by measurement occasion for all rating source groups.

Table 5

Fit Statistics for Model 2 of Measurement Invariance Testing

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	942	<.0000	3666.222	3906.521	3671.559
CFI			.883	.878	.895
TLI			.869	.863	.882
RMSEA			.070	.073	.070
SRMR			.054	.054	.051

Table 6

Rating Source Measurement Invariance Factor Correlations

2001

	Self-Ratings					Boss Ratings					Direct Report Ratings				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.74	1.0				.67	1.0				.75	1.0			
F3	.79	.81	1.0			.76	.76	1.0			.71	.64	1.0		
F4	.72	.75	.81	1.0		.66	.76	.86	1.0		.70	.75	.83	1.0	
F5	.48	.65	.51	.55	1.0	.43	.55	.50	.51	1.0	.51	.72	.36	.56	1.0

2002

	Self-Ratings					Boss Ratings					Direct Report Ratings				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.74	1.0				.69	1.0				.77	1.0			
F3	.78	.81	1.0			.79	.79	1.0			.77	.73	1.0		
F4	.74	.81	.90	1.0		.70	.79	.88	1.0		.73	.78	.86	1.0	
F5	.58	.72	.56	.60	1.0	.46	.64	.57	.58	1.0	.61	.79	.49	.63	1.0

2003

	Self-Ratings					Boss Ratings					Direct Report Ratings				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.77	1.0				.74	1.0				.75	1.0			
F3	.76	.76	1.0			.76	.75	1.0			.75	.65	1.0		
F4	.76	.80	.86	1.0		.76	.77	.85	1.0		.75	.70	.85	1.0	
F5	.54	.67	.59	.65	1.0	.49	.64	.52	.63	1.0	.58	.74	.49	.58	1.0

The third level of testing for measurement invariance, Model 3, involves freeing the factor correlations to be estimated separately by rating source group, while constraining the factor loadings to be equal across groups. See Table 7 for Chi Square statistics and fit indices and Tables 8 - 10 for factor loadings. Again, the RMSEA values are indicative of moderate, but acceptable fit. The results indicated that the factor loadings across groups are similar.

Table 7. Fit Statistics for Model 3 of Measurement Invariance Testing

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	996	<.0000	3863.755	4023.492	3878.021
CFI			.876	.875	.889
TLI			.869	.868	.882
RMSEA			.070	.072	.070
SRMR			.101	.087	.137

Table 8

Year 2001 Standardized Factor Loadings

Item	Factor	Rating Source Group		
		Self	Boss	Direct Report
1	Envision	0.568	0.584	0.763
2	Envision	0.597	0.618	0.756
3	Envision	0.656	0.598	0.780
4	Envision	0.689	0.716	0.835
5	Envision	0.651	0.680	0.807
6	Envision	0.590	0.600	0.748
7	Envision	0.615	0.680	0.779
8	Energize	0.529	0.520	0.664
9	Energize	0.609	0.604	0.706
10	Energize	0.616	0.663	0.682
11	Energize	0.609	0.690	0.749
12	Energize	0.654	0.675	0.780
13	Energize	0.664	0.696	0.794
14	Energize	0.689	0.733	0.828
15	Edge	0.617	0.653	0.790
16	Edge	0.611	0.643	0.790
17	Edge	0.657	0.708	0.768
18	Edge	0.626	0.667	0.841
19	Execute	0.602	0.601	0.757
20	Execute	0.623	0.657	0.815
21	Execute	0.672	0.661	0.829
22	Execute	0.569	0.579	0.708
23	Execute	0.615	0.655	0.773
24	Execute	0.642	0.691	0.791
25	Ethics/Character	0.778	0.830	0.885
26	Ethics/Character	0.811	0.828	0.896
27	Ethics/Character	0.642	0.655	0.742

Table 9

Year 2002 Standardized Factor Loadings

Item	Factor	Rating Source Group		
		Self	Boss	Direct Report
1	Envision	0.633	0.641	0.817
2	Envision	0.586	0.612	0.742
3	Envision	0.651	0.629	0.819
4	Envision	0.708	0.692	0.871
5	Envision	0.682	0.658	0.840
6	Envision	0.593	0.600	0.756
7	Envision	0.628	0.626	0.759
8	Energize	0.593	0.554	0.737
9	Energize	0.634	0.607	0.743
10	Energize	0.589	0.597	0.670
11	Energize	0.611	0.657	0.762
12	Energize	0.667	0.616	0.796
13	Energize	0.680	0.672	0.815
14	Energize	0.700	0.656	0.859
15	Edge	0.658	0.640	0.819
16	Edge	0.663	0.639	0.789
17	Edge	0.646	0.670	0.762
18	Edge	0.686	0.648	0.822
19	Execute	0.590	0.584	0.726
20	Execute	0.596	0.633	0.758
21	Execute	0.674	0.662	0.810
22	Execute	0.629	0.622	0.724
23	Execute	0.668	0.667	0.813
24	Execute	0.702	0.682	0.848
25	Ethics/Character	0.714	0.707	0.865
26	Ethics/Character	0.663	0.646	0.782
27	Ethics/Character	0.657	0.654	0.805

Table 10

Year 2003 Standardized Factor Loadings

Item	Factor	Rating Source Group		
		Self	Boss	Direct Report
1	Envision	0.649	0.648	0.857
2	Envision	0.663	0.690	0.857
3	Envision	0.688	0.627	0.847
4	Envision	0.718	0.714	0.869
5	Envision	0.685	0.688	0.847
6	Envision	0.620	0.617	0.818
7	Envision	0.672	0.673	0.812
8	Energize	0.526	0.507	0.693
9	Energize	0.672	0.691	0.842
10	Energize	0.572	0.587	0.645
11	Energize	0.602	0.670	0.752
12	Energize	0.636	0.635	0.811
13	Energize	0.726	0.725	0.876
14	Energize	0.721	0.733	0.895
15	Edge	0.617	0.607	0.850
16	Edge	0.674	0.672	0.863
17	Edge	0.685	0.707	0.825
18	Edge	0.690	0.669	0.863
19	Execute	0.614	0.624	0.806
20	Execute	0.634	0.631	0.844
21	Execute	0.678	0.671	0.846
22	Execute	0.597	0.617	0.740
23	Execute	0.650	0.617	0.788
24	Execute	0.715	0.706	0.856
25	Ethics/Character	0.713	0.705	0.877
26	Ethics/Character	0.656	0.666	0.827
27	Ethics/Character	0.614	0.673	0.768

Longitudinal Measurement Invariance

Model 1 of the test for longitudinal measurement invariance fit the hypothesized five-factor leadership performance model to each of the three measurement occasions. For each group, there were 588 observations, 27 items and 5 hypothesized factors. A summary of the fit statistics obtained for measurement occasions for each of the three rating source groups is presented in Tables 11 - 13. Again, while not a perfect fit to the data, the fit for Model 1 across all three measurement occasions is acceptable for all rater groups with no RMSEA values obtained above .10.

Table 11

Model 1 Fit Statistics for Self-Ratings

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	314	<.0000	889.606	981.906	923.449
CFI			.906	.892	.897
TLI			.895	.879	.885
RMSEA			.056	.060	.057
SRMR			.048	.047	.047

Table 12

Model 1 Fit Statistics for Boss Ratings

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	314	<.0000	1034.951	1057.918	912.308
CFI			.894	.887	.923
TLI			.882	.884	.913
RMSEA			.062	.063	.057
SRMR			.055	.053	.044

Table 13

Model 1 Fit Statistics for Direct Report Ratings

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	314	<.0000	1741.665	1057.918	1866.697
CFI			.860	.887	.865
TLI			.844	.884	.849
RMSEA			.088	.063	.092
SRMR			.059	.053	.061

The data provide evidence of moderate, but acceptable fit for Model 2. See Table 14 for Chi Square statistics and fit indices and Table 15 for factor correlations. The factors are intercorrelated at relatively high levels, with the exception of Factor 5, which demonstrated lower correlations with the other four factors. The correlations among the five factors do not differ by measurement occasion for all rating source groups.

Table 14

Fit Statistics for Model 2 Measurement Invariance Testing

			Rating Source Groups		
Fit Statistic	df	p	Self	Boss	Direct Report
Chi Square	986	<.0000	2957.381	3207.215	5912.078
CFI			.892	.895	.855
TLI			.884	.888	.845
RMSEA			.058	.062	.092
SRMR			.048	.052	.072

Table 15

Factor Correlations

Self-Ratings

	2001					2002					2003				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.74	1.0				.74	1.0				.77	1.0			
F3	.79	.81	1.0			.78	.81	1.0			.76	.76	1.0		
F4	.72	.75	.81	1.0		.74	.81	.90	1.0		.76	.80	.86	1.0	
F5	.48	.65	.51	.55	1.0	.58	.72	.56	.60	1.0	.54	.67	.59	.65	1.0

Boss Ratings

	2001					2002					2003				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.67	1.0				.69	1.0				.74	1.0			
F3	.76	.76	1.0			.79	.79	1.0			.76	.75	1.0		
F4	.66	.76	.86	1.0		.70	.79	.88	1.0		.76	.77	.85	1.0	
F5	.43	.55	.50	.51	1.0	.46	.64	.57	.58	1.0	.49	.64	.52	.63	1.0

Direct Report Ratings

	2001					2002					2003				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.75	1.0				.77	1.0				.75	1.0			
F3	.71	.64	1.0			.77	.73	1.0			.75	.65	1.0		
F4	.70	.75	.83	1.0		.73	.78	.86	1.0		.75	.70	.85	1.0	
F5	.51	.72	.36	.56	1.0	.61	.79	.49	.63	1.0	.58	.74	.49	.58	1.0

Finally, Model 3 also provided evidence of acceptable fit. See Table 16 for Chi Square statistics and fit indices and Tables 17 - 19 for factor loadings. The results indicated that the factor loadings across years are consistent.

Table 16

Fit Statistics for Model 3 Measurement Invariance Testing

			Rating Source Groups		
Fit Statistic	df	p	Self	Boss	Direct Report
Chi Square	996	<.0000	2862.386	3103.667	5603.919
CFI			.898	.900	.864
TLI			.892	.895	.857
RMSEA			.056	.060	.089
SRMR			.056	.075	.081

Table 17

Standardized Factor Loadings for Self-Ratings

Item	Factor	Measurement Occasion		
		2001	2002	2003
1	Envision	0.598	0.597	0.626
2	Envision	0.638	0.569	0.663
3	Envision	0.671	0.672	0.679
4	Envision	0.674	0.698	0.699
5	Envision	0.639	0.650	0.660
6	Envision	0.615	0.608	0.623
7	Envision	0.647	0.629	0.683
8	Energize	0.568	0.592	0.570
9	Energize	0.592	0.608	0.632
10	Energize	0.550	0.552	0.568
11	Energize	0.632	0.627	0.635
12	Energize	0.643	0.649	0.640
13	Energize	0.682	0.695	0.695
14	Energize	0.667	0.659	0.663
15	Edge	0.661	0.678	0.671
16	Edge	0.640	0.651	0.679
17	Edge	0.668	0.668	0.671
18	Edge	0.647	0.639	0.684
19	Execute	0.669	0.609	0.666
20	Execute	0.672	0.568	0.674
21	Execute	0.686	0.687	0.713
22	Execute	0.602	0.588	0.597
23	Execute	0.614	0.644	0.650
24	Execute	0.677	0.696	0.704
25	Ethics/Character	0.783	0.735	0.733
26	Ethics/Character	0.783	0.685	0.697
27	Ethics/Character	0.626	0.650	0.636

Table 18

Standardized Factor Loadings for Boss Ratings

Item	Factor	Measurement Occasion		
		2001	2002	2003
1	Envision	0.655	0.647	0.667
2	Envision	0.703	0.639	0.730
3	Envision	0.631	0.666	0.638
4	Envision	0.744	0.727	0.738
5	Envision	0.737	0.697	0.731
6	Envision	0.643	0.633	0.638
7	Envision	0.744	0.663	0.718
8	Energize	0.557	0.552	0.550
9	Energize	0.608	0.601	0.669
10	Energize	0.665	0.633	0.654
11	Energize	0.651	0.610	0.644
12	Energize	0.709	0.644	0.684
13	Energize	0.708	0.682	0.689
14	Energize	0.768	0.678	0.734
15	Edge	0.718	0.684	0.686
16	Edge	0.721	0.684	0.734
17	Edge	0.717	0.690	0.689
18	Edge	0.754	0.675	0.744
19	Execute	0.683	0.623	0.694
20	Execute	0.737	0.641	0.708
21	Execute	0.674	0.675	0.705
22	Execute	0.634	0.604	0.638
23	Execute	0.700	0.689	0.663
24	Execute	0.752	0.706	0.724
25	Ethics/Character	0.869	0.763	0.758
26	Ethics/Character	0.821	0.692	0.730
27	Ethics/Character	0.608	0.614	0.660

Table 19

Standardized Factor Loadings for Direct Report Ratings

Item	Factor	Measurement Occasion		
		2001	2002	2003
1	Envision	0.770	0.770	0.829
2	Envision	0.738	0.668	0.815
3	Envision	0.751	0.800	0.807
4	Envision	0.810	0.855	0.847
5	Envision	0.788	0.809	0.820
6	Envision	0.719	0.719	0.776
7	Envision	0.745	0.694	0.764
8	Energize	0.642	0.680	0.677
9	Energize	0.724	0.752	0.841
10	Energize	0.635	0.654	0.660
11	Energize	0.753	0.758	0.766
12	Energize	0.756	0.766	0.802
13	Energize	0.801	0.820	0.849
14	Energize	0.839	0.856	0.880
15	Edge	0.725	0.739	0.806
16	Edge	0.752	0.716	0.818
17	Edge	0.753	0.753	0.792
18	Edge	0.825	0.741	0.830
19	Execute	0.695	0.612	0.738
20	Execute	0.777	0.633	0.805
21	Execute	0.786	0.764	0.823
22	Execute	0.708	0.654	0.712
23	Execute	0.744	0.768	0.759
24	Execute	0.777	0.808	0.811
25	Ethics/Character	0.858	0.839	0.850
26	Ethics/Character	0.855	0.769	0.828
27	Ethics/Character	0.738	0.807	0.794

Finally, ratings for each target were combined across all rating sources and submitted to tests of longitudinal measurement equivalence. Chi square statistics and fit indices are shown in Tables 20, 21 and 23. Factor correlations are shown in Table 22 and factor loadings are shown in Table 24. Once again, Models 1-3 provided evidence of acceptable fit with no RMSEA values exceeding the .10 limit.

Table 20

Model 1 Fit Statistics for All Ratings Combined

			Year of Measurement		
Fit Statistic	df	p	2001	2002	2003
Chi Square	314	<.0000	1696.221	1909.603	1578.411
CFI			.845	.829	.873
TLI			.826	.809	.858
RMSEA			.087	.093	.083
SRMR			.066	.068	.061

Table 21

Model 2 Fit Statistics for All Ratings Combined

Fit Statistic	df	p	
Chi Square	942	<.0000	5184.266
CFI			.850
TLI			.832
RMSEA			.088
SRMR			.065

Table 22

Factor Correlations for All Ratings Combined

	2001					2002					2003				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
F1	1.0					1.0					1.0				
F2	.68	1.0				.67	1.0				.68	1.0			
F3	.71	.68	1.0			.74	.71	1.0			.70	.65	1.0		
F4	.64	.73	.85	1.0		.65	.75	.86	1.0		.69	.72	.84	1.0	
F5	.40	.60	.40	.45	1.0	.44	.69	.39	.48	1.0	.43	.63	.41	.54	1.0

Table 23

Model 3 Fit Statistics for All Ratings Combined

Fit Statistic	df	p	
Chi Square	996	<.0000	5279.942
CFI			.848
TLI			.839
RMSEA			.086
SRMR			.076

Table 24

Standardized Factor Loadings for All Ratings Combined

Item	Factor	Measurement Occasion		
		2001	2002	2003
1	Envision	0.723	0.716	0.766
2	Envision	0.744	0.677	0.787
3	Envision	0.695	0.726	0.712
4	Envision	0.811	0.812	0.827
5	Envision	0.746	0.746	0.759
6	Envision	0.696	0.696	0.705
7	Envision	0.756	0.717	0.763
8	Energize	0.609	0.613	0.605
9	Energize	0.657	0.655	0.752
10	Energize	0.620	0.634	0.637
11	Energize	0.703	0.691	0.707
12	Energize	0.726	0.703	0.727
13	Energize	0.794	0.773	0.781
14	Energize	0.799	0.768	0.804
15	Edge	0.750	0.758	0.742
16	Edge	0.756	0.754	0.813
17	Edge	0.754	0.726	0.746
18	Edge	0.774	0.737	0.811
19	Execute	0.716	0.657	0.735
20	Execute	0.739	0.650	0.759
21	Execute	0.746	0.725	0.777
22	Execute	0.625	0.600	0.638
23	Execute	0.733	0.742	0.729
24	Execute	0.795	0.798	0.796
25	Ethics/Character	0.854	0.784	0.782
26	Ethics/Character	0.847	0.743	0.763
27	Ethics/Character	0.655	0.715	0.707

Latent Growth Modeling

Measurement equivalence across the three measurement occasions was confirmed; thus, analyses continued with the evaluation of the LGM of leadership performance with multiple continuous variables defined above (see Figure 2a). In this analysis, the five factor composite scores were entered into the model as scaled manifest variables across three points in time. This model estimated growth factors, an intercept with mean fixed to equal 0 and variance freely estimated, and a slope with mean and variance freely estimated. The disturbances of the growth factors were uncorrelated and the intercept and slope were free to covary. Upon the specification of an adequate LGM, the final model for each rater group was augmented with the covariates (i.e., sector and subdivision) and the outcome variables (i.e., promotion and consensus performance). Finally, the boss group was augmented with the rater composition moderator variable.

Before performing LGM analyses using direct report ratings, a comparison was made between leaders who had ratings from the same direct report each year (i.e., consistent direct report ratings) and leaders who had ratings from different direct reports each year (i.e., inconsistent direct report ratings). If a leader had ratings from more than one rater, these ratings were averaged and the mean rating was used in further analyses. These two groups were examined for measurement equivalence using the procedure described earlier. Tests indicated that the leadership performance model was equivalent between consistent and inconsistent direct report groups. Table 25 contains the fit statistics for the three measurement equivalence models.

Table 25

Tests of Measurement Equivalence for Consistent and Inconsistent Direct Reports

			Measurement Models		
Fit Statistic	df	p	Model 1	Model 2	Model 3
Chi Square	314, 628, 655	<.0000	2917.267	3307.837	3344.873
CFI			.895	.891	.891
TLI			.883	.878	.883
RMSEA			.066	.067	.066
SRMR			.047	.050	.064

LGM analyses were then conducted for consistent direct report ratings and for inconsistent direct report ratings. Fit statistics for these models are shown in Table 26. Estimated growth curves for both groups are shown in Figures 3 and 4.

Table 26

Fit Statistics for Consistent and Inconsistent Direct Report Latent Growth Models

			Direct Report Groups	
Fit Statistic	df	p	Consistent	Inconsistent
Chi Square	100	<.0000	920.408	1246.586
CFI			.786	.753
TLI			.775	.740
RMSEA			.157	.164
SRMR			.069	.095

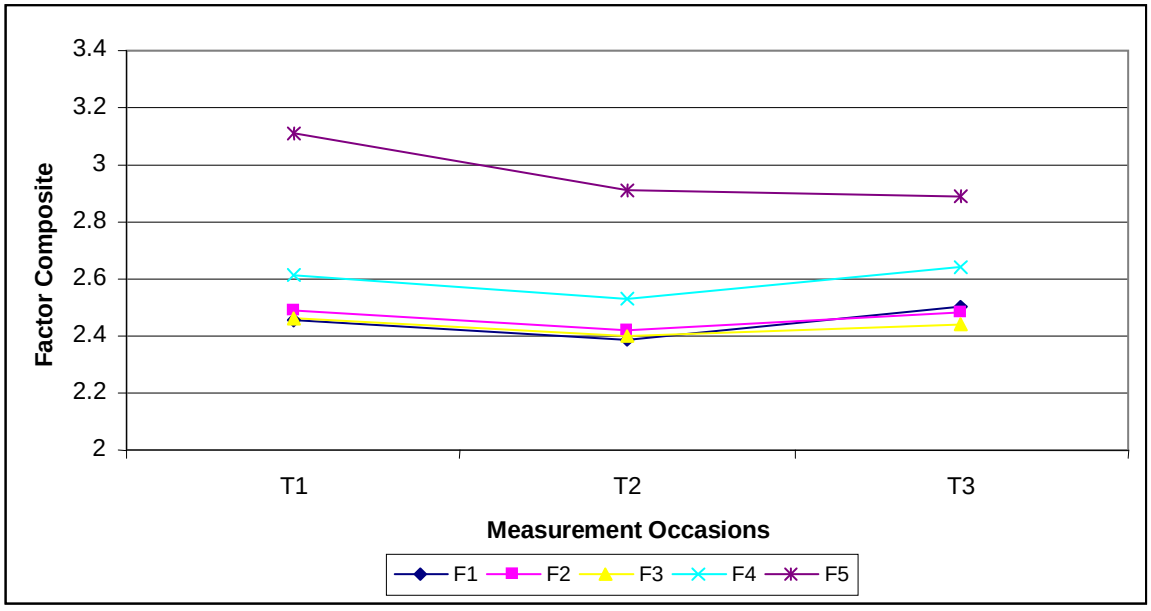

Figure 3. Growth Curves for Consistent Direct Report Ratings

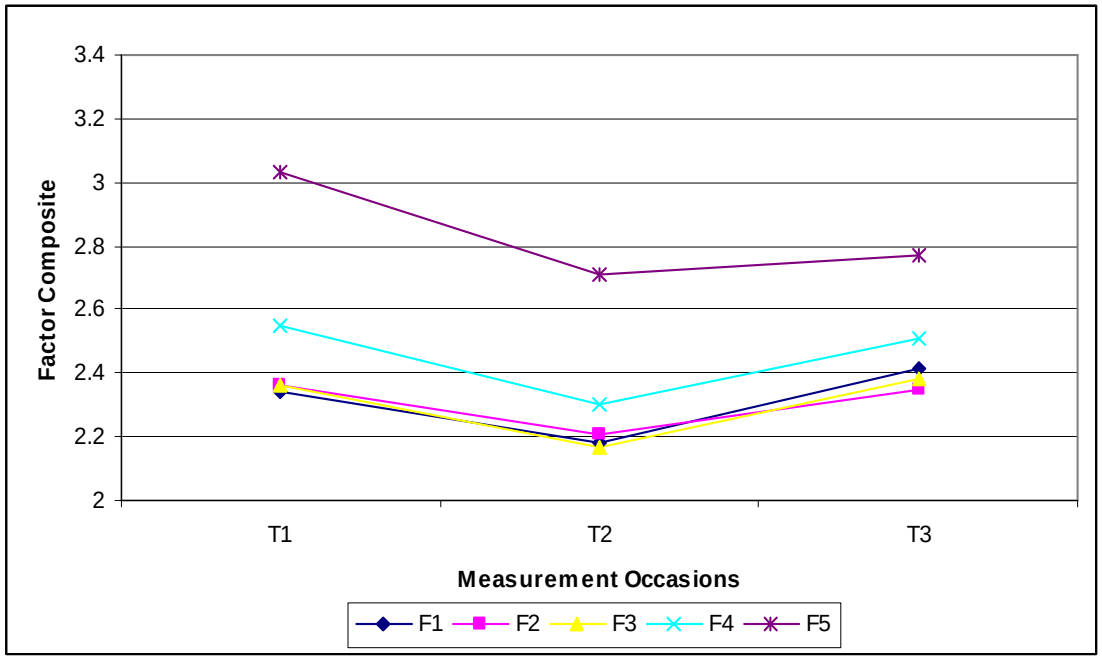

Figure 4. Growth Curves for Inconsistent Direct Report Ratings

Both direct report models exhibited poor model fit and followed similar growth curvatures as indicated by similar intercept and slope covariances (-.074 for consistent raters and -.036 for inconsistent raters). Interestingly, the consistent direct report group tended to rate leaders more highly than the inconsistent direct report group. Comparisons between each direct report group by measurement occasion are shown in Figures 5-7.

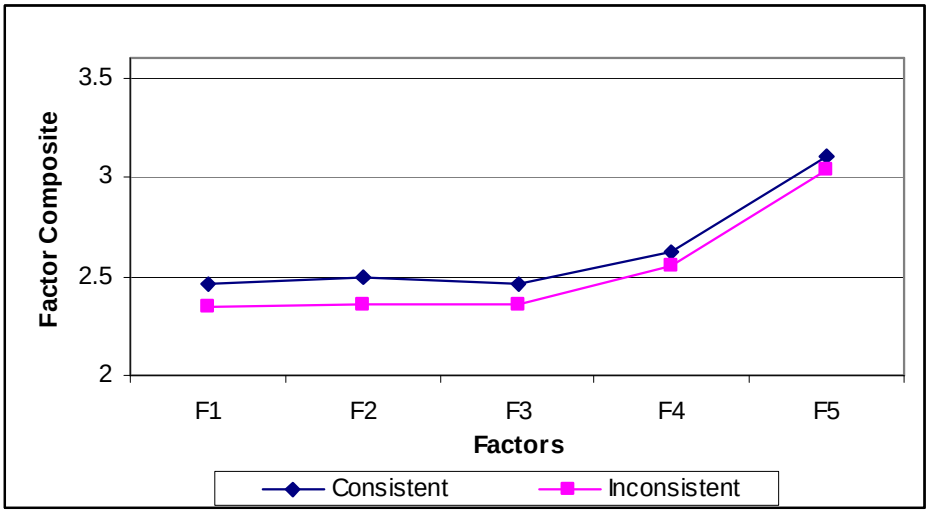

Figure 5. Direct Report Comparison at Time 1

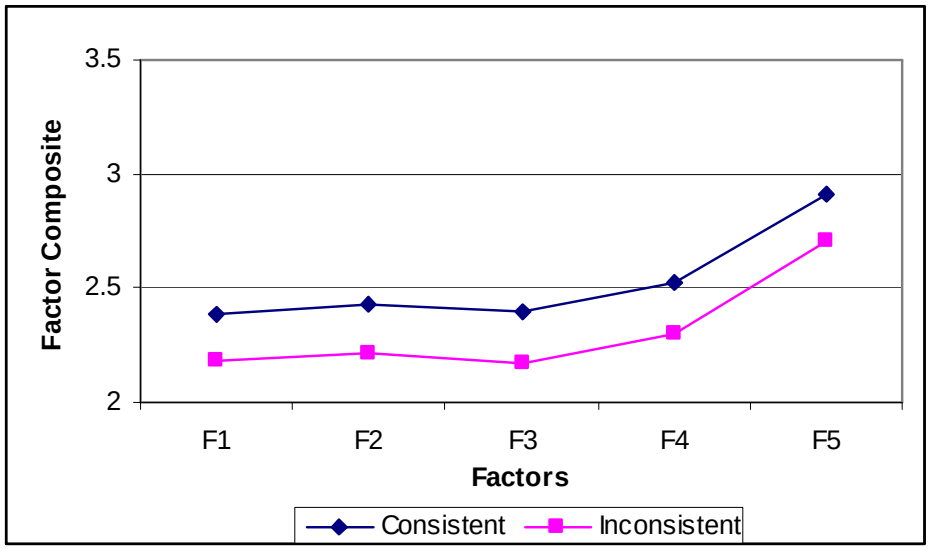

Figure 6. Direct Report Comparison at Time 2

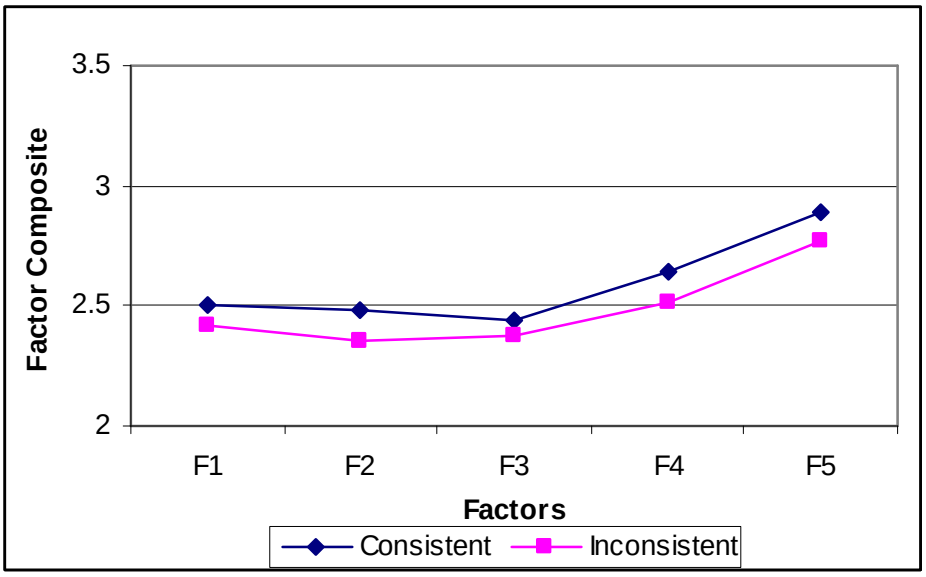

Figure 7. Direct Report Comparison at Time 3

The two direct report groups had similar model fit and similar growth curves, but the consistent direct report model had a smaller chi-square value and slightly higher CFI and TLI estimates, indicating a slightly better model fit. Therefore, analyses continued using only rating targets that had a boss rating, a self-rating, and a mean score from consistent direct reports for each measurement occasion. Excluding managers whose rater group composition changed across the three measurement occasions also made it possible to rule out such composition changes as a possible explanation for subsequent results. This reduced the sample size to 331 leaders with complete ratings.

Next, LGM analyses were performed for each rater group to address research questions 3 and 4. To review, research question 3 asked what form the latent growth curve takes for longitudinal leadership performance. Research question 4 asked whether longitudinal leadership performance trajectories are similar for different rater groups. Model fit statistics are presented in Table 27 for all rater groups. These fit statistics

provide evidence of linear growth for boss ratings, but not for self-ratings, direct report ratings, or the mean ratings across all rater groups.

Table 27

LGM Fit Statistics for all Rater Groups

			Rater Groups			
Fit Statistic	df	p	Boss	Self	Direct Report	All
Chi Square	100	<.0000	326.633	932.797	920.408	971.940
CFI			.916	.757	.786	.764
TLI			.912	.745	.775	.752
RMSEA			.083	.159	.157	.162
SRMR			.054	.071	.069	.076

Estimated growth curves for all rater groups are shown in Figures 8 - 11. Factor composite scores for all groups and all measurement occasions are shown in Table 28. Ratings on Envision (F1) for all rater groups showed a decrease from 2001 to 2002, but then an increase from 2002 to 2003. Ratings on Energize (F2), Edge (F3), and Execute (F4) showed a steady decrease in the boss rater group, but showed a steady increase in self-ratings. Ratings on Energize, Edge, and Execute from direct reports and all groups combined decreased from 2001 to 2002 and increased from 2002 to 2003. Ethics and Character (F5) ratings showed a steady decrease for all rater groups. The steady decrease in ratings on four of the five leadership performance factors in the boss rating group is the likely reason that this rater group was the only one for which a LGM fit the data.

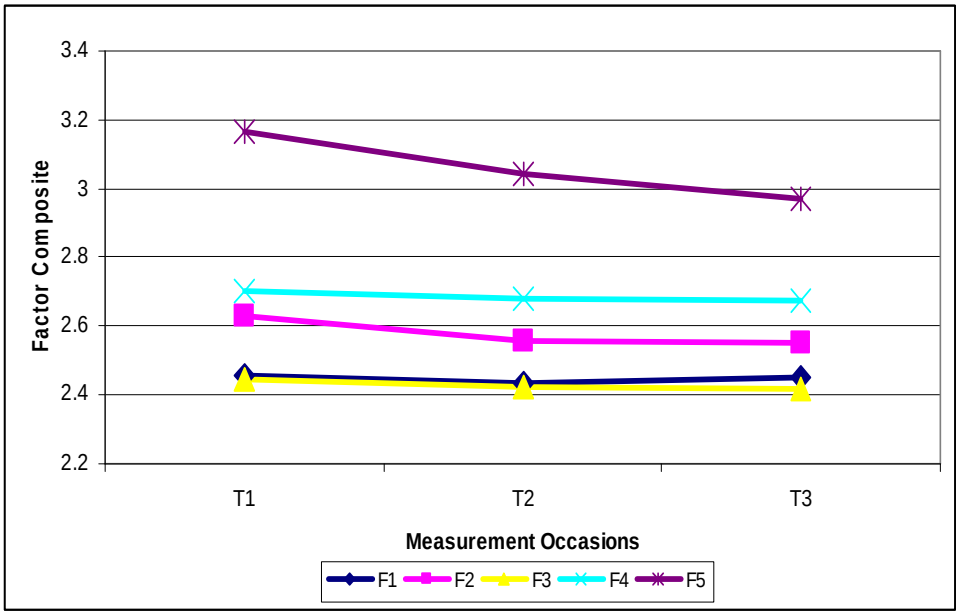

Figure 8. Growth Curves for Boss Ratings

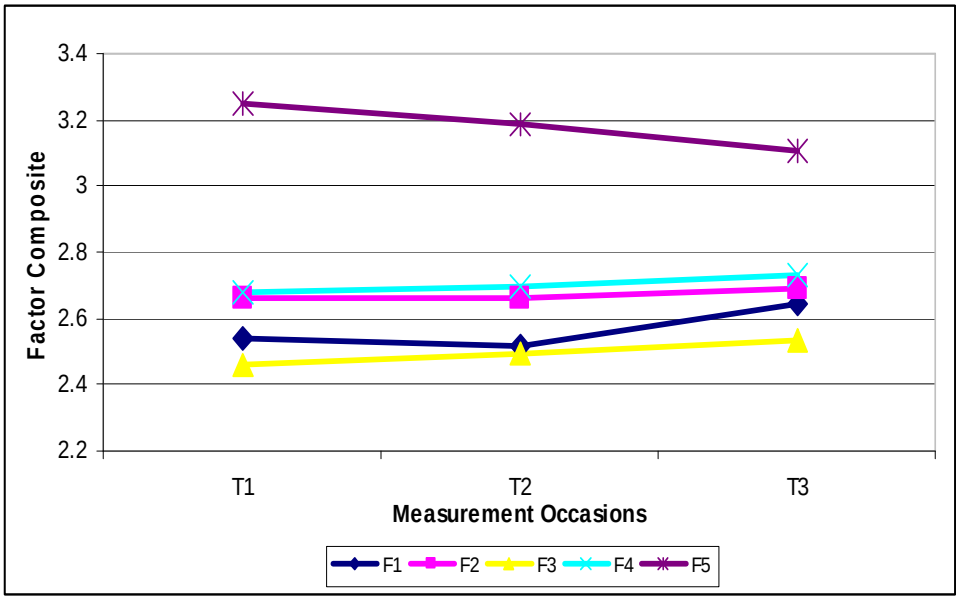

Figure 9. Growth Curves for Self-Ratings

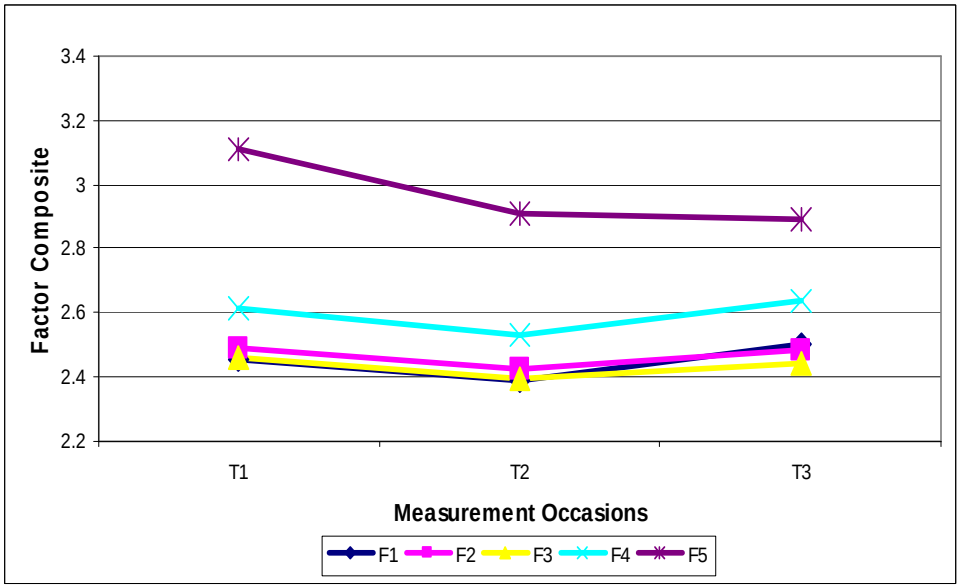

Figure 10. Growth Curves for Direct Report Ratings

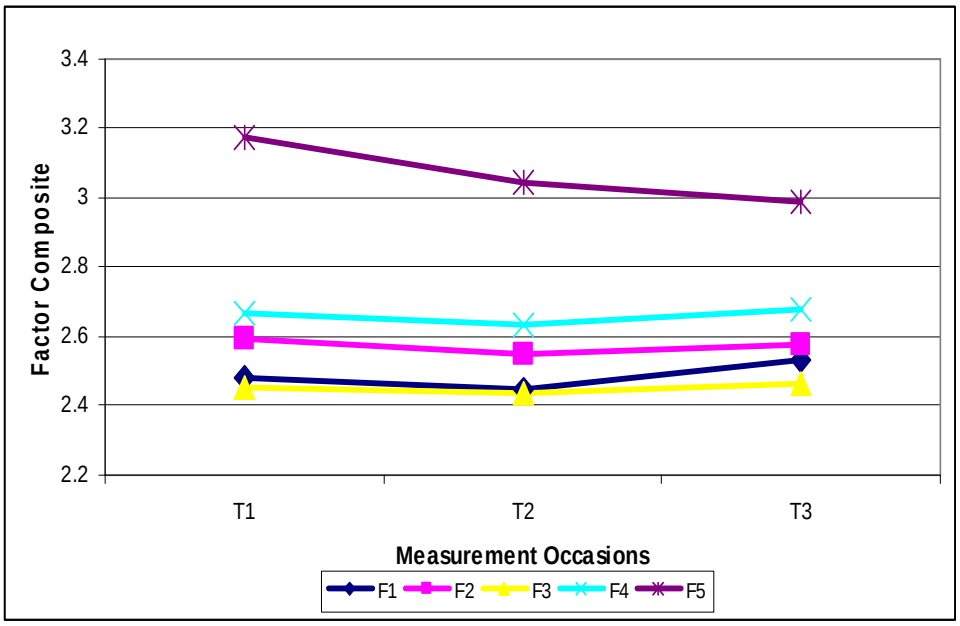

Figure 11. Growth Curves for All Ratings Combined

Table 28

Factor composite scores by measurement occasion

Boss Ratings	2001	2002	2003
F1	2.456	2.432	2.452
F2	2.631	2.558	2.552
F3	2.447	2.422	2.417
F4	2.703	2.678	2.674
F5	3.167	3.040	2.973
Self-Ratings			
F1	2.540	2.520	2.642
F2	2.663	2.664	2.689
F3	2.459	2.495	2.534
F4	2.681	2.694	2.732
F5	3.249	3.188	3.107
Direct Report Ratings			
F1	2.456	2.388	2.501
F2	2.490	2.424	2.485
F3	2.462	2.398	2.444
F4	2.616	2.530	2.638
F5	3.110	2.912	2.890
All Ratings Combined			
F1	2.484	2.447	2.532
F2	2.595	2.548	2.575
F3	2.456	2.438	2.465
F4	2.666	2.634	2.681
F5	3.175	3.047	2.990

Growth Mixture Modeling

Growth mixture modeling relaxes the assumption of a single population to allow for parameter differences in unobserved, latent subpopulations. Instead of examining individual variation around one mean growth curve, a growth mixture model (GMM) allows the subpopulations to vary around different mean growth curves. Essentially, growth mixture modeling examines a growth model for each subpopulation, or latent class (Muthen, 2004). Therefore, subpopulation membership can be inferred from different Time 1 leadership performance ratings and development trajectories. Growth mixture modeling also allows for the inclusion of covariates and distal outcomes that can help to define subpopulation membership by estimating each individual's probability of membership along with their score on the estimated growth factors.

The LGM for each rater group was next augmented with covariates to answer research questions 5 and 6. Research question 5 asked if the rate of leadership performance change was contingent upon rater composition in the boss rating group. Question 6 asked if the effects of the covariates were mediated by initial status and/or change in leadership performance.

The addition of covariates requires analyses by factor composite; thus, each leadership performance factor was analyzed separately for each rater group. Measures of classification quality based on individual class probabilities such as Bayesian information criterion (BIC) values, entropy, and latent class probabilities are considered when evaluating model fit (Muthen, 2004). A low BIC value corresponds to a high loglikelihood value and a parsimonious model. An entropy value and average latent class probabilities closer to 1.0 are indicative of better classification precision and better model

fit (Muthen, 2004). Unfortunately, because growth mixture modeling is still relatively new in the literature, tests for significant differences between models do not yet exist. Instead, model evaluation must be based upon the available statistics and the theory supporting each model (Muthen, 2004).

Growth Mixture Modeling with Covariates

A GMM was run for all rater groups without covariates, with each covariate separately, and then with all covariates simultaneously. The GMMs were run in this sequence because excluding covariates might result in a distorted analysis because the covariate(s) might have direct effects on the growth factors. This occurs because the items become incorrectly related to the latent class variable if the covariates are excluded, which is similar to leaving out an important predictor in a multiple regression analysis, thus distorting the slope of the remaining predictors. In growth mixture modeling, the covariates may be needed to correctly specify latent class membership (Muthen, 2004). Running the GMMs in this sequence allowed for comparisons across models with and without each covariate. It should not be expected that latent class size or individual membership classification remain the same when modeling with and without covariates; however "it is the model with covariates properly included that gives the better answer" (p. 355; Muthen, 2004). Occasionally, two or more models fit the data about the same leaving no statistical basis on which to declare the better model. This is not a problem, but merely indicates that there are two ways of looking at the same reality. Occasionally, modeling is more art than science and decisions about model fit should be based in the theory behind the model, predictive validity, and practical usefulness (Muthen, 2004).

Model fit results for self-ratings of the leadership performance factor Envision are presented in Table 29. The addition of the sector and subdivision covariates separately did not improve model fit, but did improve the average latent class probabilities. The addition of both covariates improved model fit and average latent class probabilities resulting in the best fitting model. However, neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model slightly reduced growth factor residual variances, but adding the subdivision covariate to the model separately did not reduce any residual variances.

In 2001, Lo, Mendell, and Rubin proposed a likelihood ratio test for evaluating the number of latent classes for GMM. A low p-value indicates that the model with fewer classes should be rejected for the model with an additional class (Muthen, 2004). The Lo-Mendell-Rubin likelihood ratio test (LMR LRT) indicated that only one latent class was present in the data. Thus, the addition of the contextual covariates did not result in the identification of separate growth classes.

Table 29

Self-Rating Growth Mixture Model with Covariates for Envision

	Covariates			
	None	Sector	Subdivision	Both
BIC	1254.495	1261.666	1263.155	1275.067
Free Parameters	11	14	14	17
Entropy	0.960	0.789	0.953	0.991
Class 1 N	4	123	102	89
Class 2 N	324	205	226	239
Class 1 Probability	0.775	0.921	0.996	0.995
Class 2 Probability	0.994	0.940	0.978	0.997

Model fit results for Energize are presented in Table 30. The addition of the sector covariate improved model fit and the average latent class probabilities. The addition of the subdivision covariate also improved both model fit and the average latent class probabilities. The addition of both covariates improved model fit and average latent class probabilities even further. However, neither covariate was a significant predictor of the slope or intercept. The addition of the covariates to the model failed to reduce measurement or growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 30

Self-Rating Growth Mixture Model with Covariates for Energize

	Covariates			
	None	Sector	Subdivision	Both
BIC	1163.571	1163.073	1173.328	1188.819
Free Parameters	11	14	14	17
Entropy	0.538	0.918	1.000	1.000
Class 1 N	14	57	83	88
Class 2 N	314	271	245	240
Class 1 Probability	0.640	0.980	1.000	1.000
Class 2 Probability	0.883	1.000	1.000	1.000

Model fit results for Edge are presented in Table 31. The addition of the sector covariate improved model fit and the average latent class probabilities. However, the addition of the subdivision covariate did not improve model fit or average latent class probabilities. The addition of both covariates also did not improve model fit or average latent class probabilities. Neither covariate was a significant predictor of the slope or intercept. Adding neither the sector nor subdivision covariates to the model reduced measurement or growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 31

Self-Rating Growth Mixture Model with Covariates for Edge

	Covariates			
	None	Sector	Subdivision	Both
BIC	1313.228	1318.926	1327.104	1335.493
Free Parameters	11	14	14	17
Entropy	0.989	1.000	0.953	0.956
Class 1 N	1	96	20	6
Class 2 N	327	232	308	322
Class 1 Probability	0.998	1.000	0.979	0.849
Class 2 Probability	0.986	1.000	0.990	0.994

Model fit results for Execute are presented in Table 32. The addition of the sector and subdivision covariates separately improved model fit and the average latent class probabilities. The addition of both covariates also improved model fit and average latent class probabilities. However, neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model substantially reduced measurement and growth factor residual variances, but adding the subdivision covariate to the model separately did not reduce residual variances. Despite the resulting fit statistics, the LMR LRT indicated that only one latent class was present in the data.

Table 32

Self-Rating Growth Mixture Model with Covariates for Execute

	Covariates			
	None	Sector	Subdivision	Both
BIC	1210.955	1213.142	1227.641	1234.649
Free Parameters	11	14	14	17
Entropy	0.744	0.820	1.000	1.000
Class 1 N	36	28	101	80
Class 2 N	292	300	227	248
Class 1 Probability	0.802	0.837	1.000	1.000
Class 2 Probability	0.947	0.962	1.000	1.000

Model fit results for Ethics and Character are presented in Table 33. The addition of the sector covariate improved model fit and the average latent class probabilities, but the model estimated three latent classes rather than the one latent class estimated without the addition of a covariate. The addition of the subdivision covariate did not improve model fit or average latent class probabilities. However, the addition of both covariates improved model fit and average latent class probabilities, resulting in the best model fit with three latent classes. Neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model substantially reduced measurement and growth factor residual variances, but adding the subdivision covariate to the model separately did reduce residual variances. The LMR LRT indicated that four latent classes

were present in the data; however, the four-class model was not identified. Therefore, the three-class model was retained.

Table 33

Self-Rating Growth Mixture Model with Covariates for Ethics and Character

	Covariates			
	None	Sector	Subdivision	Both
BIC	1280.941	1262.167	1297.613	1289.587
Free Parameters	11	18	14	22
Entropy	0.656	0.898	0.655	0.890
Class 1 N	194	25	132	25
Class 2 N	134	115	196	187
Class 3 N		188		116
Class 1 Probability	0.907	0.961	0.894	0.945
Class 2 Probability	0.887	0.988	0.902	0.942
Class 3 Probability		0.947		0.988

Intercepts, slopes, and estimated means for the best fitting model for each leadership performance factor are shown in Table 34. In four out of the five cases, the model containing both covariates fit the data best. Also in four out of five cases, the best fitting model contained two latent classes, one small class representing between 25% and 29% of the total sample and one large class representing the remaining 75% to 71%. In comparison to the LGM mean scores that increased and decreased over time, all GMM mean scores show a steady increase or decrease over time due to the model's ability to

estimate latent classes. For example, the average of all self-ratings combined for Envision (shown in Table 28) decreased from 2001 to 2002, but increased from 2002 to 2003. By using GMM (shown in Table 34), two latent classes now become apparent for Envision, a small class (27%) that begins with a higher intercept and increases at a faster rate than a second, larger class (73%). The same occurs for Energize, however it is the larger classes that have higher intercepts and faster growth rates for Edge and Execute. Ratings on Ethics and Character reveal three latent classes, one small class (8%) whose ratings increase rapidly, one class whose ratings slowly drop, and one class whose ratings decrease more quickly.

While the sizes of Class 1 for Envision, Energize, Edge, and Execute are similar, examining latent class membership revealed little consistency across classes. The majority of leaders in Class 1 for Envision came from one sector. The majority of leaders in Class 1 for Energize also came from one sector, but a different sector than Class 1 for Envision. Class 1 for Edge represented membership from several sectors, and Class 1 for Execute represented membership mainly from two sectors. All three latent classes for Ethics and Character represented membership from a variety of sectors and subdivisions.

Table 34

Best Fitting Models for Self-Ratings

	Model	N	Intercept	Slope	Mean 1	Mean 2	Mean 3
Envision							
Class 1	Both	89	2.635	.087	2.595	2.672	2.749
Class 2		239	2.551	.047	2.501	2.535	2.570
Energize							
Class 1	Both	88	2.773	.022	2.754	2.771	2.788
Class 2		240	2.637	.024	2.624	2.636	2.647
Edge							
Class 1	Sector	96	2.040	.045	2.377	2.389	2.402
Class 2		232	2.365	.026	2.493	2.541	2.589
Execute							
Class 1	Both	80	2.341	.020	2.555	2.563	2.571
Class 2		248	2.666	.032	2.724	2.754	2.784
Ethics and Character							
Class 1	Both	25	2.302	.407	2.297	2.690	3.083
Class 2		187	3.032	.001	3.027	3.016	3.005
Class 3		116	3.805	-.251	3.799	3.535	3.270

Next, GMMs were run for the direct report rater group. Model fit results for Envision are presented in Table 35. The addition of the sector covariate improved model fit but did not greatly improve the average latent class probabilities, and the model estimated three latent classes rather than the one latent class estimated without the addition of a

covariate. The addition of the subdivision covariate improved model fit and average latent class probabilities. The addition of both covariates improved model fit and average latent class probabilities, resulting in the best model fit with two latent classes. However, neither covariate was a significant predictor of the slope or intercept. Adding the sector and subdivision covariates to the model separately reduced measurement and growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 35

Direct Report Growth Mixture Model with Covariates for Envision

	Covariates			
	None	Sector	Subdivision	Both
BIC	1513.599	1429.898	1433.556	1449.296
Free Parameters	11	17	14	17
Entropy	0.812	0.851	0.855	1.000
Class 1 N	12	22	19	12
Class 2 N	316	17	309	316
Class 3 N		289		
Class 1 Probability	0.775	0.803	0.765	1.000
Class 2 Probability	0.961	0.869	0.976	1.000
Class 3 Probability		0.951		

Model fit results for Energize are presented in Table 36. The addition of neither the sector nor subdivision covariate alone improved either model fit or average latent class

probabilities. However, the addition of both covariates improved model fit and average latent class probabilities, resulting in the best model fit with two latent classes. Neither covariate was a significant predictor of the slope or intercept, but adding the sector covariate to the model slightly reduced measurement and growth factor residual variances. Adding the subdivision covariate to the model separately reduced residual variances even further. The LMR LRT indicated that two latent classes were present in the data.

Table 36

Direct Report Growth Mixture Model with Covariates for Energize

	Covariates			
	None	Sector	Subdivision	Both
BIC	1421.649	1530.399	1529.853	1539.348
Free Parameters	11	14	14	17
Entropy	0.874	0.539	0.588	1.000
Class 1 N	11	58	194	10
Class 2 N	317	270	134	318
Class 1 Probability	0.787	0.866	0.882	1.000
Class 2 Probability	0.974	0.849	0.870	1.000

Model fit results for Edge are presented in Table 37. The addition of the sector covariate improved model fit and the average latent class probabilities. The addition of the subdivision covariate did not improve model fit but did improve average latent class probabilities. The addition of both covariates improved model fit and average latent class

probabilities, resulting in the best model fit with two latent classes. However, neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model reduced growth factor residual variances, but adding the subdivision covariate to the model separately did not. The LMR LRT indicated that two latent classes were present in the data.

Table 37

Direct Report Growth Mixture Model with Covariates for Edge

	Covariates			
	None	Sector	Subdivision	Both
BIC	1443.616	1455.891	1462.014	1473.431
Free Parameters	11	14	14	17
Entropy	0.891	0.893	0.698	1.000
Class 1 N	11	15	76	65
Class 2 N	317	313	252	263
Class 1 Probability	0.793	0.816	0.832	1.000
Class 2 Probability	0.978	0.981	0.940	1.000

Model fit results for Execute are presented in Table 38. The addition of the sector and subdivision covariates separately improved model fit. However, the addition of both covariates did not improve model fit or average latent class probabilities. The resulting best model was the model including the sector covariate that fit two latent classes. However, neither covariate was a significant predictor of the slope or intercept. Neither the sector nor subdivision covariate reduced measurement or growth factor residual

variances when added to the model. The LMR LRT indicated that two latent classes were present in the data.

Table 38

Direct Report Growth Mixture Model with Covariates for Execute

	Covariates			
	None	Sector	Subdivision	Both
BIC	1345.023	1357.569	1361.920	1390.202
Free Parameters	11	14	14	17
Entropy	0.943	0.953	0.946	.702
Class 1 N	9	11	9	195
Class 2 N	319	317	319	133
Class 1 Probability	0.939	0.870	0.932	.912
Class 2 Probability	0.988	0.994	0.989	.877

Model fit results for Ethics and Character are presented in Table 39. The addition of the sector covariate improved model fit. However, the addition of the subdivision covariate did not improve model fit or average latent class probabilities. The addition of both covariates improved model fit and average latent class probabilities, resulting in the best model fit with two latent classes. However, neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model did not reduce measurement or growth factor residual variances, and adding the subdivision covariate to the model separately only very slightly reduced residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 39

Direct Report Growth Mixture Model with Covariates for Ethics and Character

	Covariates			
	None	Sector	Subdivision	Both
BIC	1532.058	1555.422	1546.702	1560.887
Free Parameters	11	14	14	17
Entropy	0.707	0.979	0.709	1.000
Class 1 N	38	3	39	104
Class 2 N	290	325	290	224
Class 1 Probability	0.803	0.691	0.813	1.000
Class 2 Probability	0.936	0.997	0.935	1.000

Intercepts, slopes, and estimated means for the best fitting model for each leadership performance factor are shown in Table 40. In four out of the five cases, the model containing both covariates fit the data best. In all cases the best fitting model contained two latent classes, one small class representing between 3% and 20% of the total sample and one large class representing the remaining 97% to 80%. In comparison to the LGM mean scores that increased and decreased over time, all GMM mean scores show a steady increase or decrease over time due to the estimation of latent classes. For example, the average of all direct report ratings combined for Envision (shown in Table 28) decreased from 2001 to 2002, but increased from 2002 to 2003. By using GMM (shown in Table 40), two latent classes now become apparent for Envision, a small class (3%) that begins with a higher intercept and has a flat trajectory and a second, large class (97%) that

begins with a lower intercept and has a positive trajectory. Ratings on Energize also reveal two latent classes, one small class (3%) whose ratings increase and a large class (97%) whose ratings decrease. Ratings for both latent classes on Edge decrease, but the large class's (80%) ratings decrease more quickly after an initial higher intercept. Ratings on Execute display two divergent latent classes. The smaller class (3%) has a very low intercept, but a rapidly increasing growth trajectory. The large class (97%) has a much higher intercept and a decreasing growth trajectory. Ratings on Ethics and Character also reveal two latent classes that both exhibit negative trajectories, but slightly different intercepts.

Examining latent class membership revealed some consistency across classes. Latent class membership for Envision and Energize were exactly the same, and most of these leaders came from one sector and one subdivision. Class 1 for Execute consisted almost entirely of leaders from two sectors. Unlike the large latent class, this small class had a very highly positive trajectory on Execute.

Table 40

Best Fitting Models for Direct Report Ratings

	Model	N	Intercept	Slope	Mean 1	Mean 2	Mean 3
Envision							
Class 1	Both	11	2.673	.041	2.471	2.471	2.471
Class 2		317	2.129	.141	2.052	2.229	2.405
Energize							
Class 1	Both	11	1.856	.133	1.955	2.114	2.273
Class 2		317	2.679	-.070	2.510	2.486	2.461
Edge							
Class 1	Both	65	1.985	.121	2.375	2.365	2.355
Class 2		263	2.292	.043	2.482	2.461	2.441
Execute							
Class 1	Sector	11	1.404	.627	1.374	1.994	2.615
Class 2		317	2.693	-.029	2.672	2.638	2.604
Ethics and Character							
Class 1	Both	104	3.215	-.143	3.201	3.092	2.983
Class 2		224	2.936	-.111	3.066	2.930	2.794

Next, GMMs were run for the boss rater group. Model fit results for Envision are presented in Table 41. The addition of the sector covariate did not improve model fit or the average latent class probabilities. The addition of the subdivision covariate slightly improved model fit and average latent class probabilities. The addition of the composition covariate did not improve model fit or average latent class probabilities. The addition of both sector and subdivision covariates did not improve model fit, nor did

the addition of both subdivision and composition covariates. The addition of both sector and composition covariates improved model fit; however, one class contained only 5 observations. The addition of all three covariates also did not improve model fit or average latent class probabilities. The resulting model with the best fit was the model with the addition of the subdivision covariate. However, none of the covariates were significant predictors of the slope or intercept. Adding the sector covariate to the model reduced some of the measurement and growth factor residual variances, adding the subdivision covariate to the model separately only very slightly reduced residual variances, and adding the composition covariate did not reduce residual variances. The LMR LRT indicated that two latent classes were present in the data.

Model fit results for Energize are presented in Table 42. The addition of the sector covariate did not improve model fit or the average latent class probabilities. The addition of the subdivision covariate slightly improved model fit and average latent class probabilities. The addition of the composition covariate did not improve model fit or average latent class probabilities. The addition of both sector and subdivision covariates did not improve model fit. The addition of both sector and composition and subdivision and composition covariates did improve model fit; however, the models were not identified. The addition of all three covariates improved model fit and average latent class probabilities, resulting in the best model fit with two latent classes. However, none of the covariates were significant predictors of the slope or intercept. Adding the sector covariate to the model reduced measurement factor residual variances, adding the subdivision covariate to the model did not reduce residual variances, and adding the

composition covariate reduced growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Model fit results for Edge are presented in Table 43. The addition of the sector covariate did not improve model fit or the average latent class probabilities. The addition of the subdivision covariate also did not improve model fit or average latent class probabilities. The addition of the composition covariate slightly improved model fit and average latent class probabilities. Combinations of two covariates were tested, with the best fitting model resulting from the addition of both the subdivision and composition covariates. The addition of all three covariates resulted in a nonidentified model. None of the covariates were significant predictors of the slope or intercept. Adding the sector and composition covariates to the model separately did not reduce measurement or growth factor residual variances, but adding the subdivision covariate to the model did slightly reduce the growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Model fit results for Execute are presented in Table 44. The addition of each covariate individually improved model fit over the model with no covariates. Combinations of two covariates were tested, and the best fitting model resulted from the addition of both the sector and composition covariates. The addition of all three covariates did not improve model fit over the inclusion of only sector and composition. None of the covariates were significant predictors of the slope or intercept. Adding the sector covariate to the model reduced some of the measurement and growth factor residual variances, but adding the subdivision and composition covariates to the model

separately did not reduce residual variances. The LMR LRT indicated that two latent classes were present in the data.

Model fit results for Ethics and Character are presented in Table 45. The addition of the sector and composition covariates individually did not improve model fit over the model with no covariates. However, the addition of the subdivision covariate did improve model fit and latent class probabilities. Combinations of two covariates were tested, and improved models resulted from the addition of both the sector and subdivision and the sector and composition covariates. The addition of all three covariates did not improve model fit. Composition was the only covariate to significantly predict the intercepts ($r = -0.243$) and slopes ($r = 0.315$). Adding the sector and subdivision covariates to the model separately reduced some of the measurement factor residual variances, and adding the composition covariate reduced some of both the growth and measurement factor residual variances. The LMR LRT indicated that three latent classes were present in the data.

Table 41

Boss Growth Mixture Model with Covariates for Envision

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All
BIC	1628.479	1651.518	1644.399	1645.201	1662.892	1662.818	1661.109	1677.272
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.811	.679	.814	.810	.696	.978	.813	.729
Class 1 N	29	77	28	29	108	5	28	130
Class 2 N	302	254	303	302	223	326	303	201
Class 1 Probability	.821	.857	.830	.825	.857	.935	.833	.912
Class 2 Probability	.964	.927	.964	.963	.903	.997	.963	.927

Table 42

Boss Growth Mixture Model with Covariates for Energize

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All
BIC	1449.586	1476.960	1470.224	1465.422	1486.839	1489.390	1497.131	1492.971
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.831	.574	.576	.810	.747	.914	.940	.878
Class 1 N	21	159	122	28	97	39	5	119
Class 2 N	310	172	209	303	234	292	326	212
Class 1 Probability	.815	.897	.833	.962	.916	1.000	.682	.970
Class 2 Probability	.964	.850	.896	.806	.915	.974	.985	.958

Table 43

Boss Growth Mixture Model with Covariates for Edge

	Covariates						
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp
BIC	1768.575	1789.021	1783.670	1784.202	1794.417	1800.971	1802.283
Free Parameters	11	14	14	14	17	17	17
Entropy	.836	.701	.835	.842	.734	.675	.981
Class 1 N	20	36	24	20	98	114	77
Class 2 N	311	295	307	311	233	217	254
Class 1 Probability	.828	.714	.784	.812	.875	.867	.998
Class 2 Probability	.964	.935	.968	.967	.945	.930	.996

Table 44

Boss Growth Mixture Model with Covariates for Execute

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All
BIC	1582.653	1606.480	1592.489	1596.665	1618.608	1610.371	1612.562	1619.226
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.709	.743	.750	.855	.871	.960	.910	.767
Class 1 N	49	91	48	21	36	39	140	46
Class 2 N	282	240	243	310	295	292	191	285
Class 1 Probability	.794	.916	.817	.778	.879	1.000	.962	.802
Class 2 Probability	.937	.940	.950	.974	.966	.991	.984	.956

Table 45

Boss Growth Mixture Model with Covariates for Ethics and Character

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All
BIC	1466.161	1555.240	1479.808	1550.873	1560.768	1564.959	1562.288	1565.665
Free Parameters	11	14	20	14	17	20	17	23
Entropy	.915	.717	.921	.578	.935	.955	.586	.842
Class 1 N	101	110	101	179	15	292	121	49
Class 2 N	163	221	162	152	316	12	210	62
Class 3 N	26		23			17		220
Class 4 N	41		45					
Class 1 Probability	.999	.881	.999	.886	.975	1.000	.819	.925
Class 2 Probability	.980	.939	.985	.860	.983	.779	.907	.953
Class 3 Probability	.934		.973			.835		.930
Class 4 Probability	.766		.768					

Intercepts, slopes, and estimated means for the best fitting model for each leadership performance factor are shown in Table 46. As opposed to the self-report and direct report ratings, the best fitting models for each leadership performance factor were not as clear in the boss ratings. In two out of the five cases, the model containing both sector and composition covariates fit the data best. Also in two out of five cases, the best fitting model contained both subdivision and composition covariates. The composition covariate was included in conjunction with at least one other covariate in all best-fitting models. In four models, two latent classes were present in the data. By using GMM, two latent classes became apparent for Envision, a small class (8.5% of total sample) that begins with a higher intercept and decreases at a fast rate and a second, large class (91.5%) with a positive trajectory. Ratings for Energize reveal two moderately sized classes, one with a positive (36%) and one with a negative (64%) trajectory. Ratings for Edge reveal a small class (23%) with a negative trajectory and a large class (77%) with a near zero rate of change. Ratings on Execute reveal a small class (12%) with a rapidly increasing trajectory and a large class (88%) with a slowly decreasing trajectory. Ratings for Ethics and Character reveal three latent classes, two small classes, one (5%) with a positive and one (4%) with a negative trajectory, and a large class (91%) with a high initial value and a negative trajectory.

Examining latent class membership revealed several interesting findings. For Edge, almost all members of Class 1, which displayed a much more rapid decline in ratings than Class 2, were from one subdivision. Most of Class 1 on Execute was from one of two subdivisions, and all were from the same sector. All leaders in the third class for Ethics

and Character received consistent ratings from their bosses, and the class was filled almost entirely with leaders from one sector and from three subdivisions. The leaders in this class were the only ones to show improvement on this factor.

Table 46

Best Fitting Models for Boss Ratings

	Model	N	Intercept	Slope	Mean 1	Mean 2	Mean 3
Envision							
Class 1	Sub/Comp	28	2.549	-.567	2.564	2.000	1.436
Class 2		303	2.421	.063	2.435	2.501	2.567
Energize							
Class 1	All	119	2.041	.290	2.487	2.525	2.564
Class 2		212	2.588	-.041	2.696	2.617	2.538
Edge							
Class 1	Sub/Comp	77	2.945	-1.315	2.491	2.437	2.383
Class 2		254	2.465	-.105	2.429	2.427	2.426
Execute							
Class 1	Sector/Comp	39	2.428	.081	2.511	2.569	2.627
Class 2		292	2.858	-.103	2.727	2.704	2.680
Ethics and Character							
Class 1	Sector/Comp	292	3.223	-.155	3.192	3.082	2.971
Class 2		12	2.689	-.178	2.847	2.652	2.457
Class 3		17	2.743	.183	2.901	3.066	3.232

Next, GMMs were run for all ratings combined. Model fit results for Envision are presented in Table 47. The addition of the sector covariate did not improve model fit;

however, the addition of the subdivision covariate did improve model fit and average latent class probabilities. The addition of both covariates improved model fit and average latent class probabilities, but the best resulting model was the one including the subdivision covariate that fit two latent classes. Neither covariate was a significant predictor of the slope or intercept. Adding the sector and subdivision covariates to the model separately slightly reduced some of the measurement and growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 47

All Raters Growth Mixture Model with Covariates for Envision

	Covariates			
	None	Sector	Subdivision	Both
BIC	589.668	588.184	599.933	608.168
Free Parameters	11	14	14	17
Entropy	.841	.832	1.000	.970
Class 1 N	11	123	101	135
Class 2 N	317	205	227	193
Class 1 Probability	.707	.930	1.000	1.000
Class 2 Probability	.968	.950	1.000	.988

Model fit results for Energize are presented in Table 48. The addition of the sector covariate did not improve model fit. However, the addition of the subdivision covariate did improve model fit and average latent class probabilities. The addition of both covariates did not improve model fit, but did improve average latent class probabilities.

The best resulting model was the one including the subdivision covariate that fit two latent classes. Neither covariate was a significant predictor of the slope or intercept. Adding the sector covariates to the model reduced the measurement and growth factor residual variances, but adding the subdivision covariate did not reduce residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 48

All Raters Growth Mixture Model with Covariates for Energize

	Covariates			
	None	Sector	Subdivision	Both
BIC	561.464	578.500	568.715	577.232
Free Parameters	11	17	14	17
Entropy	.910	.601	1.000	.852
Class 1 N	5	109	101	113
Class 2 N	323	62	227	215
Class 3 N		157		
Class 1 Probability	.635	.966	1.000	.958
Class 2 Probability	.985	.693	1.000	.949
Class 3 Probability		.697		

Model fit results for Edge are presented in Table 49. The addition of both the sector and subdivision covariates individually substantially improved model fit, but the addition of both covariates did not improve model fit. The best resulting model was the one including the subdivision covariate that fit two latent classes. Neither covariate was a

significant predictor of the slope or intercept. Adding the sector and subdivision covariates to the model separately did not reduce the measurement and growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 49

All Raters Growth Mixture Model with Covariates for Edge

	Covariates			
	None	Sector	Subdivision	Both
BIC	670.791	670.652	681.890	694.624
Free Parameters	11	14	14	20
Entropy	.400	.922	1.000	.825
Class 1 N	201	231	98	186
Class 2 N	127	97	230	122
Class 3 N				20
Class 1 Probability	.811	.976	1.000	.981
Class 2 Probability	.801	.996	1.000	.851
Class 3 Probability				.699

Model fit results for Execute are presented in Table 50. The addition of both the sector and subdivision covariates individually and jointly substantially improved model fit. The best resulting model was the one including both covariates. Neither covariate was a significant predictor of the slope or intercept. Adding the sector covariate to the model did not reduce measurement and growth factor residual variances, but adding the

subdivision covariate did reduce growth factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 50

All Raters Growth Mixture Model with Covariates for Execute

	Covariates			
	None	Sector	Subdivision	Both
BIC	527.923	521.562	542.634	547.937
Free Parameters	11	14	20	17
Entropy	.731	.861	.836	.980
Class 1 N	18	219	38	89
Class 2 N	310	109	12	239
Class 3 N			135	
Class 4 N			143	
Class 1 Probability	.693	.949	.922	.995
Class 2 Probability	.942	.968	.981	.991
Class 3 Probability			.911	
Class 4 Probability			.894	

Model fit results for Ethics and Character are presented in Table 51. The addition of both the sector and subdivision covariates individually and jointly substantially improved model fit. The best resulting model was the one including the subdivision covariate only. Neither covariate was a significant predictor of the slope or intercept. Adding neither the sector nor subdivision covariates to the model separately reduced measurement or growth

factor residual variances. The LMR LRT indicated that two latent classes were present in the data.

Table 51

All Raters Growth Mixture Model with Covariates for Ethics and Character

	Covariates			
	None	Sector	Subdivision	Both
BIC	515.158	513.161	528.580	527.362
Free Parameters	11	14	14	17
Entropy	.383	.885	1.000	.954
Class 1 N	96	51	83	69
Class 2 N	232	277	245	259
Class 1 Probability	.745	.801	1.000	.946
Class 2 Probability	.823	.991	1.000	1.000

Intercepts, slopes, and estimated means for the best fitting model for each leadership performance factor for all ratings combined are shown in Table 52. In four out of the five cases, the model containing the subdivision covariate fit the data best. In all cases, the best fitting model contained two latent classes. In comparison to LGM mean scores that both increased and decreased over time, all GMM mean scores show a steady increase or decrease over time due to the models' ability to estimate latent classes. For example, the average of all ratings combined for Envision (shown in Table 28) decreased from 2001 to 2002, but increased from 2002 to 2003. By using GMM (shown in Table 52), two latent classes now become apparent for Envision, a small class (31%) that begins with a lower intercept and increases at a faster rate than a second, large class (69%). Ratings for Energize, Edge, and Execute reveal one class with a positive trajectory and one class with a negative trajectory. Finally, ratings for Ethics and Character reveal two latent classes,

one small class (25%) whose ratings decrease more slowly than a second, large class (75%) with a lower intercept value.

Examining latent class membership revealed consistencies in class membership across leadership performance factors. Class 1 for all factors contained many of the same leaders. Class 1 for Execute was slightly less consistent due to the addition of the Sector covariate, and differed from the Class 1 assignment from the other four leadership performance factors by including more leaders from one particular sector. In all cases except for Ethics and Character, Class 1 displayed a positive trajectory.

Table 52

Best Fitting Models for All Ratings

	Model	N	Intercept	Slope	Mean 1	Mean 2	Mean 3
Envision							
Class 1	Subdivision	101	2.317	.055	2.385	2.434	2.483
Class 2		227	2.504	.008	2.518	2.525	2.531
Energize							
Class 1	Subdivision	101	2.542	.027	2.510	5.535	2.561
Class 2		227	2.608	-.027	2.619	2.592	2.564
Edge							
Class 1	Subdivision	98	2.349	.031	2.404	2.437	2.470
Class 2		230	2.461	-.011	2.472	2.461	2.451
Execute							
Class 1	Both	89	2.577	-.059	2.564	2.600	2.636
Class 2		239	2.714	-.040	2.696	2.691	2.687
Ethics and Character							
Class 1	Subdivision	83	3.222	-.086	3.222	3.137	3.051
Class 2		245	3.150	-.098	3.152	3.056	2.961

Growth Mixture Modeling with Outcome Variables

Finally, the distal outcome variable consensus performance was added to the GMMs for each rater group to answer research question 7. Question 7 asked if change in leadership performance predicts promotion and/or consensus performance. There was no variability on the promotion outcome variable; thus, it was not included in these analyses. 2.4% of the leaders had consensus performance scores of 1, 70.1% of leaders had scores

of 2, and 27.4% of leaders had scores of 3. Tables 53 - 57 show the model results for self-ratings with the addition of the consensus performance outcome variable.

Adding the consensus performance variable to the best fitting model for self-report ratings of Envision did not improve model fit. The entropy value and the latent class probability for Class 1 decreased slightly while the latent class probability for Class 2 increased. Class size for both latent classes changed very little. Model fit for self-report ratings of Energize deteriorated substantially with the addition of the consensus performance variable. While models including the moderator variables had perfect fit, the addition of the consensus performance variable resulted in an ill-fitting model. The addition of the performance variable to the best fitting model for self-report ratings of Edge also did not improve model fit. Adding the performance variable to the model with the Sector covariate resulted in a reduced entropy value and reduced latent class probabilities. The model containing both moderator variables and the performance variable was not identified, so no fit indices are available. The addition of the consensus performance variable to the best fitting model for self-ratings of Execute also did not improve model fit. The entropy value and both latent class probabilities decreased in comparison to the model containing only the moderator variables. Finally, the addition of the performance variable to best fitting model for self-ratings of Ethics and Character did improve model fit. The complete model estimated two rather than three classes with one very small class and one large class. The entropy value was slightly improved and the latent class probabilities remained high.

Table 53

Self-Report Ratings Growth Mixture Model with Covariates for Envision

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1254.495	1261.666	1263.155	1275.067	1750.994
Free Parameters	11	14	14	17	23
Entropy	0.960	0.789	0.953	0.995	0.993
Class 1 N	4	123	102	71	92
Class 2 N	324	205	226	257	236
Class 1 Probability	0.775	0.921	0.996	0.992	.991
Class 2 Probability	0.994	0.940	0.978	1.000	1.000

Table 54

Self-Report Ratings Growth Mixture Model with Covariates for Energize

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1163.571	1163.073	1173.328	1188.819	1674.890
Free Parameters	11	14	14	17	23
Entropy	0.538	0.918	1.000	1.000	0.468
Class 1 N	14	57	83	88	182
Class 2 N	314	271	245	240	146
Class 1 Probability	0.640	0.980	1.000	1.000	0.836
Class 2 Probability	0.883	1.000	1.000	1.000	0.837

Table 55

Self-Report Ratings Growth Mixture Model with Covariates for Edge

	Covariates				
	None	Sector	Subdivision	Sector/Sub	Consensus/Sector
BIC	1313.228	1318.926	1327.104	1335.493	1811.271
Free Parameters	11	14	14	17	19
Entropy	0.989	1.000	0.953	0.956	.905
Class 1 N	1	96	20	6	53
Class 2 N	327	232	308	322	275
Class 1 Probability	0.998	1.000	0.979	0.849	.953
Class 2 Probability	0.986	1.000	0.990	0.994	.970

Table 56

Self-Report Ratings Growth Mixture Model with Covariates for Execute

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1210.955	1213.142	1227.641	1236.55	1700.496
Free Parameters	11	14	18	17	23
Entropy	0.744	0.820	0.829	0.984	0.913
Class 1 N	36	28	226	83	20
Class 2 N	292	300	29	245	308
Class 3 N			73		
Class 1 Probability	0.802	0.837	0.961	1.000	0.836
Class 2 Probability	0.947	0.962	0.821	0.993	0.985
Class 3 Probability			0.903		

Table 57

Self-Report Ratings Growth Mixture Model with Covariates for Ethics and Character

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1280.941	1312.17	1297.613	1289.587	1810.497
Free Parameters	11	18	14	22	23
Entropy	0.656	0.918	0.655	0.890	0.928
Class 1 N	194	14	132	25	9
Class 2 N	134	309	196	187	319
Class 3 N		5		116	
Class 1 Probability	0.907	0.821	0.894	0.945	0.860
Class 2 Probability	0.887	0.974	0.902	0.942	0.986
Class 3 Probability		0.872		0.988	

Tables 58 - 62 show the full GMM results for direct report ratings. The addition of the consensus performance variable to the best fitting model for direct report ratings of Envision did not improve model fit. In comparison to the model with both moderator variables, the model containing all moderator and distal variables had a slightly lower entropy value and latent class probabilities, and it estimated three latent classes rather than two. Adding consensus performance to the direct report ratings model for Energize slightly decreased model fit over the best fitting model that included both covariates. The entropy value and latent class probabilities were just slightly lower in the full model, but latent class sizes differed substantially. The addition of the performance outcome to the best fitting model for Edge neither improved nor decreased model fit, because in both models the fit was perfect. However, the full model did have a significantly higher BIC and estimated very different latent class sizes. Finally, the addition of the consensus performance variable to the Energize and Ethics and Character models decreased model fit substantially.

Table 58

Direct Report Ratings Growth Mixture Model with Covariates for Envision

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1513.599	1429.898	1433.556	1449.296	1938.563
Free Parameters	11	17	14	17	28
Entropy	0.812	0.851	0.855	1.000	0.931
Class 1 N	12	22	19	12	78
Class 2 N	316	17	309	316	240
Class 3 N		289			10
Class 1 Probability	0.775	0.803	0.765	1.000	1.000
Class 2 Probability	0.961	0.869	0.976	1.000	0.968
Class 3 Probability		0.951			0.782

Table 59

Direct Report Ratings Growth Mixture Model with Covariates for Energize

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1421.649	1530.399	1529.853	1539.348	2014.705
Free Parameters	11	14	14	17	23
Entropy	0.874	0.539	0.588	1.000	0.980
Class 1 N	11	58	194	10	116
Class 2 N	317	270	134	318	212
Class 1 Probability	0.787	0.866	0.882	1.000	1.000
Class 2 Probability	0.974	0.849	0.870	1.000	0.990

Table 60

Direct Report Ratings Growth Mixture Model with Covariates for Edge

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1443.616	1455.891	1462.014	1473.431	1941.968
Free Parameters	11	14	14	17	23
Entropy	0.891	0.893	0.698	1.000	1.000
Class 1 N	11	15	76	65	187
Class 2 N	317	313	252	263	141
Class 1 Probability	0.793	0.816	0.832	1.000	1.000
Class 2 Probability	0.978	0.981	0.940	1.000	1.000

Table 61

Direct Report Ratings Growth Mixture Model with Covariates for Execute

	Covariates					
	None	Sector	Subdivision	Sector/Sub	Consensus/Sector	All
BIC	1345.023	1357.569	1361.920	1390.202	1833.648	1866.197
Free Parameters	11	14	14	17	19	23
Entropy	0.943	0.953	0.946	.702	0.946	0.735
Class 1 N	9	11	9	195	11	25
Class 2 N	319	317	319	133	317	303
Class 1 Probability	0.939	0.870	0.932	.912	0.888	0.848
Class 2 Probability	0.988	0.994	0.989	.877	0.991	0.936

Table 62

Direct Report Ratings Growth Mixture Model with Covariates for Ethics and Character

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	1532.058	1555.422	1546.702	1560.887	2016.390
Free Parameters	11	14	14	17	23
Entropy	0.707	0.979	0.709	1.000	0.703
Class 1 N	38	3	39	104	45
Class 2 N	290	325	290	224	283
Class 1 Probability	0.803	0.691	0.813	1.000	0.833
Class 2 Probability	0.936	0.997	0.935	1.000	0.932

Tables 63 - 67 show the GMM results with covariates for boss ratings. Adding the consensus performance variable with all three covariates to the All Covariate (All Cov) model of boss ratings for Envision decreased model fit. The BIC increased to 2145.532 with 27 free parameters. Entropy decreased to 0.692, the size of class 1 decreased to 64, the size of class 2 increased to 267, the latent class probability of class 1 decreased to 0.825, and the latent class probability of class 2 increased to 0.933. Adding the consensus performance variable to the best fitting model, Sub/Comp, the BIC increased to 2133.240 with 23 free parameters. Entropy decreased to 0.745, the size of class 1 increased to 49, the size of class 2 decreased to 282, the latent class probability of class 1 decreased to 0.794, and the latent class probability of class 2 decreased to 0.951.

By adding the consensus performance variable to the All Cov model for Energize, the BIC increased to 1977.327 with 27 free parameters. Entropy decreased to 0.602, the size of class 1 increased to 137, the size of class 2 decreased to 194, the latent class probability for class 1 decreased to 0.871, and the latent class probability of class 2 decreased to 0.886. Adding the consensus performance variable to the All Cov model for Edge (not shown in Table 65 due to previous model non-identification), the BIC was 2297.357 with 27 free parameters. Entropy equaled 0.594, the size of class 1 equaled 70, the size of class 2 equaled 261, the latent class probability of class 1 equaled 0.774, and the latent class probability of class 2 equaled 0.902. Adding the consensus performance variable to the best fitting model, Sub/Comp, the BIC increased to 2281.843 with 23 free parameters. Entropy decreased to 0.530, the size of class 1 slightly increased to 80, the size of class 2 slightly decreased to 251, the latent class probability for class 1 decreased to 0.763, and the latent class probability for class 2 decreased to 0.874.

Adding the consensus performance variable to the All Cov model for Execute, the BIC increased to 2100.732 with 27 free parameters. Entropy decreased to 0.629, the size of class 1 increased to 164, the size of class 2 decreased to 167, the latent class probability for class 1 increased to 0.880, and the latent class probability for class 2 decreased to 0.903. Adding the consensus performance variable to the best fitting model, Sector/Comp, the BIC increased to 2082.989 with 23 free parameters. Entropy decreased to 0.571, the size of class 1 increased to 187, the size of class 2 decreased to 144, the latent class probability for class 1 decreased to 0.874, and the latent class probability of class 2 decreased to 0.876.

Adding the consensus performance variable to the All Cov model for Ethics and Character, the BIC increased to 2056.912 with 27 free parameters and two latent classes instead of three. Entropy decreased to 0.757, the size of class 1 was 95, the size of class 2 was 236, the latent class probability for class 1 was 0.900, and the latent class probability of class 2 was 0.944. Adding the consensus performance variable to the best fitting model, Sector/Comp, the BIC increased to 2045.367 with 23 free parameters and two latent classes rather than three. Entropy increased to 1.000, the size of class 1 was 39, the size of class 2 was 292, the latent class probability for class 1 was 1.000, and the latent class probability for class 2 was 1.000.

Table 63

Boss Ratings Growth Mixture Model with Covariates for Envision

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All Cov
BIC	1628.479	1651.518	1644.399	1645.201	1662.892	1662.818	1661.109	1677.272
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.811	.679	.814	.810	.696	.978	.813	.729
Class 1 N	29	77	28	29	108	5	28	130
Class 2 N	302	254	303	302	223	326	303	201
Class 1 Probability	.821	.857	.830	.825	.857	.935	.833	.912
Class 2 Probability	.964	.927	.964	.963	.903	.997	.963	.927

Table 64

Boss Ratings Growth Mixture Model with Covariates for Energize

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All Cov
BIC	1449.586	1476.960	1470.224	1465.422	1486.839	1489.390	1497.131	1492.971
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.831	.574	.576	.810	.747	.914	.940	.878
Class 1 N	21	159	122	28	97	39	5	119
Class 2 N	310	172	209	303	234	292	326	212
Class 1 Probability	.815	.897	.833	.962	.916	1.000	.682	.970
Class 2 Probability	.964	.850	.896	.806	.915	.974	.985	.958

Table 65

Boss Ratings Growth Mixture Model with Covariates for Edge

	Covariates						
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp
BIC	1768.575	1789.021	1783.670	1784.202	1794.417	1800.971	1802.283
Free Parameters	11	14	14	14	17	17	17
Entropy	.836	.701	.835	.842	.734	.675	.981
Class 1 N	20	36	24	20	98	114	77
Class 2 N	311	295	307	311	233	217	254
Class 1 Probability	.828	.714	.784	.812	.875	.867	.998
Class 2 Probability	.964	.935	.968	.967	.945	.930	.996

Table 66

Boss Ratings Growth Mixture Model with Covariates for Execute

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All Cov
BIC	1582.653	1606.480	1592.489	1596.665	1618.608	1610.371	1612.562	1619.226
Free Parameters	11	14	14	14	17	17	17	20
Entropy	.709	.743	.750	.855	.871	.960	.910	.767
Class 1 N	49	91	48	21	36	39	140	46
Class 2 N	282	240	243	310	295	292	191	285
Class 1 Probability	.794	.916	.817	.778	.879	1.000	.962	.802
Class 2 Probability	.937	.940	.950	.974	.966	.991	.984	.956

Table 67

Boss Ratings Growth Mixture Model with Covariates for Ethics and Character

	Covariates							
	None	Sector	Subdivision	Composition	Sector/Sub	Sector/Comp	Sub/Comp	All Cov
BIC	1466.161	1555.240	1479.808	1550.873	1560.768	1564.959	1562.288	1565.665
Free Parameters	11	14	20	14	17	20	17	23
Entropy	.915	.717	.921	.578	.935	.955	.586	.842
Class 1 N	101	110	101	179	15	292	121	49
Class 2 N	163	221	162	152	316	12	210	62
Class 3 N	26		23			17		220
Class 4 N	41		45					
Class 1 Probability	.999	.881	.999	.886	.975	1.000	.819	.925
Class 2 Probability	.980	.939	.985	.860	.983	.779	.907	.953
Class 3 Probability	.934		.973			.835		.930
Class 4 Probability	.766		.768					

Tables 68 - 72 show the full GMM results for all ratings combined. The addition of the consensus performance variable to the best fitting model and the model containing both covariates for ratings of Envision did not improve model fit. In comparison to the model with only the Subdivision variable, the addition of the performance variable slightly lowered the entropy value and latent class probabilities. Adding consensus performance to the best fitting model, Subdivision, and the model containing both covariates for Energize severely decreased model fit. The entropy values and latent class probabilities were significantly lower in both models and were indicative of inadequate model fit. The addition of the performance outcome to the best fitting model, Subdivision, and the model containing both covariates for Edge did not improve model fit, because fit in the Subdivision model was perfect. However, the full model containing both covariates and the outcome variable was an improvement over the model containing only the covariates. Adding the consensus performance variable to the model with both covariates for Execute did not improve model fit. The full model had decreased entropy and latent class probability values. The addition of the outcome variable to the best fitting model, Subdivision, for Ethics and Character neither increased nor decreased model fit because both models fit perfectly; however, the model including consensus performance had a higher chi-square value. Adding the performance variable to the model with both covariates; however, did slightly improve model fit for all ratings of Ethics and Character.

Table 68

All Ratings Combined Growth Mixture Model with Covariates for Envision

	Covariates					
	None	Sector	Subdivision	Sector/Sub	Consensus/Sub	All
BIC	589.668	588.184	599.933	608.168	1083.287	1091.544
Free Parameters	11	14	14	17	19	23
Entropy	.841	.832	1.000	.970	0.956	0.816
Class 1 N	11	123	101	135	24	14
Class 2 N	317	205	227	193	304	314
Class 1 Probability	.707	.930	1.000	1.000	0.928	0.856
Class 2 Probability	.968	.950	1.000	.988	0.994	0.966

Table 69

All Ratings Combined Growth Mixture Model with Covariates for Energize

	Covariates					
	None	Sector	Subdivision	Sector/Sub	Consensus/Sub	All
BIC	561.464	578.500	568.715	577.232	1044.644	1062.512
Free Parameters	11	17	14	17	19	23
Entropy	.910	.601	1.000	.852	0.390	0.539
Class 1 N	5	109	101	113	141	129
Class 2 N	323	62	227	215	187	199
Class 3 N		157				
Class 1 Probability	.635	.966	1.000	.958	0.779	0.852
Class 2 Probability	.985	.693	1.000	.949	0.818	0.862
Class 3 Probability		.697				

Table 70

All Ratings Combined Growth Mixture Model with Covariates for Edge

	Covariates					
	None	Sector	Subdivision	Sector/Sub	Consensus/Sub	All
BIC	670.791	670.652	681.890	694.624	1147.991	1164.929
Free Parameters	11	14	14	20	19	23
Entropy	.400	.922	1.000	.825	0.893	0.939
Class 1 N	201	231	98	186	35	150
Class 2 N	127	97	230	122	293	178
Class 3 N				20		
Class 1 Probability	.811	.976	1.000	.981	0.894	0.993
Class 2 Probability	.801	.996	1.000	.851	0.981	0.969
Class 3 Probability				.699		

Table 71

All Ratings Combined Growth Mixture Model with Covariates for Execute

	Covariates				
	None	Sector	Subdivision	Sector/Sub	All
BIC	527.923	521.562	542.634	547.937	1022.681
Free Parameters	11	14	20	17	23
Entropy	.731	.861	.836	.980	0.845
Class 1 N	18	219	38	89	18
Class 2 N	310	109	12	239	310
Class 3 N			135		
Class 4 N			143		
Class 1 Probability	.693	.949	.922	.995	0.774
Class 2 Probability	.942	.968	.981	.991	0.969
Class 3 Probability			.911		
Class 4 Probability			.894		

Table 72

All Ratings Combined Growth Mixture Model with Covariates for Ethics and Character

	Covariates					
	None	Sector	Subdivision	Sector/Sub	Consensus/Sub	All
BIC	515.158	513.161	528.580	527.362	998.135	1016.713
Free Parameters	11	14	14	17	19	23
Entropy	.383	.885	1.000	.954	1.000	0.978
Class 1 N	96	51	83	69	38	158
Class 2 N	232	277	245	259	290	170
Class 1 Probability	.745	.801	1.000	.946	1.000	0.986
Class 2 Probability	.823	.991	1.000	1.000	1.000	0.997

CHAPTER FIVE: DISCUSSION

The purpose of this study was to challenge traditional beliefs that performance is a stable construct and to examine the validity of multisource repeated measure leadership performance appraisals in predicting important leader outcomes. In organizations, performance feedback is often given and received annually, yet few researchers and practitioners have concerned themselves with examining the longitudinal effects of providing multisource feedback despite evidence that performance is dynamic.

Using longitudinal ratings of performance rather than a more typical cross-sectional design, this study provides evidence for the changing nature of executive leadership performance and the predictive ability of leadership performance change on overall job performance. In addition, this study sought to investigate the effects of rater group composition, rater perspective, and rater context on longitudinal performance ratings. Prior research shows that performance is dynamic (Thoreson, et al., 2004) and that raters from different organizational perspectives are attuned to different behaviors and can provide unique information in their ratings (Scullen et al., 2000). This study confirms and extends these previous findings.

Discussion of Findings by Research Question

Seven research questions were addressed by this study. The first and second questions addressed measurement equivalence across rating source groups and measurement occasions on the leadership performance construct. Previous research results have been inconsistent regarding equivalence among rating sources. Supporters of multisource feedback systems advocate for the additional information gained by using

raters from different organizational perspectives; however, rating source inequivalence might obscure a target's longitudinal performance trend if data from all rating sources is combined and limits cross-group comparisons. Several researchers have used a CFA approach to examine rating source equivalence (Cheung, 1999; Maurer, Raju & Collings, 1998; Facteau & Craig, 2001). In these studies, results indicated that ratings from sources including self, supervisors, peers, and direct reports were equivalent. The results from this study concur with previous research.

Measurement equivalence of the leadership performance measure across rater groups and time intervals was first established in this study to ensure that the same construct was measured and could be compared before evaluating individual performance trends and their resulting outcomes. A series of CFA model comparisons were conducted to confirm measurement equivalence among rating source groups and measurement occasions. Using CFA, all models indicated moderate, but acceptable, fit. Self-ratings, boss ratings, and direct report ratings across the three measurement occasions were equivalent.

Measurement equivalence was confirmed across all rater groups and all measurement occasions. The leadership performance factors were intercorrelated at relatively high levels with the exception of Ethics and Character, which demonstrated lower correlations with the other four factors across all rater groups and measurement occasions. This may be due to the fact that Edge, Execute, Envision, and Energize describe the behaviors and actions of a leader while Ethics and Character measures the traits of integrity and emotional stability. Tests of measurement equivalence concluded that the correlations among the five factors and factor loadings on the five factors did not differ significantly by rating source group or across the three measurement occasions.

Measurement equivalence of the self, boss, and direct report ratings supported combining all ratings into a composite leadership performance score. Idiosyncratic rater effects account for a very large proportion of performance rating variance, and because of the large error component, averaging across multiple raters serves to significantly reduce the effects of bias and random error (Scullen, et al., 2000). Leadership performance ratings were combined and averaged across rater groups to create composite ratings from all raters combined. Because the measurement equivalence models indicated acceptable fit, data analysis continued by examining individual trends in the leadership performance data.

The third and fourth research questions investigated the form that latent growth curves take for longitudinal leadership performance and how these curves might be dependent upon rating sources. In previous analyses with these data (Kaiser, Craig, & Kaplan, 2002), ANOVA suggested several possible trends in overall leadership performance, including an upward trend in observed means for boss ratings from 2000 to 2002. In the current study, boss ratings were the only source ratings that adequately fit a LGM. Four of the five leadership factors indicated a negative trend over the time period 2001 to 2003. In the previous ANOVA, direct report ratings showed a downward trend from 2001 to 2002. In the current study, direct report ratings failed to support an adequately fitting LGM; however, four of the five leadership performance factors showed decreasing ratings from 2001 to 2002 and increasing ratings from 2002 to 2003. Finally, self-ratings increased from 2000 to 2001, but decreased from 2001 to 2002 in the previous study. In the current study, self-ratings failed to support an adequately fitting LGM; however, three of the five leadership performance factors showed steady increases

in self-ratings over time, one factor showed steady decreases, and one factor showed an initial decrease followed by an increase.

LGM analyses were performed for each rater group on each leadership performance factor for years 2001 through 2003. Fit statistics provided evidence of adequate model fit for boss ratings, but not for self-ratings, direct report ratings, or average ratings across rater groups. Despite model fit being adequate for only the boss rater group, latent growth curves for all rater groups were remarkably similar. Ratings on Envision for all rater groups showed a decrease from 2001 to 2002, but then an increase from 2002 to 2003. Ethics and Character ratings showed a steady decrease for all rater groups. Ratings on Energize, Edge, and Execute were more mixed. These leadership factors showed a steady decrease in the boss rater group, but a steady increase in self-ratings. Ratings on Energize, Edge, and Execute from direct reports and all groups combined decreased from 2001 to 2002 and increased from 2002 to 2003. Despite the measurement equivalence shown between rating source groups, the information provided by each rater group in this study was unique, supporting the idea that obtaining ratings from different organizational perspectives adds valuable information about target behavior.

The initial decreases in leadership performance that occurred after the 2000-2001 companywide leadership intervention should not be unexpected because leadership interventions using multisource feedback have been shown to result in performance decreases for a subset of managers in previous research (Kluger & DeNisi, 1996). Some targets who receive negative feedback might become discouraged and not motivated to improve. However, other researchers suggest that performance feedback might result in an immediate decline in performance followed by a subsequent increase in that

performance once the target's new knowledge and skills have become engrained (Klein & Ziegert, 2004).

The next two research questions investigated the effects of rater composition, rater context variables (i.e., sector and subdivision), and leader outcome variables (i.e., promotion and consensus performance). The effects of rater composition were examined in both direct report and boss ratings. Before using direct report ratings in LGM analyses, a comparison was made between consistent and inconsistent direct report ratings. Measurement equivalence tests indicated that the leadership performance model was equivalent between consistent and inconsistent direct report groups. Both direct report groups' latent growth models exhibited poor model fit yet followed similar growth curvatures. The consistent direct report group tended to rate leaders more highly than the inconsistent direct report group. Leniency in the consistent direct report group might be due to a greater liking of familiar leaders or more valid evaluation of the leaders' performance. Future research should seek to confirm and extend these findings.

The effects of rater composition for boss ratings were examined using SOF LGM. The dichotomous rater composition variable was entered into the model as a moderator variable separately and in conjunction with the context covariates. The addition of the composition covariate singularly and in conjunction with the other covariates to the LGM of Envision failed to improve model fit. The best fitting model for Energize resulted from the addition of all three covariates. The best fitting model for Edge resulted from the addition of both the subdivision and composition covariates. The best fitting model for Execute resulted from the addition of both the sector and composition covariates. Finally, the best fitting model for Ethics and Character resulted from the addition of both

the sector and composition covariates. The composition covariate also significantly predicted the intercept and slope for Ethics and Character. In four out of the five leadership performance models, the addition of the composition variable in conjunction with one or both context covariates improved model fit. Therefore, it can be concluded that rater composition and rater context for bosses play important roles in detecting linear change in leader performance.

Klein and Ziegert (2004) proposed that individual differences and organizational climate have effects on leader change and development. Organizational climate may have indirect effects on leader change by influencing how leaders experience work challenges, feedback, and training. Organizational climate may have a moderating effect by influencing leader acquisition of new skills and knowledge as a result of work challenges, feedback, and training. Additionally, organizational climate may effect raters and how they provide ratings. Certain organizational units may provide rater training or be more supportive of certain leadership behaviors. Leader performance typically plays a dominant role in leadership performance ratings, but research has found that rater characteristics also have effects on performance ratings (Cardy & Dobbins, 1994). This study included several rater characteristics to examine whether these characteristics influence judgments of leadership performance. While there were no measures of organizational climate used in this study, research has found that functional units (i.e., organizational sectors and subdivisions) can have their own unique perceptions of organizational culture (Cantwell, Mullen, & Aiman-Smith, 2007).

Adding the sector and/or subdivision covariates to the GMMs for every rater group improved model fit for each leadership performance factor. Adding both covariates to

models for self-ratings resulted in the best fitting model for four of the five models. The model for Edge improved with the addition of the Sector covariate. Importantly, using GMM with self-ratings allowed the model to estimate latent classes that clarified the direction of leadership performance growth. With simple LGM, growth trajectories were positive and negative over time. Three leadership performance factors showed steady increases over time, one factor showed steady decreases, and one factor showed an initial decrease followed by an increase.

The addition of the covariates allowed the model to separate leaders into classes with similar growth trajectories that resulted in positive trajectories for all leadership performance factors. GMM estimated two latent classes for Envision, Energize, Edge, and Execute that showed different slopes and intercepts, but all slopes were in the positive direction. Three classes were estimated for Ethics and Character with one class having a very low intercept and a highly positive growth trajectory, one class had a moderate intercept and little movement over time, and the third class had a very high intercept and a large decline in ratings over time. GMM helped to clarify the direction of growth in self-ratings over time, which resulted in clear improvements in leadership performance for most individuals. Positive growth for the majority of classes on all leadership performance factors was unique to self-ratings, possibly suggesting positive rater bias (leniency). This should not be unexpected because individuals tend to view themselves positively, which might lead them to ignore, distort, rationalize, or attach less importance to negative information they may receive from others. Also, others might be reluctant to give negative feedback to a leader, which might perpetuate that leader's belief that he or she is performing better than reality (Waldman & Atwater, 1998).

Similar to the self-ratings, adding both covariates to models for direct report ratings also resulted in the best fitting model for four models. The model for Execute improved with the addition of the Sector covariate. Adding the subdivision covariate to models for all ratings combined resulted in the best fitting model for four models. The model for Execute improved most with the addition of both covariates. The addition of both the subdivision and composition covariates to the models for boss ratings resulted in the best fitting model for Envision and Edge. Adding both the sector and composition covariates resulted in the best fitting model for Execute and Ethics and Character. Finally, the addition of all three covariates resulted in the best fitting model for Energize. In all leadership performance models for each rater group the rater context variables improved model fit over the models containing no covariates. The addition of these covariates allowed the GMMs to better estimate latent classes.

Identifying latent classes that group individual leaders into clusters with different trajectories of leadership performance growth allows researchers to identify commonalities among the leaders within a latent class. In this study, comparing the latent classes for each leadership performance factor and for each rater group identified certain organizational sectors and subdivisions that employed leaders with steeper leadership growth trajectories. For example, the first latent classes identified in the self-ratings contained a small group of leaders with a positive growth trajectory. While the size of Class 1 was similar for Envision, Energize, Edge, and Execute and the trajectories were all positive, the same leaders were not always assigned to the same latent classes for each factor. The majority of leaders in Class 1 for Envision came from the Commercial Solutions sector while the majority of leaders in Class 1 for Energize came from the

Corporate sector. Class 1 for Edge represented membership from several sectors, and Class 1 for Execute represented membership mainly from two sectors, Personal Communications and Semiconductor Products. All three latent classes for Ethics and Character represented membership from a variety of sectors and subdivisions.

Examining posterior latent class membership for direct report ratings revealed greater consistency across classes. Latent class membership for Envision and Energize was exactly the same, and most of the leaders in Class 1 came from the Integrated Electronics sector and the Automotive Communications subdivision. These leaders had no change over the three years on Envision while leaders in Class 2 had positive growth. Conversely, the leaders in Class 1 for Energize had positive growth while leaders in Class 2 showed a decline in ratings. Class 1 for Execute had a highly positive trajectory and consisted almost entirely of leaders from two different sectors, Global Telecom and Personal Communications.

Examining posterior latent class assignments for boss ratings also revealed several interesting findings. Almost all of Class 1 on Execute, which was the only class to display positive growth, was from the Commercial Solutions sector and most were from one of two subdivisions, Marketing and Sales and Technology Development. Class 3 for Ethics and Character was the only class to show improvement on this factor and was filled almost entirely with leaders from the Commercial Solutions sector and from two subdivisions, Marketing and Sales and Technology Development. These results indicate that something occurred in the Commercial Solutions sector and Marketing and Sales and Technology Development subdivisions that did not occur or did not occur to the same

degree in other sectors and subdivisions that helped leaders improve over time in Execute and Ethics and Character.

Finally, examining latent class membership for all ratings combined also revealed consistencies in class membership across leadership performance factors. Class 1 for all factors contained many of the same leaders and displayed a positive trajectory for all factors except Ethics and Character. Class 2 for Ethics and Character also displayed a negative trajectory. The majority of leaders in Class 1 for all factors came from two sectors, Personal Communications and Semiconductor Products, and two subdivisions, Semiconductor Products Marketing and Global Supply.

Across all rater sources, one sector was clearly represented in latent classes with positive trajectories. The Personal Communications sector had membership in Class 1 for all leadership performance factors for all ratings combined. Personal Communications also had membership in Class 1 on ratings of Execute for both self and direct report ratings. These results suggest that something within this sector encouraged or allowed leaders to improve in leadership performance during the time period studied, especially Envision: achieving results better and faster than competitors by utilizing innovative, proven, and rigorous management practices. The Personal Communications sector is one of the key components for this telecom company. Interestingly, while the broader stock market declined from January 1, 2001 to December 31, 2003 (the Dow Jones Industrial Average declined 4% and the Nasdaq declined 28%), the stock price for the organization used in this study declined at a much lower rate than its two closest competitors (-36% versus -50% and -85%). While no firm conclusions may be drawn from this observation, improved leadership in the Personal Communications sector might

have contributed to the relative success of this organization over its competitors during this time immediately following the "dot-com collapse."

For boss ratings, the latent class identified with a positive trajectory most often was not Personal Communications, but Commercial Solutions. Leaders from Personal Communications fell into the latent class with a negative trajectory in models of boss ratings. It is differences such as these that support the use of multisource feedback rather than single source ratings of performance.

The final research question addressed how change in leadership performance relates to objective performance outcomes. In a study conducted by Deadrick and colleagues (1997), individual differences in ability and experience were significantly correlated with performance trends but only accounted for 5% of the variance in the rate of performance change. One would expect organizations to reward and promote leaders who successfully acquire and utilize leadership skills and knowledge. However, organizations often reward and promote leaders for things such as seniority or financial performance that might have little to do with leader development (Klien & Ziegert, 2004).

Day and Lance (2004) believed that the measurement of change must allow for the identification of subgroups of individuals who share similar change patterns. Growth mixture modeling has the capability of both identifying subgroups of leaders with different leadership performance trajectories and using this longitudinal change information to predict outcomes such as job performance. In the final stage of the current research, the distal outcome variable consensus performance was added to the GMMs for each rater group. Unfortunately, the promotion outcome could not be included in these analyses because it had no variability.

Adding the consensus performance variable to the best fitting model of boss ratings and self-report ratings for Envision, Energize, Edge, and Execute did not improve model fit. However, the addition of the performance variable to the models for Ethics and Character did improve model fit. The addition of the consensus performance variable to the best fitting model for all ratings combined for Envision, Energize, Edge, and Execute did not improve model fit. The addition of the outcome variable to the best fitting model for Ethics and Character neither increased nor decreased model fit because both models fit perfectly; however, adding the performance variable to the model with both covariates did slightly improve model fit.

These results indicate that adding the consensus performance outcome variable to the models did not improve model fit for any of the leadership performance factors except Ethics and Character. Adding consensus performance as an outcome variable to the models for Ethics and Character improved model fit for self-ratings, boss ratings, and all ratings combined. In this study, leaders who showed growth in their honesty and ability to put organizational success before personal achievement received higher consensus performance scores, making this a very important leadership performance factor on which to emphasize development.

This finding is contrary to results found by Cardy and Selvarajan (2006). Using vignettes, these researchers found that ethical behavior judgments made by students were more positively biased when the target had successful rather than unsuccessful performance outcomes. However, the study design was such that ethical behaviors and performance outcomes were observed at the same point in time. The current study used a longitudinal design in which ethical performance was measured years before performance

outcomes were assessed. Therefore, one may conclude from the current study that ethical behavior led to successful job performance, not that successful performance leads to the perception of greater ethics as suggested by Cardy and Selvarajan.

Garcia-Zamor (2003) argues that ethics can be developed in individuals and cites organizations that have provided ethics awareness training to the benefit of the employees and the organization. For example, GE Industrial Systems was able to increase job satisfaction and productivity by providing ethical development courses. Garcia-Zamor also argues that the organization itself plays an important role in fostering an ethical climate.

The Ethics and Character factor used in this study reflected a leader's honesty and willingness to place the organization's goals first. This is consistent with past factor analytic research that identified two factors for ethics: integrity in dealing with others and self-serving behavior (Morgan, 1989). Morgan found integrity to be strongly related to trust, which is an important determinant of leader-follower relationship quality in leader-member exchange (LMX) theory. Integrity is also important from an implicit leadership theory standpoint. Leaders need to be perceived as displaying a level of integrity that is consistent with their followers' expectations to be most effective (Craig & Gustafson, 1998). Individuals who are perceived to be fair, believable, and honest are more likely to be described as leader-like (Lord, Foti, & DeVader, 1984). Past research has shown that integrity and ethical behavior are important characteristics in successful managers (Posner & Schmidt, 1984; Mortensen, Smith, & Cavanagh, 1989). This study reinforces these findings by showing that leaders with positive development on Ethics and Character had higher consensus performance scores.

Implications

In addition to extending the leadership and measurement literatures, the findings of this study add significantly to practice. Implications for researchers and practitioners include the usefulness of multisource feedback and the correct way to evaluate such ratings, the nature of leadership performance over time, the influence of rater context and rater group composition in leadership performance change, and the predictive ability of longitudinal change in leadership performance.

The validity of multisource feedback instruments has important implications for practical use. Supporters argue that these instruments are advantageous because of the additional information gained by having multiple perspectives on a target's performance. Supporters state that cross-source rating differences are not biases but represent true differences in perspective and opportunity to observe the target's behavior (Woehr, Sheehan, & Bennett, 2005). However, others warn that rating source inequivalence can obscure true performance ratings if different rater groups conceptualize performance differently, making comparisons and composite scores created across groups inaccurate (Bollen, 1989). This study demonstrated that it is possible and prudent to address both sides of this debate. In order to use ratings from multiple rater groups, researchers and practitioners must first establish the measurement equivalence of the performance construct across rater groups. Only then may comparisons be made and composite scores be created across groups.

Because self-ratings, boss ratings, and direct report ratings across the three measurement occasions in this study were equivalent, comparisons on the leadership performance factors were possible and supported combining all ratings into a composite

leadership performance score. Utilizing different rater groups did add to the richness of the leadership performance data. For example, boss ratings showed a consistent decline on ratings of Energize, Edge, and Execute while self-ratings on the same three factors displayed a consistent incline. When all ratings were combined, this difference was no longer apparent and would have remained hidden if multisource feedback had not been obtained. In addition, the results of this study are consistent with findings from previous research (Mount, 1984) that showed self-ratings tended to be higher than boss ratings, and boss ratings tended to be higher than direct report ratings.

Researchers have begun to acknowledge that performance changes over time. The results of this study concur with earlier findings about the dynamic nature of performance (Thoreson, et al., 2004). Longitudinal change in performance was found for every leadership performance factor in ratings from every rater group, and latent growth curves for all rater groups were remarkably similar. However, boss ratings showed the most consistent longitudinal change with four of the five leadership factors indicating a negative trend.

This study provides evidence that leadership performance is dynamic. Future leadership performance research should refrain from using a cross-sectional design and instead use a longitudinal approach. In addition, human resource managers should require recurring employee performance appraisals from consistent raters so that change in performance over time can be evaluated and reinforced.

Organizational climate might affect leader change by influencing how leaders experience work challenges, feedback, and training (Klein & Ziegert, 2004). Organizational climate might also affect how individuals rate performance based on

training and feedback or on work context. This study examined the effects of organizational sector and subdivision on leadership performance ratings and found that adding sector and/or subdivision covariates to the models for every leadership performance factor improved model fit for each rating source group. Importantly, using growth mixture modeling with the rater context covariates allowed the estimation of latent classes that clarified the direction of leadership performance growth.

Identifying latent classes allows researchers to document similarities among the leaders within a latent class. In this study, comparing the latent classes for each leadership performance factor and for each rater group identified a small number of organizational sectors and subdivisions with employees providing ratings showing steeper leadership growth trajectories. This information has important practical implications. Managers can identify sectors that show substantial improvement in performance and use these sectors as models for other sectors. Or, if certain sectors have a large number of leaders with negative trajectories, management might be able to intervene and assist the sector to improve.

Another important rating issue to consider in addition to organizational perspective and work context is whether raters consistently rate the same target year after year. This study's results indicated that direct reports who consistently rated the same leader tended to rate those leaders more highly than the direct reports who rated different leaders. This leniency might be due to greater familiarity or liking of the leader or a more valid evaluation of his or her performance. Adding the composition moderator variable to the boss rating model improved model fit for four of the five leadership performance models. The composition covariate also significantly predicted the intercept and slope for Ethics

and Character ratings. These results indicate that obtaining leadership performance ratings from the same boss for the same leader year after year plays an important role in detecting linear change in leader performance. Professionals in charge of administering an organization's annual performance appraisal system should make every effort possible to have the same employees provide ratings to the same target year after year.

Multisource feedback systems are often used for training and development purposes within organizations (Avolio, Sosik, Jung, & Berson, 2003). Leaders receive feedback from their bosses, peers, and direct reports and are hopefully given opportunities to improve their performance based on this feedback and how it compares to their own self-report ratings. Over time, one would expect a leader's performance to improve. Therefore, it was expected in this study that change in leadership performance over time would predict a leader's promotion and consensus performance outcomes. Unfortunately, there was no variability on the promotion variable; thus, this question could not be completely answered. Managers from this company tend to not differentiate among targets as much as other managers do in industry (Kaiser, et al., 2002). However, one would expect positive change in leadership performance to be considered in promotion decisions; thus, tracking longitudinal performance could have substantial benefits for succession planning. Surprisingly, only improvement on Ethics and Character predicted consensus performance.

It was also expected that utilizing consistent raters across all measurement intervals would prove critical to understanding the true change in leadership performance and its resulting predictive ability. This was only true for boss ratings of Ethics and Character. The model containing the sector and composition covariates and the consensus

performance variable resulted in a perfect fit. This model consisted of one small class with a positive slope and one large class with a negative slope. The small class consisted almost entirely of leaders with raters from one sector who had the same boss rate them annually. Past research has shown that integrity and ethical behavior are important characteristics in successful managers (Posner & Schmidt, 1984; Mortensen, Smith, & Cavanagh, 1989). This study confirms these findings by showing that leaders with positive development on Ethics and Character had higher consensus performance scores.

The results of this research have significant implications for executive succession planning and performance appraisal and management systems. Examining and tracking leadership performance growth gives management the power to estimate the rate of leadership development and how this rate can affect executive succession planning and training and development programs. Management may use this information, for example, to estimate the length of time needed for a new hire to reach the performance level of a retiring executive. In addition, examining leadership performance change, especially in Ethics and Character, and its ability to predict consensus performance validates the performance appraisal system. Although causality cannot be firmly established by the design of the current study, leaders who take part in an annual multisource feedback system and work to improve their leadership performance in Ethics and Character might well expect higher consensus performance scores.

Limitations and Future Research

While this study adds to the literature and provides practical applications for the findings, there were several limitations. The measurement instrument used in this study was not a true 360-degree performance evaluation because it did not include peer ratings.

The conclusions drawn from this study may only be applicable to self, boss, and direct report ratings. Future research should strive to include peer ratings.

This study also only examined annual leadership performance ratings for three years. This restricted the LGM analyses such that quadratic growth could not be examined. Four time points are necessary to estimate quadratic growth. More frequent ratings, such as quarterly performance ratings, or a longer timeline might have improved precision in growth curve estimation.

A major limitation of this study was that the promotion variable could not be included in the analyses because it displayed no variation. No leaders were promoted or demoted within the three-year timeline. Future research should examine more distal outcome variables. Three years might have been too short of a time period to expect promotion changes.

Finally, many leaders exhibited little change over the three years. This may be due to the collection of ratings during a maintenance stage of performance. The leaders participating in this study were in the highest levels of the organization, and were likely not learning and applying new skills to their jobs or facing any other changes that might place them in a transitory stage which typically creates large fluctuations in performance (Kanfer & Ackerman, 1989; Murphy, 1989; Ployhart & Hakel, 1998). In addition, data for this study were from only one company, and past work with this organization has shown that managers do not differentiate among rating targets as well as similar companies in their industry (Kaiser, Craig, & Kaplan, 2002). Future research would benefit from including data from multiple organizations from a variety of industries.

In sum, this study supports the use of annual multisource feedback systems with consistent raters. Future research should further investigate the relationship between the ethical conduct of business and ratings of job performance. The current study found that growth on this factor predicted higher ratings of consensus performance, a score that included subjective as well as objective measures of overall job performance. A greater understanding of how change on these important leadership performance factors is important for both individual leader development and organizational policies and practices.

References

Atwater, L.E., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings of leadership. *Personnel Psychology, 48*, 35-60.

Atwater, L.E., Waldman, D., Atwater, D., & Cartier, J. (2000). An upward feedback field experiment: Supervisors' cynicism, follow-up and commitment to direct reports. *Personnel Psychology, 53*, 275-297.

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

Avolio, B.J., Sosik, J.J., Jung, D.I., & Berson, Y. (2003). Leadership models, methods, and applications. In Daniel R. Ilgen, Walter C. Borman, & Richard J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology: Vol. 12.* New York, NY: John Wiley & Sons.

Avolio, B.J., Waldman, D.A., & McDaniel, M.A. (1990). Age and work performance in nonmanagerial jobs: The effects of experience and occupational type. *Academy of Management Journal, 33*, 407-422.

Bass, B.M. (1962). Further evidence of the dynamic nature of criteria. *Personnel Psychology, 15*, 93-97.

Bass, B.M. (1985). *Leadership and performance beyond expectations*. New York: Free Press.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., Mackenzie, M.B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Borman, W.C. (1997). 360 ratings: an analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review, 7*, 299-315.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications and Programming*. Mahwah, NE: Lawrence Erlbaum.

Cantwell, A.R., Mullen, T.R., & Aiman Smith, L. (2007). Subcultures tell the story: perceptions of innovation-capacity culture. Poster presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New York.

Chan, D. (1998). The conceptualization and analysis of change over time: an integrative approach incorporating longitudinal means and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*, 421-483.

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

Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling 9(2),* 233-255.

Cohen, E., & Tichy, N. (1997). How leaders develop leaders. *Training and Development Journal*, 58-71.

Costa, P. T., & McCrae, R. R. (1992). *NEO-PI-R Professional Manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Craig, S.B., & Gustafson, S.B. (1998). Perceived Leader Integrity Scale: an instrument for assessing employee perceptions of leader integrity. *Leadership Quarterly, 9(2),* 127-145.

Craig, S.B., & Kaiser, R. (2003). Applying item response theory to multisource performance ratings: what are the consequences of violating the independent observations assumption? *Organizational Research Methods, 6*, 44-60.

Day, D.V., & Lance, C.E. (2004). Understanding the development of leadership complexity through latent growth modeling. In David V. Day, Stephen J. Zaccaro, & Stanley M. Halpin (Eds.), *Leader development for transforming organizations: Growing leaders for tomorrow.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Deadrick, D.L., Bennett, N., & Russell, C.J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of management, 23*, 745-757.

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

Garcia-Zamor, J. (2003). Workplace spirituality and organizational performance. *Public Administration Review, 63*, 355-363.

Ghiselli, E.E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40*, 1-4.

Ghiselli, E.E., & Haire, M. (1960). The validation of selection tests in the light of the dynamic nature of criteria. *Personnel Psychology, 13*, 225-231.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*, 133-157.

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

Higgins, T.E. (1997). Beyond pleasure and pain. *American Psychologist, 52*, 1280-1300.

Hofmann, D.H., Jacobs, R., & Baratta, J.E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology, 72*, 457-462.

Hofmann, D.H., Jacobs, R., & Gerras, S.J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77*, 185-195.

Ireland, R.D., & Hitt, M.A. (1999). Achieving and maintaining strategic competitiveness in the twenty-first century: The role of strategic leadership. *Academy of management Executive, 13*, 43-57.

Kaiser, R., Craig, S.B., & Kaplan, R. (2002). More analysis of the 4 e's data.

Kanfer, R., & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657-690.

Kaplan, R.E., & Kaiser, R.B. (2003). Developing versatile leadership. *MIT Sloan Management Review, 44(4)*, 19-26.

Klein, K.J., & Ziegert, J.C. (2004). Leader development and change over time: A conceptual integration and exploration of research challenges. In David V. Day, Stephen J. Zaccaro, & Stanley M. Halpin (Eds.), *Leader development for transforming organizations: Growing leaders for tomorrow.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kluger, A.N., & DeNisi, A. (1996). The effects of feedback on performance: A historical review, a meta-analysis, and a preliminary feedback intervention. *Psychological Bulletin, 119*, 254-284.

Lance, C.E., Vandenberg, R.J., & Self, R.M. (2000). Latent growth models of individual change: the case of newcomer adjustment. *Organizational Behavior and Human Decision Processes, 83*, 107-140.

Lance, C.E., & Woehr, D.J. (1986). Statistical control of halo: a clarification from two models of the performance appraisal process. *Journal of Applied Psychology, 71*, 679-685.

Lo, Y., Mendell, N.R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

Lord, R.G, Foti, R.J., & De Vader, C.L. (1984). A test of leadership categorization theory: internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance, 34*, 343-378.

Marshall-Mies, J.C., Fleishman, E.A., Martin, J.A., Zaccaro, S.J., Baughman, W.A., & McGee, M.L. (2000). Development and evaluation of cognitive and metacognitive measures for predicting leadership potential. *Leadership Quarterly, 11(1)*, 135-153.

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

Morgan, R.B. (1989). Reliability and validity of a factor analytically derived measure of leadership behavior and characteristics. *Educational and Psychological Measurement, 49*, 911-919.

Mortensen, R.A., Smith, J.E., & Cavanagh, G.F. (1989). The importance of ethics to job performance: An empirical investigation of managers' perceptions. *Journal of Business Ethics, 8*, 253-260.

Motowidlo, S.J. (2003). Job performance. In Daniel R. Ilgen, Walter C. Borman, & Richard J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology: Vol. 12*. New York, NY: John Wiley & Sons.

Mount, M.K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37*, 687-702.

Mullen, T.R., Kroustalis, C., Meade, A.W., & Surface, E.A. (2006). Assessing change in perceived organizational support due to training. Poster presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas.

Murphy, K.R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2*, 183-200.

Muthén, B.O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29(1)*, 81-117.

Muthén, B.O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.

Northouse, P.G. (2004). *Leadership: Theory and practice*. Thousand Oaks, California:

Sage Publications.

Ployhart, R.E., & Hakel, M.D. (1998). The substantive nature of performance variability: predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51*, 859-901.

Podsakoff, P.M., MacKenzie, S.B., Moorman, R.H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *Leadership Quarterly, 1*, 107-142.

Posner, B.Z., & Schmidt, W.H. (1984). Values and the American manager: An update. *California Management Review, 26*, 202-216.

Reise, S. P., Widaman, K. E. & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114 (3)*, 552 – 566.

Salam, S., Cox, J.F., Sims, H.P. Jr. (1997). In the eye of the beholder: How leadership relates to 360-degree performance ratings. *Group and Organization Management, 22*, 185-209.

Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432-439.

Schmitt, N., Cortina, J.M., Ingerick, M.J., & Wiechmann, D. (2003). Personnel selection and employee performance. In Daniel R. Ilgen, Walter C. Borman, & Richard J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology: Vol. 12*. New York, NY: John Wiley & Sons.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85(6)*, 956-970.

Stogdill, R.M. (1974). *Handbook of leadership: A survey of theory and research.* New York: Free Press.

Thoresen, C.J., Bradley, J.C., Bliese, P.D., & Thoresen, J.D. (2004). The big five personality traits and individual job performance growth trajectories in maintenance and transitional job stages. *Journal of Applied Psychology, 89(5)*, 835-853.

Uggerslev, K.L., & Sulsky, L.M. (2002). Presentation modality and indirect performance information: Effects on ratings, reactions, and memory. *Journal of Applied Psychology, 87*, 940-950.

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

Waldman, D.A., & Atwater, L.E. (1998). *The power of 360° feedback: How to leverage performance evaluations for top productivity.* Houston, TX: Gulf Publishing Company.

Wherry, R.J., Sr., & Bartlett, C.J. (1982). The control of bias in ratings: a theory of rating. *Personnel Psychology, 35*, 521-551.

Woehr, D.J., Sheehan, M.K., & Bennett, W. (2005). Assessing measurement equivalence across rating sources: a multitrait-multirater approach. *Journal of Applied Psychology, 90*, 592-600.