A LARGE-SCALE, LONGITUDINAL INVESTIGATION OF THE ANTECEDENTS, MODERATORS, AND CONSEQUENCES OF PERFORMANCE IMPROVEMENT FOLLOWING MULTISOURCE FEEDBACK

by

SEAN PATRICK BALDWIN

(Under the Direction of Brian J. Hoffman)

ABSTRACT

Despite near ubiquitous use among Fortune 2000 organizations (Ewen & Edwards, 2001), the effects of multisource feedback on subsequent performance improvement have rarely been examined outside of pretest-posttest examinations. In the current study we conduct a large-scale (N=5,128 ratees) longitudinal investigation that spans three years and four feedback administrations, testing a theoretical model of the antecedents (managerial experience, developmental activities, initial status), moderators (performance dimension, rater source), and consequences (promotion rate) of performance change following multisource feedback. Results show substantially weaker estimates of performance change than previous reviews (Smither, et al., 2005), however, much stronger rates of change for identifiable subgroups. Specifically, novice and initially weak performing employees demonstrated the strongest rate of improvement.

INDEX WORDS: Multisource feedback, 360-degree feedback, performance improvement, leader development, rater source

A LARGE-SCALE, LONGITUDINAL INVESTIGATION OF THE ANTECEDENTS, MODERATORS, AND CONSEQUENCES OF PERFORMANCE IMPROVEMENT FOLLOWING MULTISOURCE FEEDBACK

by

SEAN PATRICKBALDWIN

Bachelor of Science, University of Georgia, 2007

Master of Science, University of Georgia, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PSYCHOLOGY

ATHENS, GEORGIA

2012

A LARGE-SCALE, LONGITUDINAL INVESTIGATION OF THE ANTECEDENTS, MODERATORS, AND CONSEQUENCES OF PERFORMANCE IMPROVEMENT FOLLOWING MULTISOURCE FEEDBACK

by

SEAN PATRICK BALDWIN

Major Professor: Brian J. Hoffman

Committee: Karl Kuhnert
Charles Lance

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

I dedicate this paper to Brian Hoffman, who provided guidance when I was distraught, evidence when I was doubtful and insight when I needed direction. You introduced me to the field of IO Psychology and taught me (by example) how to relentlessly work towards personal and professional goals. Undoubtedly, I would not be where I am without your mentorship.

Dude- Thank you very much.

ACKNOWLEDGEMENTS

I would like to acknowledge Calene, my loving wife, who has been a constant source of support, calm, and understanding regardless of my antics. The last five and half years have been trying, but throughout it she acted as an ever-present reminder of all the things more important than work, goals, and degrees. We did it Buccia!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Multisource feedback (MSF) systems incorporate raters from traditional (i.e., supervisors) as well as non-traditional sources (e.g. self, peers, direct reports, and customers) to appraise the performance of a target ratee. By considering unique perspectives of performance not assessed via traditional, single source tools, the primary benefit of MSF is proposed to stem from the confluence of unique perspectives, resulting in an increase in performance relevant feedback. MSF has been proposed to be a valuable tool to a cross section of management functions, providing (a) managers with a valid assessment for developmental, strategic, and administrative purposes, (b) employees with a voice in the decision-making process and developmental feedback from face-valid sources, and (c) the organization as a whole with richer performance information to increase employee motivation and enhance quality control for promotions (Edwards & Ewen, 1996). Given these proposed benefits of MSF, their widespread popularity should come as no surprise with over 95% of Fortune 2000 companies already using some form of MSF in their practice (Ewen & Edwards, 2001).

Although feedback is widely believed to play a key role in individual and organizational improvement, both the broader feedback literature and the MSF literature suggests inconsistent but generally weak and positive effects of MSF on performance improvement (Kluger & DeNisi, 1996; Smither, London, & Reilly, 2005). Kluger and DeNisi's review showed that over a third of the feedback interventions actually resulted

in a *decrease* in subsequent performance (Kluger & DeNisi, 1996). On a similar note, Smither, et al. (2005) summarized investigations of MSF performance improvement across two time points and found a weak, positive trend with substantial variability about the estimate. To the extent that not all feedback recipients improve their performance following a MSF administration, it is important to identify the individual differences and situational characteristics that most likely result in performance improvement (Smither et al., 2005).

To further complicate interpretation of these effects, there are a variety of methodological issues associated with the analysis of change that make clear conclusions difficult in this literature (Day, 2011). Although several studies have examined MSF's influence on performance improvement, the vast majority have utilized research designs with only two repeated measurements (e.g. Hazucha, Hezlett, & Schneider, 1993; Hegarty, 1974; Johnson & Ferstl, 1999; Seifert, Yukl, & McDonald, 2003; Smither, London, Vasilopoulos, & Reilly, 1995). Yet, such approaches are limited in assessing change over time (Day, 2011). Specifically, these investigations assume linear change in the focal variable, confound substantive change with measurement error, and do not allow for the testing of models of the antecedents and consequences of change (Ployhart & Vandenberg, 2010). With only three exceptions (Dai, De Meuse, & Peterson, 2010; Reilly, Smither, & Vasilopoulos, 1996; Walker & Smither, 1999), all investigations of performance change following MSF fall into this category and thus, have either been unable to examine or have potentially arrived at inaccurate conclusions regarding: (a) the existence of performance change following MSF, (b) the growth trajectory of performance change, and (c) antecedents, consequences, and moderators of performance

change. Given these limitations, it is important to investigate performance change following MSF using at least three repeated measures and modern longitudinal modeling techniques.

The present study extends the MSF literature by conducting a large scale study (N=5,128 target ratees, N~33,000 total raters per time point) using latent growth modeling to examine antecedents, moderators, and an outcome of MSF performance change over three administrations during a four year time span (see Figure 1). By incorporating modern statistical techniques to understand performance change, this study will help to clarify four primary questions regarding change in performance following MSF. First, does MSF result in performance change after accounting for measurement error? Second, assuming there is evidence for change, is the pattern of change linear or non-linear over successive MSF administrations? Third, what role do characteristics of the performance rating, including the rater source (self vs. peer vs. supervisor) and the performance dimension (technical/administrative vs. interpersonal performance) have on change following MSF? Finally, what are the antecedents (initial status, managerial experience, and participation in a tuition assistance program) and consequences (promotion rate) of change in MSF? In short, this study applied state-of-the-art longitudinal modeling techniques to re-evaluate the role of MSF in facilitating performance change and the antecedents, moderators, and consequences of change in performance over a four year interval.

CHAPTER 2

LITERATURE REVIEW AND HYPOTHESES

Change Assessment in Leader Development

Day (2011) noted that researchers "*must* measure change, especially when the focus is on development, regardless of whether it is child development or leader development" (p.561; original emphasis). This is especially important in leader development, which is viewed as a long-term process (Day, 2011). To date, however, relatively few examinations of leader development have utilized the longitudinal research designs, methods, and/or analyses capable of assessing change over time (Day, 2011). Rather, the overwhelming majority of studies investigating the influence of leadership development on performance change have used pretest-posttest designs to measure change in a focal variable and/or not taken advantage of modern methodological techniques for examining performance (Day, 2011). In this respect, much of the MSF literature (cf. Antonioni, 1995; Atwater, Roush, & Fischthal, 1995; Atwater, Waldman, Atwater, & Cartier, 2000; Bailey & Austin, 2006; Bailey & Fletcher, 2002; Hazucha, et al., 1993; Hegarty, 1974; Heslin & Latham, 2004; Johnson & Ferstl, 1999; Nemeroff & Cosentino, 1979; Rosti Jr. & Shipper, 1998; Seifert, et al., 2003; Smither, et al., 1995;Waldman & Atwater, 2001) is consistent with the broader trend in the leader development literature. Several authors (Day, 2011; Ployhart & Vandenberg, 2010; Singer & Willett, 2003), however, have noted disadvantages to these research methods when assessing change in a focal variable. Below we briefly discuss disadvantages of

using: (a) two time points to assess performance change as well as (b) traditional approaches that do not account for measurement error. Subsequently, we address the advantages of latent growth modeling to understanding performance change following MSF.

With regard to assessing change in a focal variable, many have noted that while two waves of data are better than one, they are not that much better (Day, 2011; Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003). That is, "two data points provide minimal information about individual change and also constrains the estimation of change to a linear trend" (Day, 2011, p.563). Stated another way, it is impossible to characterize nonlinear growth trajectories of individuals and groups over time with these research designs (Rogosa, 1995). This can lead to incomplete or potentially incorrect conclusions (Ployhart & Vandenberg, 2010). Recognizing these shortcomings, a third repeated measurement of the focal variable is oftentimes the defining criterion of "true-longitudinal research," while studies only utilizing two repeated measures are categorized as "quasi-longitudinal" (Day, 2011; Singer & Willett, 2003).

To our knowledge, only three true-longitudinal investigations of MSF on performance change have been conducted (Dai, et al., 2010; Reilly, et al., 1996; Walker & Smither, 1999), and these have had relatively small sample sizes with as few as 16 managers participating in later assessments. Their combined sample size of 422 seems surprisingly minimal given the wide-spread use and popularity of MSF in feedback settings. Further, the results regarding the pattern of performance improvement following multiple administrations differed across the three true-longitudinal studies with some findings showing performance change to plateau or even disappear as ratings near

organizational standards, and others exhibiting less predictable growth. Thus, although some knowledge has accumulated, the pattern of the change and the stability of the change following MSF is yet unknown.

In addition, research using true-longitudinal designs has not accounted for measurement error when analyzing change following MSF. This issue is not specific to true-longitudinal designs; yet, we are unaware of any quasi-longitudinal research that additionally accounts for measurement error. Thus, existing estimates of MSF's influence on subsequent performance change potentially confound change with measurement error (Ployhart & Vandenberg, 2010; Rogosa, et al., 1982; Singer & Willett, 2003). That is, measurement error can bias the estimated effect size of change by suppressing scores at one time point but inflating them at the other. Through the lens of classical test theory, all psychological measures of unobservable constructs are to some degree imperfect, and therefore, studies failing to recognize and model measurement error may capitalize on chance in estimating model parameters (Crocker & Algina, 1986). Next, Vandenberg and Lance (2000) suggested that any repeated measures assessment should test for measurement invariance across time to ensure that study participants' conceptualizations are comparable across time points before testing the focal hypotheses. Again, all existing studies of MSF performance change over time have to our knowledge failed to test this assumption. To this point, one of the three true-longitudinal studies used different performance measures across time points in their study because only 12 items were consistently used across all time points (Walker & Smither, 1999). It is possible then that observed changes in performance over time were indicative of perceptual changes in the rater or actual changes in the measure rather than "true" changes in ratee behavior.

Together, the extant literature on performance change following MSF is limited in understanding change. The vast majority can only specify linear change over time, and neither true nor quasi-longitudinal studies have accounted for measurement error or substantiated measurement invariance. Further, the few true-longitudinal investigations support varying growth trajectories of change following MSF (Dai, et al., 2010; Reilly, et al., 1996; Walker & Smither, 1999), making it difficult to judge the continued value of MSF to organizations. Given the fairly small effect sizes reported meta-analytically (Smither, et al., 2005) and empirically (Reilly, et al., 1996; Walker & Smither, 1999), it is possible that previous results of quasi and true-longitudinal studies were statistical artifacts rather than evidence supporting MSF. Although extant MSF literature has taken steps toward understanding the role of MSF in performance change, the resulting parameter estimates can only be characterized as weak, variable, and potentially confounded. Given these limitations of past research, it is clear that more attention is needed to examine the role of MSF in performance change.

Latent Growth Modeling

Before proceeding, it is necessary to briefly discuss latent growth modeling (LGM) in order to clarify key terms and advantages of this approach to investigating change in the context of MSF. LGM is a latent variable approach used to examine change in one or more focal variables over at least three points in time (Bentein, Vandenberghe, Vandenberg, & Stinglhamber, 2005; Bollen & Curran, 2006; Chan, 1998; Lance, Vandenberg, & Self, 2000; Meredith & Tisak, 1990). This approach models two latent variables: initial status and change. Initial status represents the starting point of the focal variable (in this case: the initial MSF ratings from each source). Change is modeled as the

rate of change observed in the focal variable over time (in this case: change in MSF ratings over the three time periods). This rate of performance change can be visually represented as a growth trajectory for individuals, as well as the sample as a whole, and also has the potential to exhibit linear or nonlinear change over time. The mean and variability of each latent variable are also estimated. Positive mean estimates for the initial status and change latent variables would suggest a positive starting value for the focal variable and a positive increase in that variable over time. Positive variability estimates for the initial status and change latent variables would suggest that within the sample there was variability about the starting values for the focal variable and also variability about the rates of change over time.

LGM offers several benefits when modeling change over time (Lance, et al., 2000; Lance, Meade, & Williamson, 2000; Singer & Willett, 1994; Vandenberg & Self, 1993). First, initial status and change latent variables in LGMs can be used as predictors or outcomes of other variables of interest; thus, utilizing LGM permits examinations of moderators, antecedents, and consequences of performance change. Second, in contrast to other popular methods of assessing change in the MSF literature (e.g. mean differences, repeated measures regression analysis, hierarchical linear modeling), second order factor latent growth modeling (SLGM) allows for the parameterization of measurement error. This variation allows change in the focal variable to be measured at the true-score level of analysis and thus, interpretations of change over time are unencumbered by measurement error. Third, unlike methodologies using only two time points which assume change to be linear, LGM permits nonlinear growth trajectories to be modeled (Lance, et al., 2000). Finally, nonlinear LGMs can be tested via nested comparisons with

linear models in order to identify the best fitting, most parsimonious model of change for the focal variable.

MSF: Change Over Time

As noted above, an SLGM is an analytical approach that has never been applied to performance change following MSF. Thus, although the MSF literature consistently suggests a weak yet positive effect of MSF on subsequent performance change (Dai, et al., 2010; Reilly, et al., 1996; Smither, et al., 2005; Walker & Smither, 1999), these effects have yet to be substantiated after accounting for measurement error. However, the theoretical propositions of Control Theory (Carver & Scheier, 1982) suggest that managerial performance will change following MSF administrations. In the following, we discuss the evidentiary basis underlying feedback in organizations with an overarching focus on changes in performance following feedback from MSF systems.

First, Control Theory proposes that all behavioral actions and reactions can be conceptualized in terms of a feedback loop (Carver & Scheier, 1982). In the context of MSF initiating behavioral change in feedback recipients, Control Theory suggests that ratees compare their performance ratings with a pre-specified criterion/referent. If the ratees perceive a difference between the referent and their rated behavior, the perceived discrepancy will result in an increased effort to change performance in order to meet the performance criterion. Thus, performance change will likely occur on average for all individuals targeted by the leader development initiative and there will likely be some degree of variability about the observed improvement; existing pretest-posttest studies actually show this very pattern of results (Smither, et al., 2005). Together, we expect that after accounting for measurement error, performance will improve following the

administration of MSF over time and that there will be variability in the estimated growth trajectories for individuals in the sample.

Hypothesis 1a: Performance will improve following MSF after accounting for measurement error.

Hypothesis 1b: There will be variability in performance change following MSF after accounting for measurement error.

The growth trajectory of performance change following MSF administrations has rarely been investigated. Our review of the literature only revealed three studies capable of characterizing the growth trajectory associated with performance change following MSF (Dai, et al., 2010; Reilly, et al., 1996; Walker & Smither, 1999), and none used modern methods to operationalize change. This gap in the literature is especially critical considering that the rate of change for performance improvement following feedback interventions is frequently noted in the theoretical literature. According to Control Theory, as feedback recipients improve their performance, the discrepancy between their performance and the organizational referent necessarily decreases over time; thus, individuals will demonstrate less effort to improve that specific performance behavior (Carver & Scheier, 1982). In this regard, Control Theory points to a nonlinear growth trajectory of performance change following feedback interventions. Alternatively, Locke and Latham (1990) suggest that Control Theory largely excludes internal motivations from their model and that individuals experience motivation not solely from a performance discrepancy with a performance referent. Rather, discrepancy and ultimately motivation are experienced whenever a specific, challenging goal is set. To this point, the rate of change may be directly related to the setting of challenging goals for the target

managers. In this particular sample, a goal setting meeting was required after each MSF administration with the direct supervisor, thus one might expect linear rather than nonlinear growth.

Using traditional approaches, three true-longitudinal studies of MSF have demonstrated somewhat conflicting representations of the performance change growth trajectory following MSF. Dai, et al. (2010) and Reilly et al. (1996) showed the strongest performance improvements occurring between the first and second administration, with the rate of change declining thereafter, supporting the pattern proposed by the tenants of Control Theory. Consistent with this finding, Smither et al.'s (2005) meta-analysis of repeated measures designs shows that the strongest performance gains are made when MSF administrations happened within less than a year of one another. Walker and Smither (1999), however, showed stronger improvements occurring after the second and third administrations rather than the first. Together, despite MSF's widespread use in modern organizations as well as strong theoretical and empirical evidence suggesting that MSF will result in increased performance change over time, the developmental nature of the appraisal system is still not agreed upon. As such, the MSF literature would benefit from a large-scale investigation using a more rigorous methodological approach to definitively describe the growth trajectory of performance change following MSF.

Research Question 1: What pattern of change over time will best describe performance improvement following MSF?

Factors Affecting Performance Change Following MSF

As noted above, past research on moderators of performance change suffer from a variety of methodological limitations. In terms of MSF characteristics that might

moderate perceived performance change, we examine the rater source and performance dimensions. We then turn our attention to variables which mark one's readiness for change. These include one's managerial experience, initial status, and enrollment in a tuition assistance program (TAP).

Rater source. Much MSF research has focused on the interpretation of source effects in performance ratings (Conway, 1996; Conway & Huffcutt, 1997; Hoffman, Lance, Bynum, & Gentry, 2008; Lance, Baxter, & Mahan, 2006; Woehr, Sheehan, & Bennett, 2005), and although there remains some contention over the meaning of these effects (Viswesvaran, Schmidt, & Ones, 2005; Wherry & Bartlett, 1982), there now exists substantial evidence that source-based differences in ratings do not reflect error, but instead, reflect valid differences in perceptions of performance. Indeed, extant literature suggests that (a) systematic rater source variance constitutes a sizeable portion of MSF ratings (Hoffman, et al., 2010), (b) it evidences theoretically grounded relationships with other variables (Hoffman & Woehr, 2009), and (c) non-traditional sources provide incremental information on managerial performance (Conway, Lombardo, & Sanders, 2001). To this point, in Smither et al.'s (2005) quantitative summary, they noted comparably weak yet different effect sizes by rater source (Self = -.04; Supervisor = .15; Peer = .04; Direct Report = .15). Similarly, Taylor, Russ-Eft, and Taylor (2009) also found that rating sources perceived varying degrees of training influence on subsequent behaviors. Importantly, source specific differences in growth trajectories for MSF performance have never been examined in true-longitudinal research. Dai et al. (2010) aggregated peers, direct reports and supervisors, and did not analyze self-report data. Reilly et al. (1996) only examined upward feedback (direct

reports' ratings) and self ratings. Although the authors tested for decreasing self-other agreement over time, their analyses did not include a comparison of self versus direct-report's perceived improvement. Finally, Walker and Smither (1999) only examined upward feedback (direct reports' ratings) and did not collect self ratings.

Although no true-longitudinal study has examined the source-specific growth trajectories of perceived performance change over time, a close look at the extant MSF literature suggests that perceptions of change in performance depend on the rating sources. First, past research suggests that different sources' ratings capture unique aspects of the criterion domain (Lance, Hoffman, Gentry, & Baranik, 2008; Hoffman, et al., 2010). For instance, target ratees may be more motivated to show improvement when in the presence of their supervisor because supervisors can provide valued rewards, resulting in higher levels of performance improvement in bosses' relative to other sources' ratings. Additionally, supervisory ratings are more reliable than other sources' ratings (Conway & Huffcutt, 1997). Similarly, evaluating direct reports' performance is explicitly part of supervisors' job, resulting in greater attention to ratee performance relative to other sources. Thus, supervisors may be more sensitive to improved performance behaviors in the target ratee. On the other hand, peers are proposed to have the greatest opportunity to observe target ratees' typical behaviors (Murphy & Cleveland, 1995; Hoffman & Woehr, 2009; Hooijberg & Choi, 2000). In contrast to the maximal effort exerted in interactions with their supervisors, ratees are likely to exert less effort in the frequent, day-to-day interactions with their peers. In this regard, peer performance ratings might be considered a reflection of ratees' typical (rather than maximal) performance and, therefore, would be expected to be more variable relative to

supervisors' ratings. Smither et al.'s (2005) review supported this trend when change in MSF was gauged using peers' ratings. To the extent that peers' performance ratings are less reliable, true increases in ratee performance may be less likely to be noted by this source. In the context of managerial training programs, supervisors perceive stronger behavioral changes following training than do peers (Taylor, et al., 2009).

On the other hand, we expect self ratings will show the least change. Many existing studies recognize that self ratings are higher relative to other sources (Atwater & Yammarino, 1997; Harris & Schaubroeck, 1988; Nilsen & Campbell, 1993). Given that self ratings typically start with a relatively high initial status, it may be difficult to exhibit improvement if one consistently rates oneself favorably. Based in conceptual and empirical differences underlying supervisor, peer, and self ratings, we hypothesized:

Hypothesis 2a: Supervisors' ratings will exhibit greater performance improvement over time than peers and self raters.

Hypothesis 2b: Peers' ratings will exhibit greater performance improvement over time than self raters.

Performance dimensionality. Although most MSF instruments are multi-dimensional, past research has not investigated the role of performance dimension in performance change. In contrast, the past 40 years (Austin & Villanova, 1992) have seen substantial progress in models of performance. For instance, the distinction between task and interpersonal performance consistently emerges across a variety of domains in conceptual models, factor analytic research, and nomological network analyses (Borman & Motowidlo, 1997; Fleishman, 1957; Hoffman, Blair, Woehr, & Meriac, 2007; Kram, 1985; Noe, 1988; Shore, Thornton, & Shore, 1990). Within the context of leadership

development and MSF particularly, a clear understanding of the role of performance dimension would provide practical information to practitioners and also expand the knowledgebase concerning the psychometric quality of ratings by pointing to areas that are more (or less) likely to improve over time. As an example, Zimmerman, Mount, and Goff (2008) found that theoretically grounded source-dimension rating combinations exhibited stronger criterion-validity than other combinations. For example, supervisors' ratings of consideration behaviors and others' ratings of initiating structure behaviors were each found to be the most predictive combinations of goal-performance. These findings suggest that some sources' perspectives may be more developmental depending on the performance dimension. Further, these results are aligned with theoretical work suggesting that performance dimensions may be attended to differently by various rating sources (Lance, et al., 2006).

Despite the dearth of research directly examining the role of performance dimension in the MSF-performance change relationship, related fields suggest that perceptions of improvement will be stronger for task performance. Feedback Intervention Theory (Kluger & DeNisi, 1996) identifies situations in which performance is most likely to improve via the attention processes of the target ratee. Specifically, when the target ratees' attention is focused on task learning processes (e.g. *How can I execute this task more effectively?*) performance is proposed to increase most. These individuals are motivated to identify the behavioral causes for substandard performance and the behavioral changes they can make to increase performance on this task. Although a helpful process for task-oriented performance, this process may prove more difficult for feedback on interpersonal dimensions. In their model of behavior changeability,

Hellervik, Hazucha, & Schneider, (1992) posited that the ability to change is based on the complexity of the behavior. Thus, a simple task-oriented skill (e.g. operating a cash register) may be easier to develop than a broad ability domain such as interpersonal skills (Hellervik, et al., 1992; Rupp, Snyder, Gibbons, & Thornton, 2006).

Hypothesis 3: Greater change in performance will be perceived for task, rather than interpersonal-oriented performance.

Antecedents of Change Following MSF

Given the weak and variable effects uncovered in their review, Smither et al. (2005) urged further research exploring "conditions and for whom…multisource feedback [is] likely to be beneficial" (p. 60). Day and Sin (2011) collectively referred to variables antecedent to positive performance change as indicative of a broader, umbrella construct labeled as readiness for change. For instance, quasi-longitudinal research has conceptualized personality characteristics such as feedback orientation (London & Smither, 2002), continuous learning (Vicere & Fulmer, 1998), and openness to experience (Dominick, Reilly, & Byrne, 2004) as individual differences reflecting readiness for change. Although past research has begun to examine readiness for change, much of this research is hindered by methodological limitations. Thus, the present study extends past research on readiness for change using the aforementioned methodological advances and by operationalizing readiness for change with initial status and managerial experience. Also, aligned with previous work identifying the role of behavioral developmental engagement following feedback interventions (Woo, Sims, Rupp, & Gibbons, 2008), we use engagement in voluntary developmental activities (enrollment in a tuition assistance program) as an index of readiness for change.

Managerial experience. Several models of leadership specify some form of past experience as predictors of leader effectiveness (Bass, 1990; Locke, 1991; Mann, 1965; Yukl, 2006). Although early work examining the relationship of experience to leader effectiveness concluded that experience was not as important to successful job performance as had been previously thought (e.g. Fiedler, 1970; Fiedler, 1992), several subsequent quantitative summaries of the literature suggest a weak-moderate relationship between experience and job performance (Hoffman, Woehr, Maldagen-Youngjohn, & Lyons, 2011; Hunter & Hunter, 1984; McDaniel, Schmidt, & Hunter, 1988) characterized by substantial variability (Quiñones, Ford, & Teachout, 1995).

A common conceptualization of experience is tenure in one's current managerial position (Quiñones et al., 1995). Managers with more experience in the role have a larger set of past situations and behavioral scripts to draw from when resolving new situations, (Zaccaro, Gilbert, Thor, & Mumford, 1991). Thus, seasoned veterans of the managerial workforce who are well acquainted with their current role in the organization, and have perhaps been the target of several professional development programs as a manager, may have higher levels of initial status relative to less experienced managers. In essence, there might be a learning curve associated with being a manager. From this perspective, novice managers may be at a performance-related disadvantage and welcome feedback concerning their current performance and suggestions on how they might improve more so than their more seasoned colleagues. Existing literature supports this idea, showing that time-oriented measures of experience evidence a weak-moderate influence on measures of job performance (ρ=.27; Quiñones et al., 1995). Thus, managerial experience could be a marker of employee readiness for change. To this end, the leader development

literature supports a negative relationship between time-oriented measures of managerial experience and learning orientation (DeRue & Wellman, 2009), such that less experienced managers have a stronger learning goal orientation; thus, novice managers may be more ready for change, and therefore, more likely to exhibit behavioral change following feedback interventions. Consistent with these arguments, the strongest MSF performance improvements occur: (a) between the first and second MSF administrations with the rate of change for improvement decaying thereafter (Dai, et al., 2010; Reilly, et al., 1996) and (b) for the initially weakest performing ratees (Reilly, et al., 1996; Walker & Smither, 1999). Thus, experienced managers may have already made behavioral changes from past MSF administrations and, as a result, be performing at stronger levels than novice mangers. Accordingly, we offered the following hypotheses:

Hypothesis 4a: Managerial experience will be related to initial status.

Hypothesis 4b: Managerial experience will be negatively related to performance improvement.

Discretionary developmental activities. Smither et al. (2005) suggests that "performance improvement is likely *only* for feedback recipients who take appropriate action…in response to their feedback" (p.56; emphasis added). In light of the substantial number of cases in which feedback interventions result in *lowered* employee performance, some have concluded that for performance gains to be made, it is important that the feedback recipients proactively work to improve their performance (Kluger & DeNisi, 1996). Aligned with previous conceptual and empirical research (Dominick, et al., 2004; London & Smither, 2002; Vicere & Fulmer, 1998), feedback recipients that are focused on improving behavioral task performance are expected to make the greatest

performance gains. To this point, research examining developmental activities has shown many benefits of executive coaching following MSF administrations: direct reports (Luthans & Peterson, 2003; Smither, London, Flautt, Vargas, & Kucine, 2003) and supervisors (Smither, et al., 2003) perceive stronger performance improvement, target ratees set more specific goals and solicit ideas for improvement (Smither, et al., 2003), and the target ratees also engage in behavioral influence tactics learned in the coaching session (Seifert & Yukl, 2010). However, many existing operationalizations of developmental activities overlook the importance of internal motivation to improve (Hazucha, et al., 2003; Luthans & Peterson, 2003; Seifert & Yukl, 2010; Seifert, et al., 2003; Smither, et al., 2003). To this end, all of the above examinations were mandatory developmental programs and thus, motivation to improve was unexamined.

Aligned with much of the leader development literature that characterizes developmental activities as an outcome of ratee reactions (e.g. feedback acceptance, satisfaction; Kudisch, 1997; Smither, et al., 2005), research examining the role of discretionary activities in leader development has primarily characterized them as an outcome variable (Woo et al., 2008). Other work, however, has characterized discretionary developmental activities as a predictor of subsequent markers performance improvement. Along these lines, Walker and Smither (1999) found that managers in a five year study who chose (of their own accord) to solicit feedback from direct reports increased their performance over several years, more so than those who did not. That is, target ratees who were ready for change were likely to proactively seek out developmental experiences that might improve their performance. Tuition assistance programs (TAPs) are discretionary activities for career development that have not

received attention in the extant literature. These programs offer full or partial funding to employees seeking additional education and involvement is at the discretion of the employee. In a society where a master's degree is estimated to earn its owner (over their lifetime) an estimated $1.3 million more than someone with only a high school diploma (Day & Newburger, 2002), individuals seeking to work their way into higher-paying, more prestigious levels within or outside of their current organization are likely to enroll in these types of developmental programs. Thus, to the degree that discretionary involvement in TAPs is another indicator of readiness for change, we expect that these motivated individuals will exhibit strong initial performance and subsequent performance change over time. As such, we expect that individuals who were recently or currently enrolled in the discretionary TAP during their MSF administration were ready for change and, therefore, more likely to exhibit strong performance and performance improvement over time. Accordingly, we hypothesize:

Hypothesis 5a: Enrollment in TAP will be positively related to performance initial status.

Hypothesis 5b: Enrollment in TAP will be positively related to performance improvement.

Initial performance. To the extent that performance change is initiated by perceiving dissonance between one's current performance and the organization's performance standards (cf. Carver & Scheier, 1982), ratees receiving poor performance feedback will be those most likely to be motivated to make behavioral changes. Similar to recommendations in the training needs assessment literature (McGehee & Thayer, 1961; Wexely, 1984), if the individual does not show a performance deficit, they may not need to work/attend training to develop that competency. More pragmatically, one has

less room to improve on a 4.5 relative to a 3.5; in essence, range restriction makes performance improvement challenging. Thus, initial MSF performance may be a strong marker of readiness for change and predictive of subsequent performance change. Several quasi-longitudinal MSF studies have examined the effect of initial status on performance change and in doing so have consistently shown that the ratees who perform poorly are the most likely to improve (e.g. Hazucha, et al., 1993; Hegarty, 1974; Smither, et al., 1995). Yet, these studies have been unable to examine the influence of initial status on nonlinear performance change or change over multiple points in time, and this research has rarely compared source specific differences. The existing true-longitudinal examinations of performance change following MSF also show that the strongest improvements are generally made by those initially poor performing managers (Reilly, et al., 1996; Walker & Smither, 1999). However, again, these results are to some degree limited in that they have not accounted for measurement error and could only test the influence of initial status as a moderator of performance change; not as a mediating variable in a larger theoretical model. In the current study, initial status will partially-mediate the influence of two antecedents onto performance change, and also have an indirect effect on an outcome variable- advancement to leadership roles (Figure 1). Given that longitudinal (Reilly, et al., 1996; Walker & Smither, 1999), pretest-posttest (Smither, et al., 1995), and meta-analyses of quasi-longitudinal (Smither, et al., 2005) examinations of MSF performance change over time support the propositions of Control Theory, we offer the following hypothesis:

Hypothesis 6: There will be a negative relationship between initial status and performance improvement.

Outcomes

As previously mentioned, little research has examined outcomes of change in performance following MSF, and no true-longitudinal studies have been done. Below we use Attribution Theory (Heider, 1958; Martinko & Gardner, 1987; Martinko, Harvey, & Douglas, 2007) to hypothesize the role that developmental trajectory might play in promotion rate.

Promotion rate. MSF systems are ultimately organizational appraisals used for leadership development. However, MSF studies oftentimes examine other measures of performance as the primary outcome (cf. Smither, et al., 2005), rather than outcomes directly associated with leader development. This line of research has been quite fruitful in identifying the concurrent validity of MSF, evidencing MSF-ratings' relationship with assessment center performance (Atkins & Wood, 2002; Helland, Hoffman, & Smith, 2003; Warech, Smither, Reilly, Millsap, & Reilly, 1998), annual performance review ratings (Smither & Walker, 2004), performance on a structured interview (Darr & Catano, 2008), measures of objective performance (e.g. production, profit; Conway, et al., 2001), and even turnover and service quality (Church, 2000). Yet, direct investigations of MSF as a predictor of managerial outcomes, such as promotion rate, have been examined far less often. Church (1997), for example, found no relationship between self-other congruence in ratings to management level, but did not directly investigate MSF performance ratings' relationship with management level. Ostroff, Atwater, and Feinberg (2004) supported a weak yet positive relationship between MSF ratings and compensation and organizational level. To date though, no true-longitudinal

investigation of MSF has examined change in MSF performance on developmental outcomes.

The sparse, cross-sectional studies on this topic represent a substantial gap in the MSF literature because MSF systems are ultimately organizational appraisals used for leadership development. That is, professional development programs are not enacted solely to develop employees but rather to groom the next set of organizational leaders. In doing so, MSF provides performance-relevant information to ratees so that they can hone their skills in order to ascend the organizational hierarchy. In other words, although it is important to understand the role that MSF plays in performance improvement, from an organizational perspective it is likely equally important to understand whether MSF influences subsequent promotion rate. Thus, it is important to understand MSF within the context of developing organizational leaders. This study reflects an initial attempt to do so by using promotion rate as a criterion.

Promotion rate quantifies employees' ascension through an organizational hierarchy into roles marked by increasing degrees of prestige and responsibility; for this reason, it is an appropriate outcome of MSF (Hoffman & Baldwin, 2011). Despite the sparse work in this area of the MSF literature, an examination of existing theory and empirical work in related fields suggests that change in MSF performance will be related to promotion rate. First, many other measures of leader performance demonstrate positive relationships with promotion rate. The assessment center literature, for example, consistently exhibits positive relationships between performance ratings and promotion-rate (Arthur, Day, McNelly, & Edens, 2003; Klimoski & Brickner, 1987). On a similar note, employees who are the target of other developmental initiatives, such as mentoring

programs, are generally considered more promotable by others (Gentry & Sosik, 2010). To the extent that leader performance ratings capture variance of managerial performance (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Conway et al., 2001), and that strong performers are more likely than others to be identified for and rewarded with promotions, we expect that all sources' ratings of MSF initial status should exhibit a positive relationship with promotion rate. Stated differently, more effective managers, as indicated by their MSF, will be promoted at a faster rate relative to other managers. We supplement this question by comparing differences across sources in promotion rate and the impact of pattern of change on promotion rate; questions previously unexplored in this area. To the extent that feedback recipients' primary means of attaining promotions is via their direct supervisor, we expect that the supervisory ratings' influence on promotion rate will be stronger than others'. Accordingly, we offered the following hypothesis: *Hypothesis 7a: MSF initial status will be positively related to promotion rate with supervisory ratings demonstrating the strongest relationship.*

In addition, we expect an individual's pattern of change over time to have an important impact on outcomes. According to Attribution Theory (Heider, 1958; Martinko, et al., 2007), raters seek to explain or attribute observed behaviors by means of underlying motivations. Specifically, effort is oftentimes ascribed to situations in which employees' performance exhibits systematic change over time: increased effort is ascribed to consistent improvement (Dweck & Leggett, 1988) and lack of effort to consistent deterioration (Jones, Rock, Shaver, Goethals, & Ward, 1968). In terms of managerial advancement, managers who exhibit rapid development, more so than their colleagues, may be perceived as being more ready for change and, therefore, placed on

the organizational fast track to leadership positions. Performance management research supports this notion, showing that raters who note an upward trend in ratee performance attribute the performance improvement to ability (Reb & Greguras, 2010); thus, these individuals may be more likely to be identified for promotion opportunities. Indeed, performance trends predict employee compensation above employees' average performance (Barnes, Reb, & Ang, 2012). On the other hand, our analyses will also allow us to examine antecedents of career plateaus. That is, individuals whose MSF performance is marked by a constant decline, or perhaps unusually high inconsistency, may be more likely to be passed over for promotions. Further, to the extent that supervisors ultimately decide whether the feedback recipient is promoted or not, we expect supervisory perceived performance improvement will exhibit a stronger relationship with promotion rate than other sources'. As such, we offered the following hypothesis:

Hypothesis 7b: MSF performance improvement over time will be positively related to promotion rate with supervisory perceptions demonstrating the strongest relationship.

CHAPTER 3

METHOD

Sample

The sample was drawn from the managerial staff at a large packaging company in the southeast United States whose work spans a variety of managerial levels with a variety of specialties (e.g., mangers of small to large teams of less-specialized to highly-specialized employees). Their managerial performance was measured over four years and three MSF administrations (2005, 2006, and 2008). Technical difficulties concerning the storage and maintenance of performance data within the organization precluded the collection of MSF performance ratings in 2007. Further, the individual-level performance data set for 2008 was only available with performance ratings aggregated up to the rater-source and rating construct-level. Thus, we justify the performance ratings' aggregation up to rater-source and construct-level via the 2005 and 2006 un-aggregated datasets in the Analyses section below. Next, we specified inclusion criteria across years, such that eligible employees must have been rated by both all three rater groups at all three time points and also had at least 75% of the MSF measure completed by all sources at each time point. We utilized conservative inclusion criteria to ensure that techniques employed for data missingness would not unduly influence results. Of the initial 7,334 ratee sample, 5,128 ratees met these criteria and were eligible for inclusion in the present study. These ratees were rated by over 32,000 total raters at each time point. Each feedback recipient received feedback each year from an average of 4.12 peers and 1.29 supervisors (Table

1). Together, our sampling technique qualified the present study as a true-longitudinal design in that it utilized at least three repeated measures of performance, an outcome with values that change systematically over time and a sensible metric for tracking time (Day, 2011; Singer & Willett, 2003). The sample of ratees was mostly white (71%) and male (76%).

Procedure

Prior to the annual performance review, each manager identified peers and supervisors who might provide relevant information about their performance. After identifying potential raters, the target ratees' direct supervisor approved the list of raters based on their ability to observe the target manager's performance. All raters (self, supervisors, and peers) rated the target ratee on a variety of behavioral dimensions. One to three months after each administration, target ratees were required to meet with their immediate supervisor to discuss their performance ratings, receive developmental coaching, and set behavioral goals for the following year. The MSF ratings were used for developmental purposes only.

Measures

MSF. The MSF measure was a 37 item survey, designed by the sponsor organization to measure eight managerial performance constructs. See Appendix A for definitions of the performance constructs. Aligned with applications of Socioanalytic Theory to managerial performance (Hogan & Holland, 2003), the eight performance constructs were expected to measure two broad domains of behavior- those that facilitate personal success and those that facilitate group success. Activities that facilitate personal success include behaviors of technical skill, such as problem solving, formulating short-

and long-term goals, and providing technical advice to others (Borman & Brush, 1993). Accordingly, ratees' task performance was measured with Problem Analysis, Results Orientation, and Quality Improvement. On the other hand, behaviors that facilitate group success include aspects of interpersonal performance like communicating, representing the organization, maintaining working relationships, and influencing others (Borman & Brush, 1993); thus, interpersonal performance was measured with Relationship Skills, Teamwork, Communication, Managing Conflict, and Developing Others. Together, we expected the 37 manifest indicators to load onto eight performance constructs, which in turn were expected to load onto two overall dimensions of performance (Figure 2). For all performance items, respondents were asked to rate their agreement with statements that ratees regularly engaged in the interpersonal and task behaviors. Responses were marked on a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) with a sixth option indicating "I don't know."

Managerial experience. Managerial experience was operationalized as the number of years of managerial status within the company. A newly hired manager or an employee who was promoted to managerial status the year of the first data collection (2005) was coded as a 0.

Tuition assistance program. The assistance program covered $150 per credit hour at the bachelor's and master's-level and required the recipient to submit justification identifying the alignment between the program of study and the organization's business strategy to the human resources department. Enrollment in the organization's TAP was operationalized as the financial sum covered by the organization for the employee's tuition between 2000 and 2008 divided by one thousand. We used the financial sum

because it reflects, in dollars and cents, the investment of the TAP in the employees' development. We also specified 2000-2008 time range to ensure that we accounted for those completing a four year bachelors program in the year preceding the first data collection.

Promotion Rate. To categorize jobs based on hierarchical level, we adopted a four category hierarchy, such that promotion rate was reflected by upward movement through any of these four categories. These categorizations (from lowest authority to highest) included supervisor, manager, staff, and senior staff. The distribution of this variable was positively skewed and violated assumptions of normality, marked by the few observed promotions (19% of the sample). Thus, before proceeding with substantive tests, we transformed the promotion variable with the following equation:

$$y = log10(x + 1)$$

where y is the transformed promotion variable and x is the observed number of promotions. This transformation yielded a distribution for the promotion rate outcome that was more representative of a normal distribution; thus, the transformed outcome variable was used in subsequent analyses.

Analyses

Data management. First we examined the degree and patterns of missingness within the datasets. Participants who failed to respond to 25% or less of the performance items were treated as missing at random cases. For these, we estimated missing values' parameters using the full information maximum likelihood option in M*plus* (Muthén & Muthén, 1998-2007). Provided the time required to attain managerial status, the perks received after becoming a manager, and the mandatory nature of the MSF system, it was

expected that attrition in the current study would covary strongly with retirement; thus, ratees who dropped out of the study before its completion were classified as missing not at random and deleted from further analyses (Newman, 2009). Next, cases exhibiting more than 25% of missing data within any time point were also classified as missing not at random and deleted. For instance, if the rater in question represented the only respondent for a rater source (e.g. supervisor ratings were unavailable for one year), the entire case's data was eliminated from subsequent analyses.

Measurement model. The two factor model for self raters, peers, and supervisors (Table 2; Models 2,4 and 6, respectively) provided a significantly closer fit to the data relative to the more parsimonious one factor models (Table 2; Models 1, 3, and 5) in terms of the $\Delta\chi^2$ test for the 2005 administration. Similar results were found for the 2006 and 2008 administrations[1]. However, given the sample size of the present study, the $\Delta\chi^2$ test has enough power to detect even minute differences in nested models' fit to the data. On the other hand, the remaining fit indices were quite similar, and the average relationship between the latent interpersonal and task factors was near unity (ψ=.98), leading us to adopt the one factor model.

Aggregation prerequisites. Consistent with past MSF research (Hoffman et al., 2010; Smither, et al., 2005), we aggregated individual, peer, and supervisor raters within sources. To justify aggregation, we used a variation of James, Demaree, and Wolf's (1984) interrater agreement statistic $r^*_{wg(j)}$ (Lindell, Brandt, & Whitney, 1999) to estimate interrater agreement. Average $r^*_{wg(j)}$s are reported in Table 3. Results support sufficient agreement ($r^*_{wg(j)} \geq .70$; Brown & Hauenstein, 2005; Lance, Butts, & Michels, 2006). Provided that these statistics are fairly consistent over time, we expect that these findings

additionally provide support for the aggregated 2008 dataset. Further, we aggregated by performance construct, such that subsequent SLGMs could account for overall performance at each time point as measured by eight indicators of performance.

CHAPTER 4

RESULTS

Table 4 presents correlations and descriptive statistics for self, peer, and supervisor sources. Before proceeding with the substantive tests of the present study, we next tested the three prerequisites to conducting LGM: Measurement invariance over time and substantiating change and variance in change for the focal variable (e.g. Bentein, et al., 2005). To test for invariance over time in the focal variable (Williams, Edwards, & Vandenberg, 2003), we used the procedures outlined by (Vandenberg & Lance, 2000). A nested comparison of self raters', peers', and supervisors' configural invariance models (Models 1, 5, and 9, respectively) with their respective metric invariance models (Table 5; Models 2, 6, and 10) revealed a significant $\Delta\chi^2$. However, the ΔCFI statistics were below Cheung and Rensvold's (2002) recommended cutoff for invariance analyses of .01, supporting metric invariance. Using the same fit criteria, subsequent invariance tests showed that the observed data failed to support scalar invariance (Table 5); thus we proceeded in testing for partial scalar invariance. Results showed that in allowing item-level intercepts for Relationship Skills, Teamwork, and Problem Analysis to be free to vary across time points for self and peer raters, partial scalar invariance was supported. For supervisors, Quality Improvement, Problem Analysis, and Developing Others were allowed to be free to vary. We used these partial scalar invariance models (Table 5; Model 4, 8, and 12) as Baseline models in subsequent tests of change over time. Given that these results exceed the criteria necessary to justify

measurement invariance for longitudinal analyses, we proceeded with estimating change over time for each rater source.

We next turned our attention to establishing the form of performance change. To test for the presence of nonlinear change in the focal variable, we made nested comparisons between the baseline models of change (Table 6; Models 1, 3, and 5) with an optimal change model that allowed for nonlinear growth trajectories for each source. The optimal change model failed to converge for self raters and revealed an inadmissible solution for peers and supervisors. Thus, results across sources indicated that the nonlinear specification was an inappropriate representation of the observed datasets. As such, we answered Research Question 1 in noting that each source perceived linear performance change.

Next, we compared competing conceptualizations of the uniquenesses underlying the measurement model at each time point (cf. Bentein, et al., 2005; see Table 6). A nested comparison between a heterogeneous (Figure 3) and homogenous (Figure 4) conceptualization of uniquenesses over time for each source favored heterogeneous uniquenesses for self raters ($\Delta\chi^2$ = 3,542.21, Δdf = 16, ΔCFI = -.03), peers ($\Delta\chi^2$ =6,125.13, Δdf = 16, ΔCFI = -.03), and supervisors ($\Delta\chi^2$ = 16,816.35, Δdf = 16, ΔCFI = -.11). Together, a linear model of change with uniquenesses modeled as heterogeneous, yet correlated over time was used in subsequent analyses for self raters, peers, and supervisors.

Having supported the assumptions underlying LGM and characterized the underlying measurement model of change over time, we next estimated change and variance in the change estimates in our Baseline correlated uniqueness models of linear

change (Table 6; Models 1, 3, and 5 for self raters, peers, and supervisors, respectively). Results suggested that self raters (κ = -.02, σ^2 = .01) and peers (κ = .004, σ^2 = .004) perceived slight performance change over time; however, supervisors did not perceive performance change (κ = -.002, σ^2 = .01; Table 7). These preliminary models of change substantiate performance change, but only for self raters and peers. Thus, Hypothesis 1a was partially supported. On the other hand, there was variability in performance change for all sources, providing full support for Hypothesis 1b. Together, the results show that performance change across the three rater groups was minimal over the four year time span, and in the expected direction for only peers. There was, however, variability in individual change trajectories suggesting potential moderators of performance change.

Moderators of performance change. To determine the influence of rater source on performance change (Hypotheses 2a-b), we constructed a 95% confidence interval around the average change parameter estimated from the each sources' LGM (Table 7). The confidence intervals for self raters (-.027 to -.018), but not peers (.001 to .008) and supervisors (-.007 to .003) were nonoverlapping with other sources', suggesting that only the self raters' perception of change was significantly different from the other sources'. Thus, we failed to support Hypothesis 2a because supervisors perceived the weakest amount of performance improvement in absolute terms. However, peers perceived stronger performance improvement in comparison to self raters, and the confidence intervals for the two sources were nonoverlapping, providing full support for Hypothesis 2b. Because a unidimensional performance model was supported, we were unable to investigate differences by type of performance. Yet, given that raters did not distinguish

between types of performance in their raters, it is safe to conclude that performance dimensions did not moderate performance change, leading us to reject Hypothesis 3.

Antecedents and consequences of performance change. We next turned our attention to investigating the antecedents and consequences of performance change following MSF. We first tested a measurement model including initial status, performance change, managerial experience, developmental activities, and promotion rate for self raters, peers, and supervisors (Table 8; Models 1,4 and 7, respectively), which showed adequate fit. We then specified structural parameters to test our theoretical model (Table 8; Models 2, 5, and 8 for self raters, peers, and supervisors, respectively) and Hypotheses 4-7. The structural model provided an adequate fit to the observed dataset for self raters, peers, and supervisors. In testing the theoretical model for each rater source, we noted that some hypothesized paths were statistically non-significant (see Table 9), and similar underlying models were found for all rating sources. In order to retain statistical power in subsequent tests, we eliminated non-significant paths and made a nested comparison of the resulting model (Table 8; Models 3, 6, and 9) with the theoretical model for each rater source. The nested comparison for self raters, peers, and supervisors suggested that the models with the removed paths provided comparable fit to the theoretical model (ΔCFI<.01) for all three sources, thus the more parsimonious models were retained for interpretation (Table 8; Models 3, 6, and 9; Figures 5 – 7).

In answering our remaining hypotheses, we first turned our attention to predictors of initial status. Managerial experience was significantly related to initial status for peers (β = .19) and supervisors (β = .15) but not for self raters (β = .003, *ns*), suggesting that those with more managerial experience were evaluated more favorably during the initial

administration. Demonstrating a similar pattern, TAP enrollment was weakly related to initial performance for peers (β = .05) and supervisors (β = .07), but not for self raters (β = -.01, p = *ns*), suggesting that those who enrolled in the company's TAP had slightly higher initial performance. Despite the statistically significant findings, it is noteworthy that TAP and managerial experience explained minimal variance in initial status (2.7% - 3.7%). Together, because neither managerial experience nor TAP predicted self raters' initial status, Hypotheses 4a and 5a were only partially supported.

Next, we turned our attention to the proposed predictors of performance change. Of the three antecedents, initial status had the strongest effect on performance change, lending full support to Hypothesis 6. Specifically, the relationship between initial status and change was consistently negative and moderate across sources. This pattern suggests that lower performers are the most likely to improve for peers, whereas self raters' and supervisors' ratings demonstrated negative change over time; thus the moderate negative influence of initial status on change is rightfully interpreted such that lower initial performers demonstrate more performance change over time (albeit negative).

In order to better understand initial status' moderating influence on performance change by rater source, we used the methods proposed by Lance (2005) to plot the moderating effect for each source. This method first identifies the predicted initial status and change equations for a rating source. For self raters, for example, these equations are recognized as the following:

$$IS' = \alpha \tag{1}$$

$$CH' = \alpha + \beta_{CH,IS} IS + \beta_{CH,TAP} TAP \tag{2}$$

The beta weights and latent variable means associated with equations 1 and 2 can be filled in through the results of the rating source's supported structural LGM. Further, standard deviations of the estimated initial status latent mean can be used to identify equations for ratees that are high and low on the moderating variable. We described these subgroups as those exhibiting one and a half standard deviations above and below the initial status mean. Thus, high and low groups on the initial status moderator could be described through the following:

$$LO - IS' = 3.97 - 1.5 * .34 = 3.46 \quad (3)$$

$$HI - IS' = 3.97 + 1.5 * .34 = 4.48 \quad (4)$$

$$LO - CH' = -.02 - .071 * 3.46 - .002 * 2.20 = -.27 \quad (5)$$

$$HI - CH' = -.02 - .071 * 4.48 - .002 * 2.20 = -.35 \quad (6)$$

These estimated initial status and change values can be used in the linear LGM equations, which hold initial status' influence on performance at unity over time and specify change as increasing by whole integers at each time point. Thus, the high and low initial status moderator groups would demonstrate two unique growth trajectories over time with each time point's predicted performance calculated by the following:

$$Y_0 LO = 1.0 * 3.46 - 0.0 * .27 = 3.46 \quad (7)$$

$$Y_1 LO = 1.0 * 3.46 - 1.0 * .27 = 3.19 \quad (8)$$

$$Y_3 LO = 1.0 * 3.46 - 3.0 * .27 = 2.64 \quad (9)$$

$$Y_0 HI = 1.0 * 4.48 - 0.0 * .35 = 4.48 \quad (10)$$

$$Y_1 HI = 1.0 * 4.48 - 1.0 * .35 = 4.14 \quad (11)$$

$$Y_3 HI = 1.0 * 4.48 - 3.0 * .35 = 3.45 \quad (12)$$

This method was employed for all predictors of change for each rating source. The moderator plots for initial status' influence on change over time are depicted in Figures 8 – 10. All sources showed the high initial status subgroup as demonstrating a slightly stronger negative rate of change over time than the low initial status subgroup. Thus, weak performers' ratings decreased less rapidly over time.

Next, we examined managerial experience as a predictor of change. When initial status was in the model, managerial experience was not predictive of change in performance for any of the sources, leading to the rejection of Hypothesis 4b. These findings suggest that managerial experience has a significant indirect effect on change through its influence on initial status for peers (β = -.15), but not supervisors (β = -.01, *ns*). On the other hand, enrollment in the organization's TAP demonstrated a weak negative effect on performance change for self raters (β = -.09) and peers (β = -.17), but not supervisors (β = -.01, *ns*), suggesting that TAP moderated the performance change (Lance, 2005). Figures11-12 represent this relationship by rating source and show the quite similar growth trajectories associated with high and low enrollment TAP ratees. Additionally, results suggest a weak, indirect effect of TAP through initial status onto performance change for peers (β = -.04) and supervisors (β = -.03), but not self raters. These results failed to support Hypothesis 5b. Together, these three antecedents accounted for 11%, 19%, and 70% of the variance in change in self, supervisor, and peer ratings, respectively with initial status being the primary predictor for all three sources.

In terms of consequences, results showed a weak, negative relationship between initial status and promotion rate for self raters (β = -.01) and peers (β = -.04), but not

supervisors (β = .001, *ns*), suggesting that those with high initial status were marginally less likely to be promoted. Given that this relationship was opposite of our predictions, we rejected Hypothesis 7a. Also opposite to predictions, performance change was significantly, weakly, and negatively related to promotion rate for all three sources, meaning that individuals with less change in performance were more likely to be promoted. Based on these findings, lower performers and those who perform more consistently were more likely to be promoted. These results failed to support Hypothesis 7b. Together, initial status and change accounted for 1.2% – 8.4% of the variance in promotion rate.

Supplemental Analyses

Latent class analysis. The unexpected direction of change for self and supervisor rater sources warranted further investigation of the data. To test if there were other meaningful growth trajectory trends within the sample, we utilized a latent clustering technique to identify systematic groups of ratees. Latent Class Analyses (LCAs) are used to identify unobserved heterogeneity within the population by identifying groups with similar rating patterns. For each rater source in the current sample, we conducted six LCAs. Because LCAs are non-nested models, we used the BIC and entropy statistics as decision criteria. Specifically, low BIC statistics or BIC statistics that demonstrate a large drop between consecutive solutions characterize better fitting models (Nylund, Asparouhov, Muthen, 2008). On the other hand, the entropy statistic characterizes the utility of the LCA classification and ranges in value from zero to one with higher values indicating stronger utility (Celeux & Soromenho, 1996). Although there exists no recommended cutoffs for clear class delineation, the entropy statistic can be used in

comparing LCA solutions (Pastor, Barron, Miller, & Davis, 2007). Given that fit indices, especially in the context of non-nested model comparisons, can at times point to different solutions, the above fit indices were examined in conjunction with theoretical plausibility, whether the additional class added meaningful information, and if the class included at least 100 observations to ensure the stability of LGM class results. All LCAs were conducted on a first order factor version of each rater source's supported measurement model of performance change[2].

Results of the LCAs suggested that peers' and supervisors' ratings characterized two distinct classes of ratees, yet self raters only distinguished a single class (Table 10). Means for the classes on the study variables are presented in Table 11. First, although demonstrating strong fit indices, the two class solution for self raters included a class with only eleven ratees; thus, the single class solution was favored for stability. Additionally, a two class solution was retained for the peer rater source because LCAs with three or more classes returned inadmissible solutions. Nonetheless, LCA results show that the vast majority of ratees were rated relatively high and maintained relatively high ratings across the three administrations for all rater sources; this suggests that most ratees were being evaluated as engaging in the target behaviors for the duration of the study, with little notable change in performance. However, for a small subset, performance improved following MSF for peers (9 %) and supervisors (10%; see Figures 13-14), and consistent with the earlier results (Reilly, et al., 1996; Walker & Smither, 1999), those classes that evidenced performance improvement tended to be lower performers initially. Further, supervisor and peer raters perceived these improvers as less experienced. In addition, it is noteworthy that for both sources, the class of individuals

with a higher promotion rate (Class 1) was typically more experienced and more likely to have exhibited a strong initial status.

Difference scores. The current study's use of LGM is a key strength over past examinations of performance change following MSF; however, we additionally calculated change over time using two more common methods to compare our results. First, we conducted a repeated measures analysis of variance for each rater source, the results of which suggested significant differences within each source (Table 12). Post hoc comparisons suggested mean differences between all mean comparisons except those exhibiting the weakest changes- peer raters between 2006 and 2008 and supervisor raters between 2005 and 2006. Second, we calculated effect sizes for each rater source in order to make direct comparisons with the effect sizes reported in the Smither et al. (2005) review. Using the formulas presented by Morris and DeShon (2002) we calculated performance change effect sizes for each rater source as:

$$d = \frac{M_{time2} - M_{time1}}{SD_{Change\ Scores}}$$

The standard deviation of change scores was calculated via the equation provided by Smither, et al. (2005) as:

$$SD_{Change\ Scores} = \sqrt{SD^2_{Time1} + SD^2_{Time2} - (2)(r_{Time1,Time2})(SD_{Time1})(SD_{Time2})}$$

Results suggest that this form of calculating change over time may yield vastly different estimates of change between years (Table 12). This set of results suggest that self raters (d = .04) and peers (d = .16) perceived a positive change in performance between 2005 and 2006, and supervisors perceived a slight decrement in performance (d = -.03). Yet the 2006-2008 effect sizes yield different effects- self raters (d = -.09), peers (d = -.04), and supervisors (d = -.13) all perceived a decrease in ratee performance. Notably, the

direction of change flipped depending on the administration years for self raters and peers and also, the magnitude of change differed by year for all sources. The difference in results between our estimates of change and these highlight the importance of conducting longitudinal research when characterizing change over time and also accounting for measurement invariance.

CHAPTER 5

DISCUSSION

This large-sample study reexamines the value of MSF in a process-based model of performance change over a four year time period using modern approaches to investigating change. These techniques, along with LCA, allowed unprecedented insight into performance change following MSF. The results are discussed in light of three overarching implications for the MSF-performance change literature. The most salient implication, that evidence for performance change was minimal, gives cause to reexamine the value and approaches underlying MSF. Yet, in supporting a theoretical model of the antecedents of change, the results point to the situations and individuals for whom MSF is most likely benefit over successive administrations of MSF, supporting the value of a targeted approach to MSF. Finally, the influence of MSF on promotion rates reveals relatively weak diagnostic value of MSF, but additionally shows that higher, more consistently performing employees are more likely to progress through the levels of management. These findings are discussed with reference to the theoretical underpinnings of performance change and the role of MSF in leadership development.

Primary Findings

Weak performance change. The first overarching contribution estimates the magnitude of change in MSF ratings in a sample that nearly doubles existing meta-analytic estimates, in an organizationally administered annual MSF process, using modern approaches to analyzing change. Slight performance improvement was evidenced

in peers' ratings, and slight declines were evident in self and supervisor ratings. In contrast, Smither et al. (2005) found stronger and positive effects for all three sources. There are a few reasons for the divergence from Smither et al.'s meta-analysis.

First, as emphasized by Smither et al., key contextual variables can impact the efficacy of MSF on performance improvement. In the present study, employees were required to meet with their managers to discuss their feedback, providing an opportunity for counseling and potentially signaling to the employee and manager the importance of the developmental process; both key features of MSF systems thought to increase the potential for change (London & Smither, 2002; Smither et al., 2005). In addition, the scale was tailored to organizational competencies and was separate from administrative performance appraisals.

On the other hand, we used a true longitudinal design spanning over four years and three administrations whereas past research has used relatively short-term time intervals. Given the weaker effects of MSF revealed over longer time intervals (Smither et al., 2005), this might have contributed to the weaker effects. Although short-term performance improvement is certainly an approach to support stronger effects, in organization-wide MSF systems, long-term performance improvement and the ability to diagnose advancement are arguably the key criteria in evaluating system efficacy. In this way, the results of this longitudinal study are potentially more useful for evaluating the efficacy of MSF systems relative to studies of shorter term systems.

The use of LGM and three time intervals further strengthens the results by accounting for measurement error, a lack of invariance, and non-linear change- factors that could potentially cause spurious results when estimating change (cf. Ployhart &

Vandenberg, 2010). As indicated in the supplemental analyses, had we used only two time intervals as has past research, the results would have differed regarding the amount and pattern of change and in a few cases, the direction of change. Thus, our findings are potentially more informative relative to past studies.

Finally, the high mean and weak variability in performance may have attenuated the detection of performance change. However, similar psychometric characteristics seem to be the rule rather than the exception with MSF ratings (Hoffman et al., in press). In other words, although it is possible that the characteristics of the ratings attenuated evidence for change, this is a common issue across MSF scales and thus, does not seem a likely explanation for the differences in results relative to past studies.

Nevertheless, that most employees were evaluated as regularly engaging in the target MSF behaviors it is possible that the null results for change simply indicate that most employees were performing up to standards. Future MSF research attending to approaches to enhance psychometric properties of MSF using recent advances in scale design, rater training, and statistical approaches is clearly a pressing need (Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001; Hoffman, Gorman, Meriac, Blair, Overstreet, & Atchley, in press; Jelley & Goffin, 2001) as well as research attending to rater motivation in MSF settings (Harris, 1994; Murphy, 2004; Tziner, Murphy, & Cleveland, 2005). Whether these findings reflect rater bias or true performance, they point to limitations in the potential usefulness of MSF for most feedback recipients and roadblocks to documenting the usefulness for practitioners. In other words, if most employees are given feedback that they are performing adequately, and performance does not improve following MSF, it is reasonable to question whether the MSF in this sample

was beneficial to feedback recipients and, whether MSF systems, as typically applied, warrant their cost.

Characterizing Antecedents of Change. The next overarching contribution is the articulation of a model of performance change that largely replicated across peer and supervisor raters. Based on this model, engagement in developmental activities and managerial experience had an indirect influence on change through their effect on initial status, and initial status had a negative effect on subsequence change.

First, all rater sources' ratings of initial status demonstrated a moderate to strong negative relationship with performance change suggesting that the weakest performers were those most likely to change. However, contrary to our expectations, the moderation plots for this relationship suggest that across sources high and low initial performers are likely to decrease their performance over time. A close examination of peers' and supervisors' latent classes (Figure 12) provides insight to this finding, showing that the weak growth trajectory evidenced in the baseline LGMs results from the majority (90%) of the sample demonstrating weak change. Yet, a small subset of the sample (6%) demonstrated the pattern proposed by Control Theory (Carver & Scheier, 1982) and the extant MSF literature (Reilly, et al., 1996; Walker & Smither, 1999) - a substantially weaker initial status and a stronger rate of improvement. Thus, those most in need of improvement tended to improve over time, supporting the efficacy of MSF, at least for this subset of managers. MSF practicing organizations then, may consider abiding a general recommendation in the training literature: Target HR efforts only towards those employees in need of improvement (McGehee & Thayer, 1961; Wexely, 1984).

Next, the present model shows that less experienced managers and managers with less participation in voluntary developmental activities had lower performance and, over the course of the study, exhibited linear increases over the four year span of the study. The results of both the peer and supervisor models showed that experienced managers were higher performers initially, and initial higher performers were less likely to show significant increases in performance. On the other hand, less experienced managers tended to receive lower ratings and those with lower initial performance were more likely to show performance improvement. Thus, although performance did not substantially change across administrations, our results point to those individuals particularly likely to benefit from MSF.

Interestingly, TAP enrollment was less strongly related to initial status than was experience, and TAP enrollment actually had a negative effect on change. At first glance, these results may appear contradictory to existing research substantiating a positive relationship between developmental activities and performance improvement (e.g. Luthans & Peterson, 2003; Smither, et al., 2003; Taylor, et al., 2009). However, much of the existing research on developmental activities and MSF focuses on activities that are directly tied to improving MSF ratings (e.g. executive coaching after MSF), rather than activities that assist in long-term career achievement. Enrollees of higher education may be motivated by personal, educational, and long-term professional goals in addition to improving MSF performance. To the extent that more efficacious and motivated individuals seek higher education (Hackett & Betz, 1989; Lent, Brown, & Larkin, 1984, 1986; Pajares & Miller, 1995), the TAP enrollees in the present sample may have generally been internally motivated employees who consistently exhibit strong

performance. Thus, at least in terms of impact on subsequent performance, the tuition program does not seem to be a particularly effective approach to improving employee performance. However, the availability of such programs potential has other benefits, such as increased organizational commitment (Luthans & Peterson, 2003).

Identification and Growth of High Potential Employees. Practically, MSF is thought to provide two core functions: diagnostic value and as an input to employee development and improvement. Our results point to somewhat limited value of MSF in both regards. First, initial status and performance change were relatively unrelated to promotion rate, diverging from past results that support a positive relationship between MSF ratings and ratings of promotability (Gentry, Gilmore, Shuffler, & Leslie, 2012). In contrast, the present study utilized actual promotion decisions as an outcome rather than perceptions of promotability; thus, our results may provide a more realistic estimate of MSF's influence on organizational outcomes. In other words, MSF may not capture information predictive of employees' progression through organizational levels.

The weak promotion relationships evidenced in our main findings are complemented by those of the supplemental analyses. Specifically, the supervisor and peer LCAs showed that the class most likely to be promoted was characterized by comparatively stronger initial status, more managerial experience, and less performance change. Thus, experienced, strong and consistent performers were the most likely to be promoted. These results are aligned with research on dynamic criteria, suggesting that performance consistency incrementally predicts compensation after accounting for maximal performance (Barnes & Morgeson, 2007). An avenue for future research then may be in examining other leader development programs' (e.g. developmental

assessment centers) influence on promotion decisions, which would allow a direct comparison and provide context for the present study's findings.

Together, whereas MSF ratees most in need of improvement were the most likely to improve, MSF ratings were not diagnostic in predicting promotion. To the extent that MSF ratings are not positively related to operationalizations of leader advancement, MSF researchers should question (a) MSF's effectiveness in developing future leaders and (b) the quality of the ratings currently used to characterize developmental needs. Our results suggest that MSF may be better used in practice as a stop gap initiative for managing lower performers, rather than a traditional leader development program. To this end, future research should focus on identifying other distal outcomes of MSF administrations.

Moderation by rater source. The present investigation also extends the existing MSF literature by presenting source-specific estimates of change in performance over time. Results revealed similar underlying models across all three sources, and the supplemental LCAs suggest that all sources' ratings tended to result in comparable classifications of ratees' change over time. Specifically, all sources distinguished a high performing class that demonstrated little change over time, and supervisors and peers distinguished a stronger improving group that started the study with weak performance. Additionally, self raters' perception of a slight decrement in performance over time may indirectly support MSF's purported effect of making self ratings more aligned with others' (Atwater, et al., 1995). Thus, our results provide indirect support of MSF's utility in changing ratees' perceptions, although the change was minimal.

Together, the results suggest that MSF is potentially over-applied and given to large numbers of raters that did not see a detectable benefit. These findings emphasize the

need to consider the developmental needs of managers and to appropriately target administrations to those employees who are most likely to benefit. For instance, inexperienced managers, recently promoted managers, or managers with performance problems might be nominated by their supervisor or anonymously by peers for participation. Other employees might be allowed to voluntarily request feedback. To the extent that novice or low performing employees are the most likely to improve, our results speak to the legitimacy of the continued use and investment in targeted MSF leader development programs. Furthermore, the present study provides practitioners with actionable recommendations for targeting MSF systems to employees who can potentially benefit most.

Limitations and Directions for Future Research

No study is without limitations and ours is no exception. First, although there were several benefits to the present study's design (e.g. longitudinal, large sample, rigorous modeling technique), our study could not account for the ratees' effort to improve on specific areas of performance. Aligned with the tenets of Goal Setting Theory (Locke & Latham, 1990), it is common in MSF administrations to focus ratees' attention onto a single performance dimension or even a critical behavior for subsequent development, rather than the full MSF performance scale. However, the strong relationship exhibited amongst the performance subscales in the current study may have hindered the effectiveness of ratees' goal-setting meetings. Nonetheless, a strong relationship between interpersonal and task-oriented performance measures is an often found pattern in the performance appraisal literature (e.g. Hoffman, et al., 2007). As others have noted, the inability to distinguish between dimensions within the context of

MSF appraisals likely results for a variety of reasons, such as the raters being untrained (Woehr & Huffcutt, 1994), existing working or personal relationships with the target ratees (Hoffman & Baldwin, 2011), or even alternative rating goals (Murphy, et al., 2004; Murphy, 2008; Wong & Kwong, 2007). Thus, although not uncommon to MSF administrations, the ambiguity of performance dimensions may have led to less effective performance goals and ultimately attenuation of performance improvement.

Next, we were only able to concretely determine upward moves through four broad categories of management, potentially introducing range restriction. Future longitudinal research may consider reexamining the influence of MSF initial status and performance change onto promotion rate in an organizational context with clearer delineations between managerial-levels.

A common limitation to any non-fully-crossed research design (i.e., all raters rate all ratees) is the inability to clearly differentiate between variance components in ratings. These methods confound rater effects and rater-ratee interaction effects in performance ratings (Putka, Lance, Le, & McCloy, 2011; Putka, Le, McCloy, & Diaz, 2008). The present study did not model the full multitrait-multimethod structure; however, our primary interest was in examining cross-source differences, consistent with the predominant use of MSF in applied settings. To this point, future research modeling the latent structure of multitrait-multimethod ratings might more effectively isolate sources of performance change.

Next, MSF is at its core a feedback system that increases leaders' knowledge of how to interact effectively with their team. Given the communication that takes place in an MSF administration via rating patterns and written comments, indicators of group

cohesion and team functioning might prove to be more important outcomes in the MSF process. That is, MSF may result in minimal individual-level, behavioral change, but may alternatively help to foster a more collaborative, open organizational culture. For example, a MSF administration might result in minimal improvements on a leader's Communication score; however, the few items that the ratee improved upon may be directly tied to improving communication amongst team-members or subordinates with whom she regularly comes into contact. Alternatively, team-members and followers that acknowledge even minimal change in a leader's behaviors might be more likely to engage in collaborative efforts, knowing their opinions have been heard. Together, future research might investigate the effects of MSF with regards to team-, rather than individual-level measures of performance change.

Finally, future research should continue to examine the reasons for small improvements for this widely adopted leader development tool (Ewen & Edwards, 2001). Although the current study forwards several ratee-specific reasons for weak performance change following MSF, the MSF literature would benefit from further longitudinal research addressing other antecedents of performance change over time (e.g. MSF system characteristics, organizational sponsorship of change; Smither, et al., 2005) in addition to alternative consequences of initial status and performance change (e.g. team cohesion, team performance).

Summary and Conclusion

The present examination nearly doubles the available data on performance change following MSF and does so using a longitudinal design accompanied by state of the art modeling techniques to describe individual growth trajectories and test a model of

performance change. Most centrally, the results paint a less optimistic picture of the efficacy of MSF relative to Smither et al.'s (2005) review by pointing to minimal change in employee performance during the course of the study. In addition, MSF and performance trajectory were largely unrelated to an objective criterion of management success. However these results do not necessarily signal all bad news for MSF practitioners; for less experienced managers and lower performers, evidence of performance improvement following MSF was found. In addition to this key finding, we tested the first theoretical structure of performance improvement following MSF and supported initial status as a direct antecedent of change and manager experience and to a lesser extent TAP participation as distal antecedents of change. We encourage future studies to incorporate modern methods to investigate the dynamics of change and attend to the long-term utility of MSF systems in the identification and development of leaders.

Footnotes

[1] The pre-aggregated 2008 dataset could not be tested as a second order factor measurement model. The results presented in Table 2, therefore, include the second order factor measurement models for 2005 and 2006, but a construct-level measurement model for 2008.

[2] We attempted to examine LCAs using second-order factor models; however, the computing power necessary for these analyses was prohibitive. Provided that these analyses were post hoc and intended to provide a deeper understanding of general forms of change within the data set, we proceeded in examining first order factor LCAs.

REFERENCES

Antonioni, D. (1995). Problems associated with implementation of an effective upward appraisal feedback process: An experimental field study. *Human Resource Development Quarterly, 6*, 157-171.

Arthur, W. A. Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–149.

Atkins, P. W. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*, 871-904. doi: 10.1111/j.1744-6570.2002.tb00133.x

Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings of leadership. *Personnel Psychology, 48*, 35-59. doi: 10.1111/j.1744-6570.1995.tb01745.x

Atwater, L. E., Waldman, D. A., Atwater, D., & Cartier, P. (2000). An upward feedback field experiment: Supervisors' cynicism, reactions, and commitment to subordinates. *Personnel Psychology, 53*, 275-297. doi: 10.1111/j.1744-6570.2000.tb00202.x

Atwater, L. E., & Yammarino, F. J. (1997). Self-other rating agreement: A review and model. In G. R. Ferris (Ed.), *Research in personnel and human resources management, Vol. 15.* (pp. 121-174). US: Elsevier Science/JAI Press.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77,* 836-874. doi: 10.1037/0021-9010.77.6.836

Bailey, C., & Austin, M. (2006). 360 Degree feedback and developmental outcomes: The role of feedback characteristics, self-efficacy and importance of feedback dimensions to focal managers' current role. *International Journal of Selection and Assessment, 14,* 51-66. doi: 10.1111/j.1468-2389.2006.00333.x

Bailey, C., & Fletcher, C. (2002). The impact of multiple source feedback on management development: Findings from a longitudinal study. *Journal of Organizational Behavior, 23*, 853-867. doi: 10.1002/job.167

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance*, *20*, 259-274. doi: 10.1080/08959280701333289

Barnes, C. M., Reb, J., & Ang, D. (2012). More than just the mean: Moving to a dynamic view of performance-based compensation. *Journal of Applied Psychology*. doi: 10.1037/a0026927

Bass, B. M. (1990). *Handbook of leadership: A survey of theory and research*. New York: Free Press.

Bentein, K., Vandenberghe, C., Vandenberg, R., & Stinglhamber, F. (2005). The role of change in the relationship between commitment and turnover: A latent growth modeling approach. *Journal of Applied Psychology, 90*, 468-482. doi: 10.1037/0021-9010.90.3.468

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley.

Bommer, W. H., Johnson, J., Rich, G. A., & Podsakoff, P. M. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605. doi: 10.1111/j.1744-6570.1995.tb01772.x

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1-21. doi: 10.1207/s15327043hup0601_1

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965-973. doi: 10.1037/0021-9010.86.5.965

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-109. doi: 10.1207/s15327043hup1002_3

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement Reconsidered: An Alternative to the r_{wg} Indices. *Organizational Research Methods, 8*, 165-184. doi: 10.1177/1094428105275376

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological Bulletin, 92*, 111-135. doi: 10.1037/0033-2909.92.1.111

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195-212.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255. doi: 10.1207/s15328007sem0902_5

Church, A. H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 82,* 281-292. doi: 10.1037/0021-9010.82.2.281

Church, A. H. (2000). Do higher performing managers actually receive better ratings? A validation of multirater assessment methodology. *Consulting Psychology Journal: Practice and Research. 52.* 99-116.

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22,* 139-162. doi: 10.1177/014920639602200106

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360. doi: 10.1207/s15327043hup1004_2

Conway, J. M., Lombardo, K., & Sanders, K. C. (2001). A meta-analysis of incremental validity and nomological networks for subordinate and peer rating. *Human Performance, 14*, 267-303. doi: 10.1207/s15327043hup1404_1

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Dai, G., De Meuse, K. P., & Peterson, C. (2010). Impact of multi-source feedback on leadership competency development: A longitudinal field study. *Journal of Managerial Issues, 22*, 197-219.

Darr, W., & Catano, V. M. (2008). Multisource assessments of behavioral competencies and selection interview performance. *International Journal of Selection and Assessment, 16*, 68-72. doi: 10.1111/j.1468-2389.2008.00410.x

Day, D. V. (2011). Integrative perspectives on longitudinal investigations of leader development: From childhood through adulthood. *The Leadership Quarterly, 22*, 561-571. doi: 10.1016/j.leaqua.2011.04.012

Day, J. C., Newburger, E. C., Bureau of the Census, W. D. C. E., & Statistics, A. (2002). The big payoff: Educational attainment and synthetic estimates of work-life earnings. Special Studies. Current Population Reports.

Day, D. V., & Sin, H. P. (2011). Longitudinal tests of an integrative model of leader development: Charting and understanding developmental trajectories. *The Leadership Quarterly, 22*, 545-560. doi: 10.1016/j.leaqua.2011.04.011

DeRue, D. S., & Wellman, N. (2009). Developing leaders via experience: The role of developmental challenge, learning orientation, and feedback availability. *Journal of Applied Psychology, 94*, 859-875. doi: 10.1037/a0015317

Dominick, P. G., Reilly, R. R., & Byrne, J. C. (2004, April). *Individual differences and peer feedback: Personality's impact on behavior change.* Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, Illinois.

Dweck, C. S., & Leggett, E. L. (1988). A social– cognitive approach to motivation and personality. *Psychological Review, 95,* 256–273.

Edwards, M. R., & Ewen, A. J. (1996). 360-degree feedback: Royal fail or holy grail? *The Career Development International, 1,* 28-31. doi: 10.1108/13620439610118537

Ewen, A. J., & Edwards, M. R. (2001). Readiness for multisource feedback. In D. W. Bracken, C. W. Timmreck & A. H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes.* (pp. 33-47). San Francisco, CA US: Jossey-Bass.

Fiedler F. E. (1970). Leadership experience and leader performance: Another hypothesis shot to hell. *Organizational Behavior and Human Performance, 5,* 1-14.

Fiedler, F. E. (1992). Time-based measures of leadership experience and organizational performance: A review of research and a preliminary model. *The Leadership Quarterly, 3,* 5-23. doi: 10.1016/1048-9843(92)90003-x

Fleishman, E. A. (1957). A comparative study of aptitude patterns in unskilled and skilled psychomotor performances. *Journal of Applied Psychology, 41*, 263-272. doi: 10.1037/h0041763

Gentry, W. A., Gilmore, D. C., Shuffler, M. L., & Leslie, J. B. (2012). Political skill as an indicator of promotability among multiple rater sources. *Journal of Organizational Behavior, 33*, 89-104. doi: 10.1002/job.740

Gentry, W. A., & Sosik, J. J. (2010). Developmental relationships and managerial promotability in organizations: A multisource study. *Journal of Vocational Behavior, 77,* 266-278. doi: 10.1016/j.jvb.2010.04.010

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. Journal of Vocational Behavior, 18, 326-336.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20,* 737-756. doi: 10.1016/0149-2063(94)90028-0

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43-62. doi: 10.1111/j.1744-6570.1988.tb00631.x

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management, 32,* 325-351.

Hegarty, W. H. (1974). Using subordinate ratings to elicit behavioral changes in supervisors. *Journal of Applied Psychology, 59,* 764-766. doi: 10.1037/h0037507

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Helland, K. R., Hoffman, B. C., & Smith, E. M. (2003, April). *What does it all mean? Differences in multisource ratings.* Presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Hellervik, L. W., Hazucha, J. F., & Schneider, R. J. (1992). Behavior change: Models, methods, and a review of evidence. In M. D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 823–895). Palo Alto, CA: Consulting Psychologists Press.

Heslin, P. A., & Latham, G. P. (2004). The effect of upward feedback on managerial behavior. *Applied Psychology: An International Review, 53,* 23-37. doi: 10.1111/j.1464-0597.2004.00159.x

Hoffman, B. J. & Baldwin, S. P. (2011). *The assessment of managers: A review and integration of assessment center and multisource performance rating research and practice.* In Assessment Centers: A Review. Edited by Novah, N. & Thornton, G. Gower: New York, NY.

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92,* 555-566. doi: 10.1037/0021-9010.92.2.555

Hoffman, B. J., Gorman, A., Meriac, J., Blair, C., Overstreet, B., & Atchley, E.K., (in press). Evidence for the effectiveness of an alternate multi-source feedback measurement methodology. *Personnel Psychology*.

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151. doi: 10.1111/j.1744-6570.2009.01164.x

Hoffman, B. J., & Woehr, D. J. (2009). Disentangling the meaning of multisource performance rating source and dimension factors. *Personnel Psychology, 62,* 735-765. doi: 10.1111/j.1744-6570.2009.01156.x

Hoffman, B. J., Woehr, D. J., Maldagen-Youngjohn, R., & Lyons, B. D. (2011). Great man or great myth? A quantitative review of the relationship between individual differences and leader effectiveness. *Journal of Occupational and Organizational Psychology, 84*, 347-381. doi: 10.1348/096317909x485207

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100-112. doi: 10.1037/0021-9010.88.1.100

Hooijberg, R., & Choi, J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. *The Leadership Quarterly, 11*, 341-364. doi: 10.1016/s1048-9843(00)00044-8

Hunter J. E., Hunter R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85-98.

Jelley, R. B., & Goffin, R. D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology, 86*, 134-144. doi: 10.1037/0021-9010.86.1.134

Johnson, J. W., & Ferstl, K. L. (1999). The effects of interrater and self-other agreement on performance improvement following upward feedback. *Personnel Psychology, 52*, 272-303. doi: 10.1111/j.1744-6570.1999.tb00162.x

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10*, 317–340.

Klimoski, R. J. & Brickner, M. (1987) Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, *40,* 243–260.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284. doi: 10.1037/0033-2909.119.2.254

Kram, K. E. (1985). *Mentoring at work*. Glenview, IL: Scott, Foresman and Company.

Kudisch, J. (1997). Factors related to participant's acceptance of developmental assessment center feedback. *Dissertation Abstracts International*, *58*(6-B), 3349 (UMI No.AAT 9735329).

Lance, C. (2005, April). *Exogenous effects on latent change variables are moderators in disguise.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology. Los Angeles, California.

Lance, C. E., Baxter, D., & Mahan, R. P. (2006). Evaluation of alternative perspectives on source effects in multisource performance measures. In W. Bennett, Jr., C. E. Lance & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges*. (pp. 49-76). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods, 9,* 202-220. doi: 10.1177/1094428105284919

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18,* 223-232. doi: 10.1016/j.hrmr.2008.03.002

Lance, C. E., Meade, A. W., & Williamson, G. M. (2000). We *should* measure change—And here's how. In G. M.Williamson & D. R. Shaffer (Eds.), *Physical illness and depression in older adults: Theory, research, and practice* (pp.201-235). New York: Plenum.

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.

Lent, R. W., Brown, S. D., & Larkin, K. C. (1984). Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology, 31,* 356-362.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23,* 127-135.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ US: Prentice-Hall, Inc.

London M. & Smither J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review,12*, 81-100.

Luthans, F., & Peterson, S. J. (2003). 360 degree feedback with systematic coaching: Empirical analysis suggests a winning combination. *Human Resource Management, 42*, 243-256. doi: 10.1002/hrm.10083

Mann, F. C. (1965). Toward an understanding of the leadership role in formal organizations. In R. Dubin, G. C. Homans, F. C. Mann & D. C. Miller (Eds.), *Leadership and productivity* (pp. 68–103). San Francisco, CA: Chandler.

Martinko, M. J., & Gardner, W. L. (1987). The leader/member attribution process. *Academy of Management Review, 12*, 235-249. doi: 10.2307/258532

Martinko, M. J., Harvey, P., & Douglas, S. C. (2007). The role, function, and contribution of attribution theory to leadership: A review. *The Leadership Quarterly, 18*, 561-585. doi: 10.1016/j.leaqua.2007.09.004

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology, 73*, 327-330.

McGehee, W., & Thayer, P. W. (1961). *Training in business and industry*. Oxford England: Wiley.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107-122.

Morris, S. B., & DeShon R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Journal of Applied Psychology. 7*, 105-125.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148-160. doi: 10.1111/j.1754-9434.2008.00030.x

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA US: Sage Publications, Inc.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*, 158-164. doi: 10.1037/0021-9010.89.1.158

Muthén, L. K. and Muthén, B. O. (1998-2007*). Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Nemeroff, W. F., & Cosentino, J. (1979). Utilizing feedback and goal setting to increase performance appraisal interviewer skills of managers. *Academy of Management Journal, 22*, 566-576. doi: 10.2307/255745

Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences.* (pp. 7-36). New York, NY US: Routledge/Taylor & Francis Group.

Nilsen, D., & Campbell, D. P. (1993). Self-observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management, 32*, 265-281.

Noe, R. A. (1988). An investigation of the determinants of successful assigned mentoring relationships. *Personnel Psychology, 41,* 457-479.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling, 14*, 535-569.

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology, 57,* 333-375. doi: 10.1111/j.1744-6570.2004.tb02494.x

Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology, 42*, 190-198.

Pastor, D. A., Barron, K. E. Miller, B. J. & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32*, 8-47.

Ployhart, R. E. & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7*, 27-65.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36,* 94-120. doi: 10.1177/0149206309352110

Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A cautionary note on modeling multitrait-multirater data arising from ill-structured measurement

designs. *Organizational Research Methods, 14,* 503-529. doi: 10.1177/1094428110362107

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93,* 959-981. doi: 10.1037/0021-9010.93.5.959

Quiñones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology, 48,* 887-910. doi: 10.1111/j.1744-6570.1995.tb01785.x

Reb, J., & Greguras, G. J. (2010). Understanding performance ratings: Dynamic performance, attributions, and rating purpose. *Journal of Applied Psychology, 95,* 213-220. doi: 10.1037/a0017237

Reilly, R. R., Smither, J. W., & Vasilopoulos, N. L. (1996). A longitudinal study of upward feedback. *Personnel Psychology, 49,* 599-612. doi: 10.1111/j.1744-6570.1996.tb01586.x

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change.* (pp. 3-66). Mahwah, NJ: Lawrence Erlbaum.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90,* 726-748.

Rosti Jr., R. T., & Shipper, F. (1998). A study of the impact of training in a management development program based on 360 feedback. *Journal of Managerial Psychology, 13,* 77.

Rupp, D. E., Snyder, L. A., Gibbons, A. M., & Thornton, G. C., III. (2006). What should developmental assessment centers be developing? *Psychologist-Manager Journal, 9,* 75-98. doi: 10.1207/s15503461tpmj0902_3

Seifert, C. F., & Yukl, G. (2010). Effects of repeated multi-source feedback on the influence behavior and effectiveness of managers: A field experiment. *The Leadership Quarterly, 21,* 856-866. doi: 10.1016/j.leaqua.2010.07.012

Seifert, C. F., Yukl, G., & McDonald, R. A. (2003). Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. *Journal of Applied Psychology, 88,* 561-569. doi: 10.1037/0021-9010.88.3.561

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43,* 101-116.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY US: Oxford University Press.

Smither, J. W., London, M., Flautt, R., Vargas, Y., & Kucine, I. (2003). Can working with an executive coach improve multisource feedback ratings over time? A quasi-experimental field study. *Personnel Psychology, 56,* 23-44. doi: 10.1111/j.1744-6570.2003.tb00142.x

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58,* 33-66. doi: 10.1111/j.1744-6570.2005.514_1.x

Smither, J. W., London, M., Vasilopoulos, N. L., & Reilly, R. R. (1995). An examination of the effects of an upward feedback program over time. *Personnel Psychology, 48,* 1-34. doi: 10.1111/j.1744-6570.1995.tb01744.x

Smither, J. W., & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time? *Journal of Applied Psychology, 89,* 575-581. doi: 10.1037/0021-9010.89.3.575

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Transfer of management training from alternative perspectives. *Journal of Applied Psychology, 94,* 104-121. doi: 10.1037/a0013006

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and Rater Factors Affecting Rating Behavior. *Group & Organization Management, 30*, 89-98. doi: 10.1177/1059601104267920

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. doi: 10.1177/109442810031002

Vandenberg, R. J., & Self, R. M. (1993). On assessing newcomers' changing commitments to the organization during the first six months of work. *Journal of Applied Psychology,* 78, 557‐568

Vicere, A. A. & Fulmer, R. M. (1998). *Leadership by design: How benchmark companies sustain success through investment in continuous learning*. Boston: Harvard Business Review.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108-131. doi: 10.1037/0021-9010.90.1.108

Waldman, D. A., & Atwater, L. E. (2001). Attitudinal and behavioral outcomes of an upward feedback process. *Group & Organization Management, 26,* 189-205. doi: 10.1177/1059601101262004

Walker, A. G., & Smither, J. W. (1999). A five-year study of upward feedback: What managers do with their results matters. *Personnel Psychology, 52,* 393-423. doi: 10.1111/j.1744-6570.1999.tb00166.x

Warech, M. A., Smither, J. W., Reilly, R. R., Millsap, R. E., & Reilly, S. P. (1998). Self-monitoring and 360-degree ratings. *Leadership Quarterly, 9,* 449-473.

Wexley, K. N. (1984). Personnel training. *Annual Review of Psychology, 35,* 519-551. doi: 10.1146/annurev.ps.35.020184.002511

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551. doi: 10.1111/j.1744-6570.1982.tb02208.x

Williams, L. J., Edwards, J. R., & Vandenberg, R. J. (2003). Recent advances in causal modeling methods for organizational and management research. *Journal of Management, 29,* 903-936. doi: 10.1016/s0149-2063(03)00084-9

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189-205. doi: 10.1111/j.2044-8325.1994.tb00562.x

Woehr, D. J., Sheehan, M. K., & Bennett, W., Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90,* 592-600. doi: 10.1037/0021-9010.90.3.592

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92,* 577-585. doi: 10.1037/0021-9010.92.2.577

Woo, S. E., Sims, C. S., Rupp, D. E., & Gibbons, A. M. (2008). Development engagement within and following developmental assessment centers: Considering feedback favorability and self-assessor agreement. *Personnel Psychology, 61,* 727-759. doi: 10.1111/j.1744-6570.2008.00129.x

Yukl, G. A. (2006). *Leadership in organizations* (6th ed.). Upper Saddle River, NJ: Pearson, Prentice-Hall.

Zaccaro, S. J., Gilbert, J. A., Thor, K. K., & Mumford, M. D. (1991). Leadership and social intelligence: Linking social perspectives and behavioral flexibility to leader effectiveness. *The Leadership Quarterly, 2,* 317-342. doi: 10.1016/1048-9843(91)90018-w

Zimmerman, R. D., Mount, M. K., & Goff, M., III. (2008). Multisource feedback and leaders' goal performance: Moderating effects of rating purpose, rater perspective, and performance dimension. *International Journal of Selection and Assessment, 16,* 121-133. doi: 10.1111/j.1468-2389.2008.00417.x

APPENDICES

Appendix A
Performance Dimension Definitions

Performance Dimension	Definition
1. Problem Analysis	Identifies, solves, anticipates and prevents problems. Identifies and assesses alternatives, determines solutions, and communicates rationale, implications and approach for implementing solutions. Uses judgment and discretion to make decisions and take actions that are sound and timely.
2. Results Orientation	Performs above and beyond others' expectations. Consistently delivers results and high quality work on schedule. Goes beyond role expectations.
3. Quality Improvement	Identifies process failures and the means by which to improve them. Recognizes the importance of measuring success and sets quality improvement objectives as well as measurement standards to characterize quality/process improvement.
4. Relationship Skills	Develops strong working relationships with team members and coworkers across business groups and functions. Values getting to know others and recognizes business decisions' impact on others. Understands and values differences and diversity of others.
5. Teamwork	Values and promotes teamwork in the workplace and across business groups and functions. Recognizes the benefits of approaching organizational goals via team-based work and encourages team-oriented solutions when appropriate.
6. Communication	Clearly, confidently and professionally communicates with others yet also listens attentively and respectfully. Ensures others have the required information to be effective in their position. Seeks input, perspective and clarification through information exchange, dialogue and asking questions.
7. Managing Conflict	Proactively takes action to resolve conflict situations by encouraging others to express disagreements and seeking out resolutions or comprises. Responds constructively and in a timely manner to conflict situations.
8. Developing Others	Recognizes and maintains people development as a priority. Actively ensures that our people are continually developed for current and future roles in the organization. Clearly defines performance standards for others and identifies solutions, strategies, or developmental opportunities to improve performance.

Figure 1
Structural model

Moderators	
Rater Source	*Self*
	Supervisory
	Peers
Performance Dimension	*Task Performance*
	Interpersonal Performance

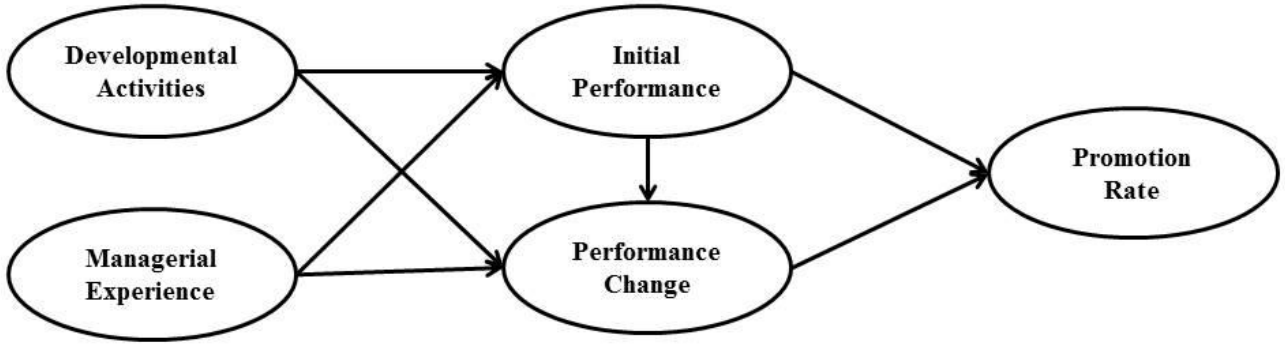
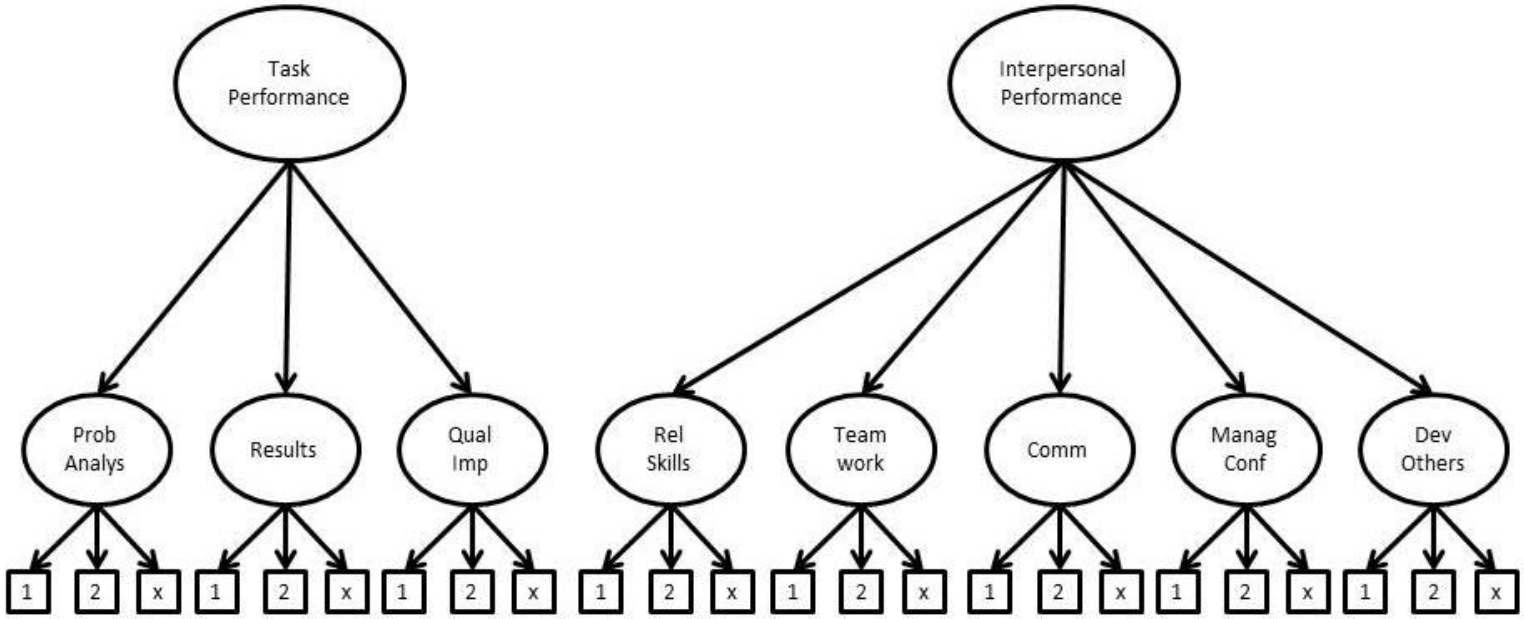
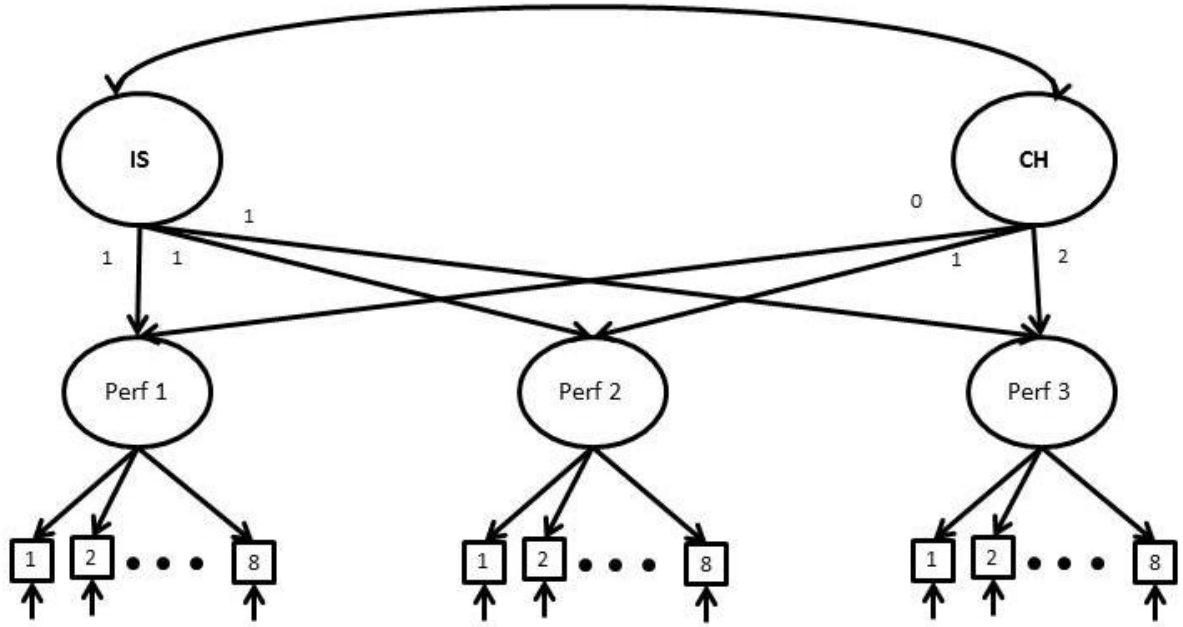


Figure 2
Theoretical Measurement Model



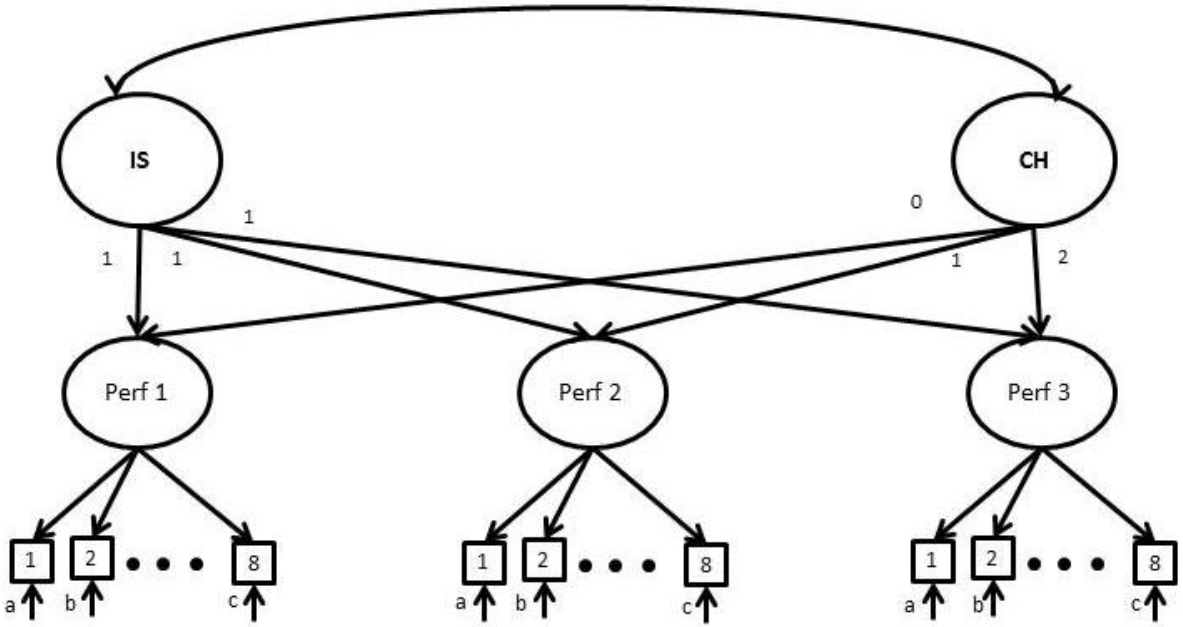
Note:Prob Analys denotes Problem Solving and Analysis; Qual Imp denotes Quality Improvement; Rel Skills denotes Relationship Skills; Comm denotes Communication; Manag Conf denotes Managing Conflict; Dev Others denotes Developing Others; boxes represent the 37 manifest indicators, the eight circles represent the eight performance constructs, and the two larger circles represent the two overarching domains of performance.

Figure 3
Latent Growth Model, Heterogenous Uniquesses Across Time



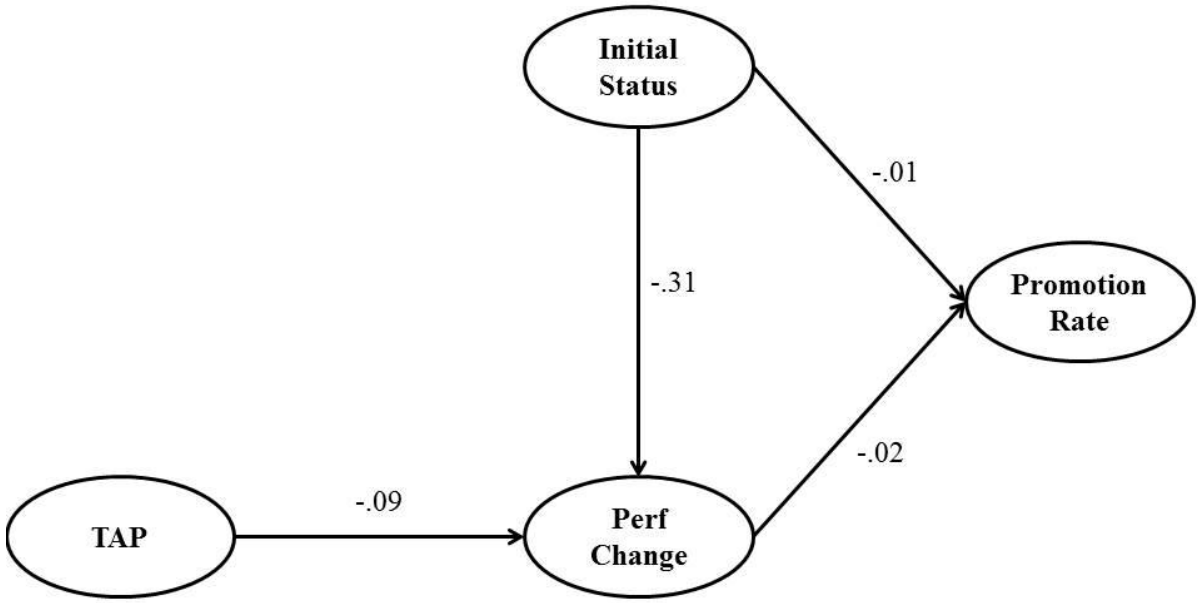
Note: IS denotes Initial Status; CH denotes Change; Perf1-3 denotes Overall Performance measured at 2005, 2006, and 2008, respectively; Squares numbered 1-8 denote the eight performance constructs comprising Overall Performance at each time point; Vertical arrows below each performance construct denotes heterogeneously estimated uniquenesses for each performance construct.

Figure 4
Latent Growth Model, Homogenous Uniquesses Across Time



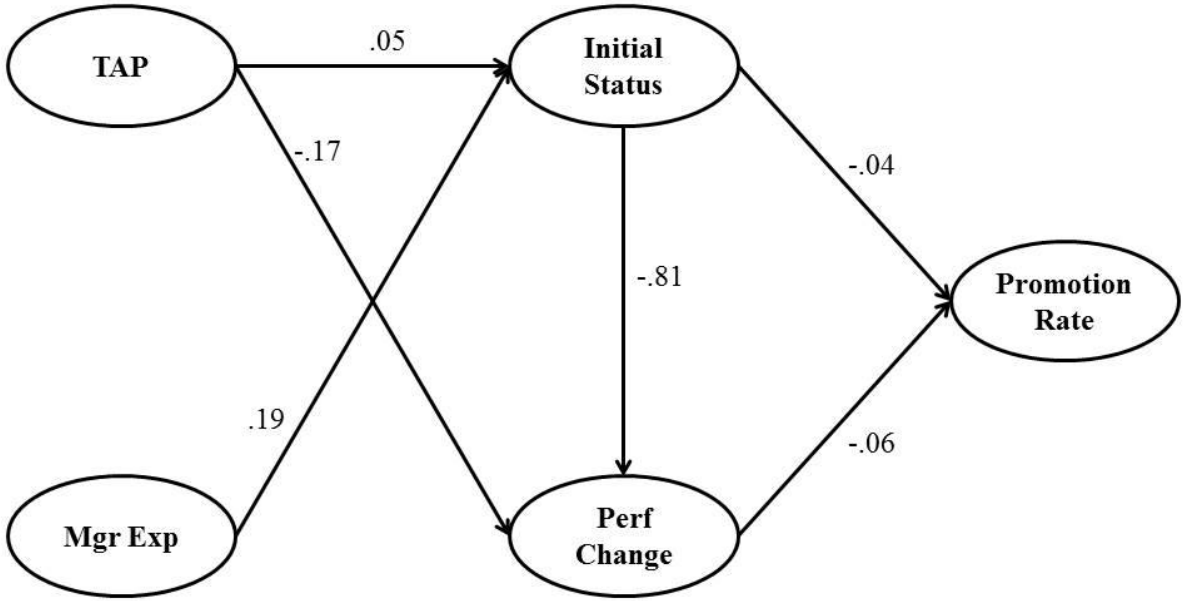
Note: IS denotes Initial Status; CH denotes Change; Perf1-3 denotes Overall Performance measured at 2005, 2006, and 2008, respectively; Squares numbered 1-8 denote the eight performance constructs comprising Overall Performance at each time point; A, B, and C each denote the homogenously estimated uniqueness of each of the eight performance constructs at each time point.

Figure 5
Supported Structural Model for Self Raters



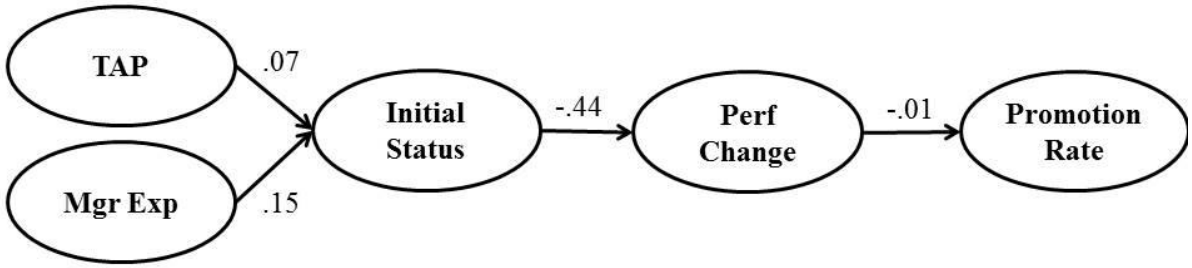
Note: All parameter estimates were statistically significant (p<.05); TAP denotes Tuition Assistance Program; Perf Change denotes performance change.

Figure 6
Supported Structural Model for Peer Raters



Note: All parameter estimates were statistically significant (p<.05); TAP denotes Tuition Assistance Program; Perf Change denotes performance change.

Figure 7
Supported Structural Model for Supervisor Raters



Note: All parameter estimates were statistically significant (p<.05); TAP denotes Tuition Assistance Program; Perf Change denotes performance change.

Figure 8
Self Raters: IS Moderating Performance Change

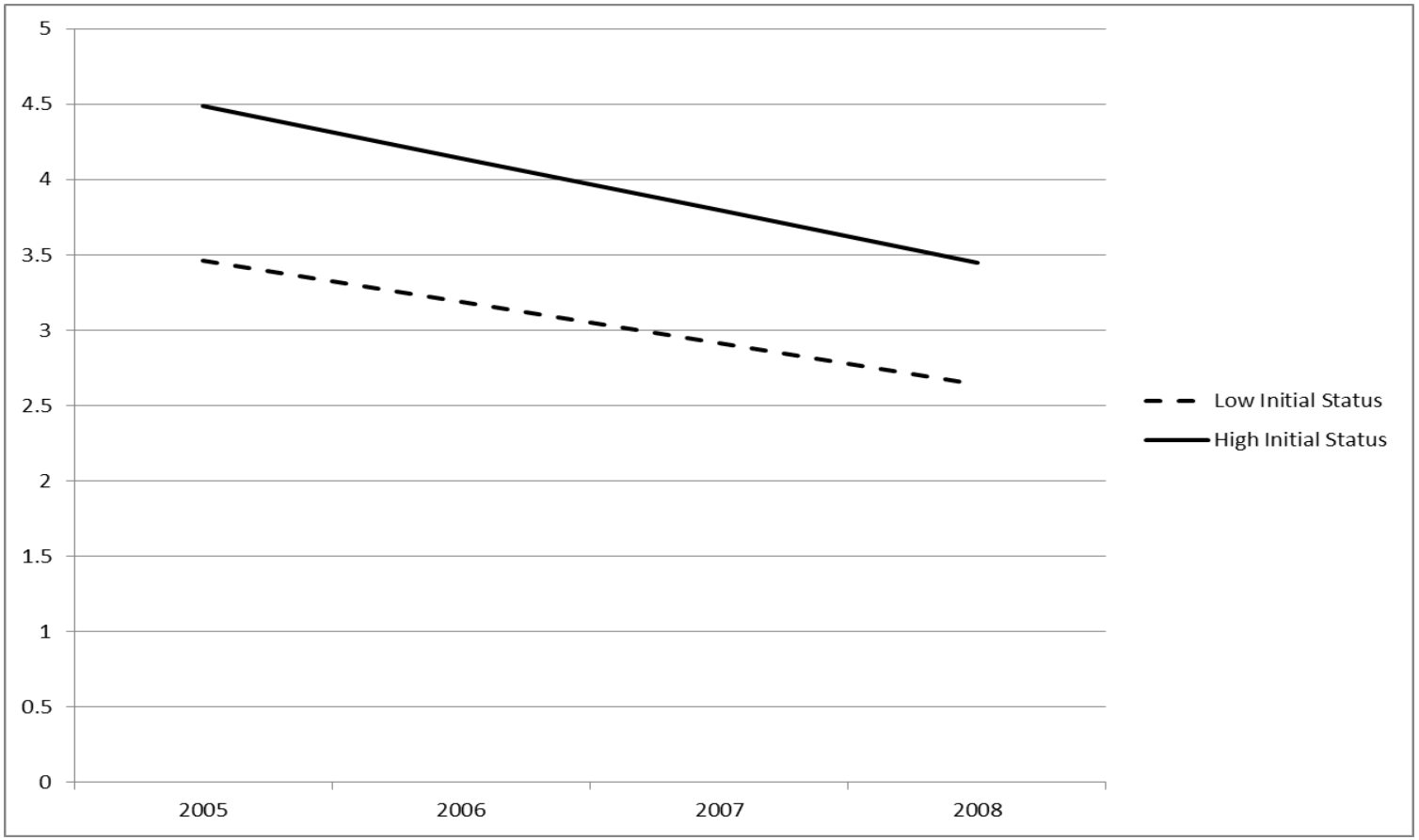


Figure 9
Peers: IS Moderating Performance Change

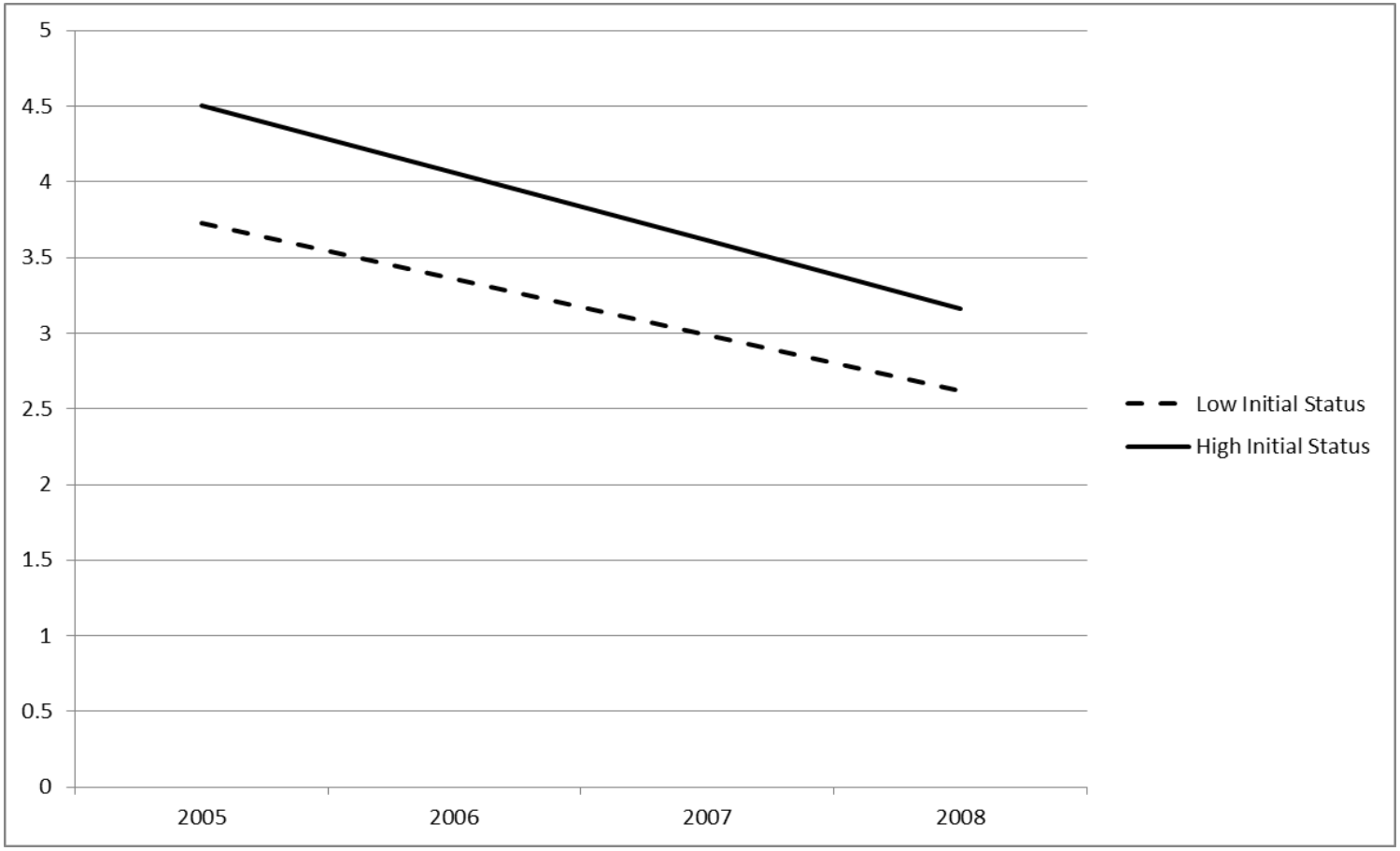


Figure 10
Supervisors: IS Moderating Performance Change

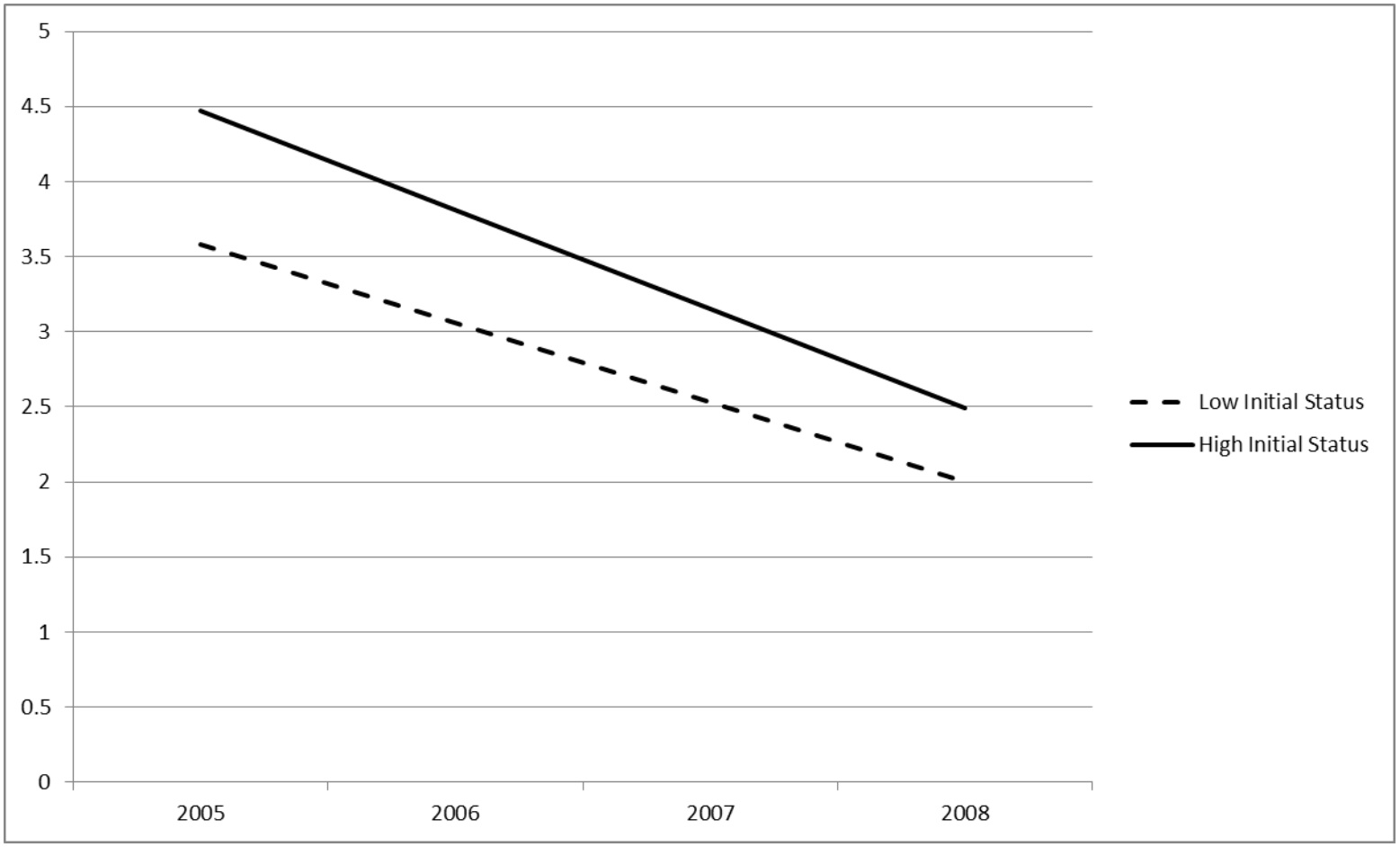


Figure 11
Self Raters: TAP Moderating Performance Change

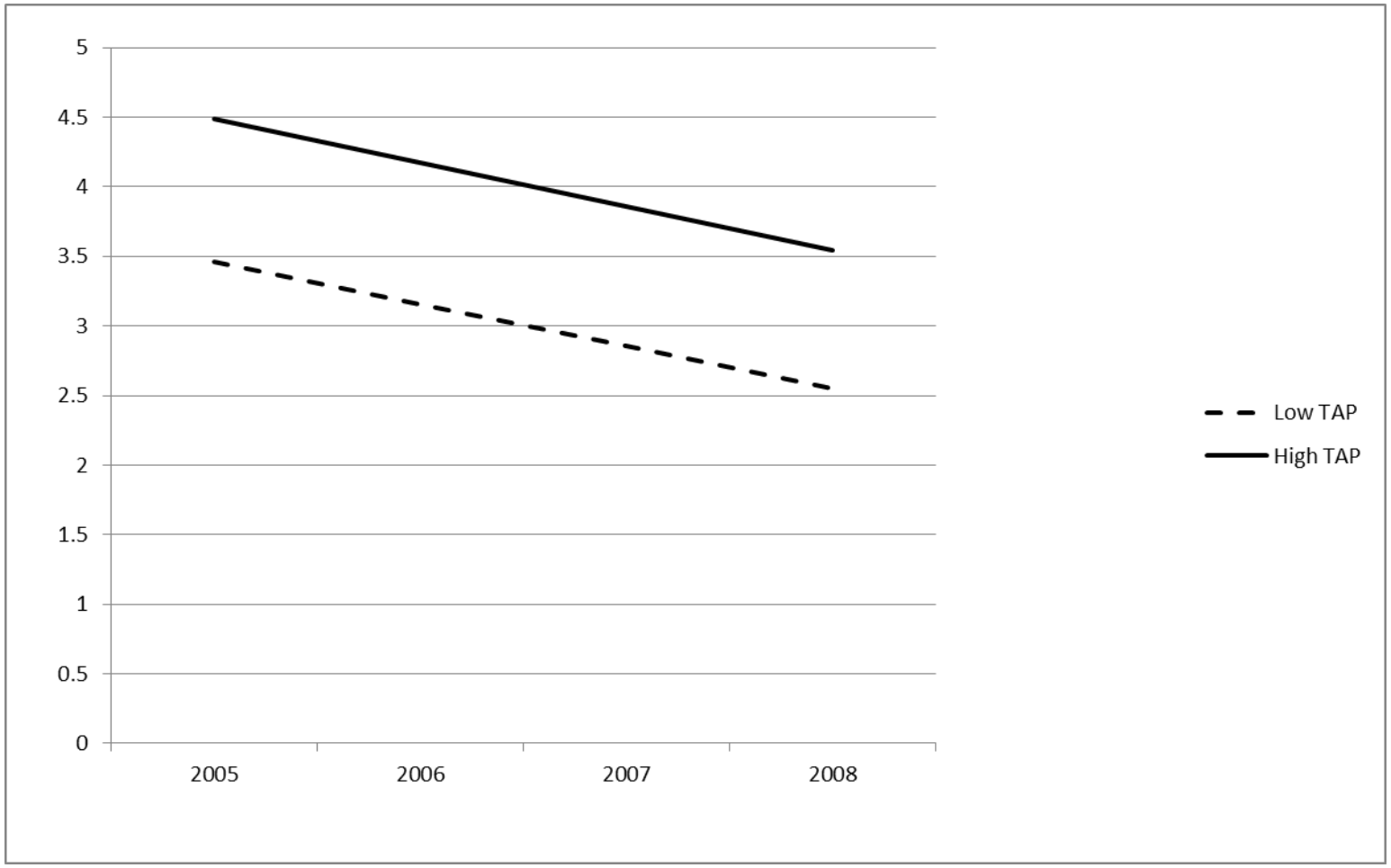


Figure 12
Peers: TAP Moderating Performance Change

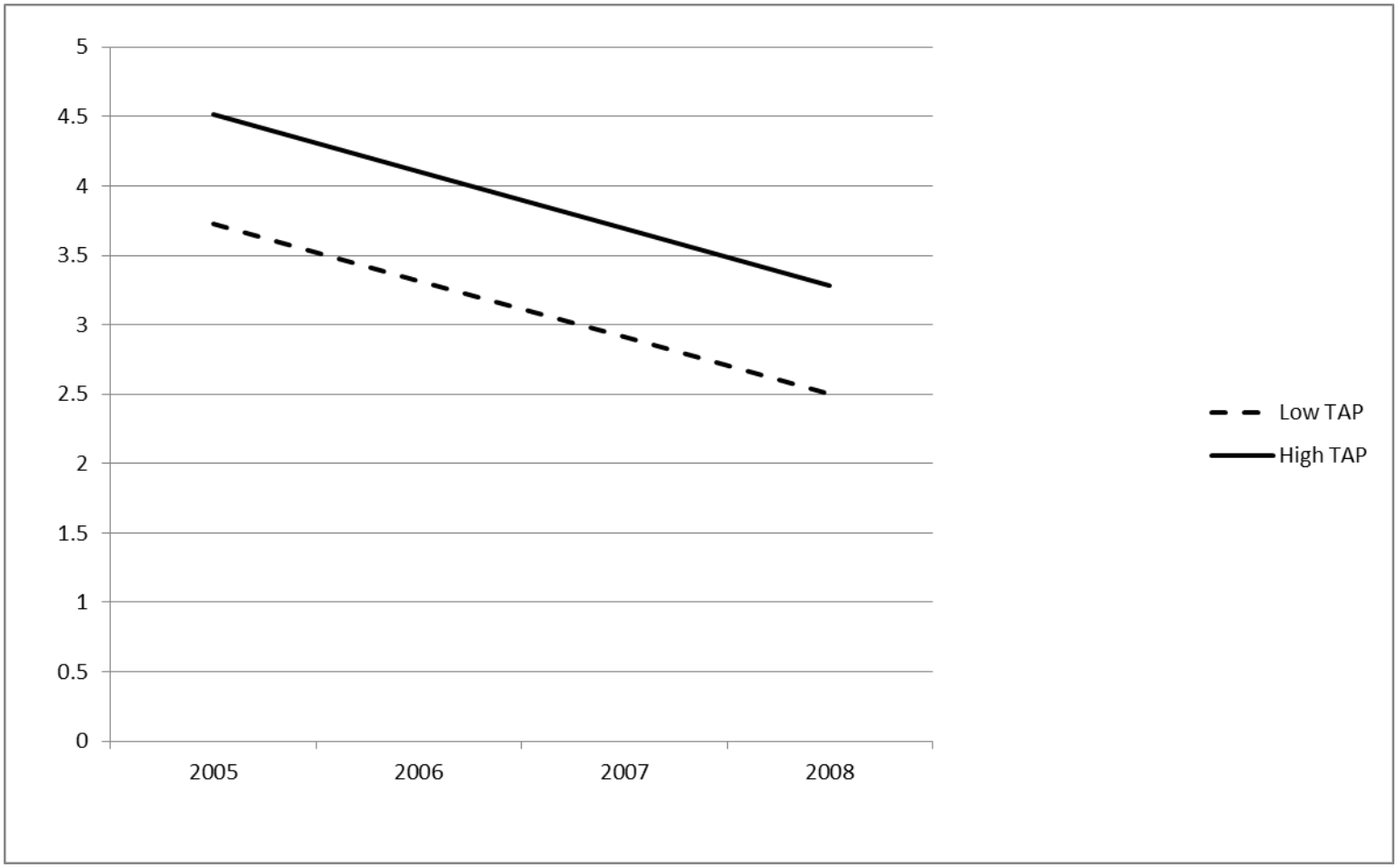


Figure 13
Peers: Growth Trajectories of a Two Class LCA

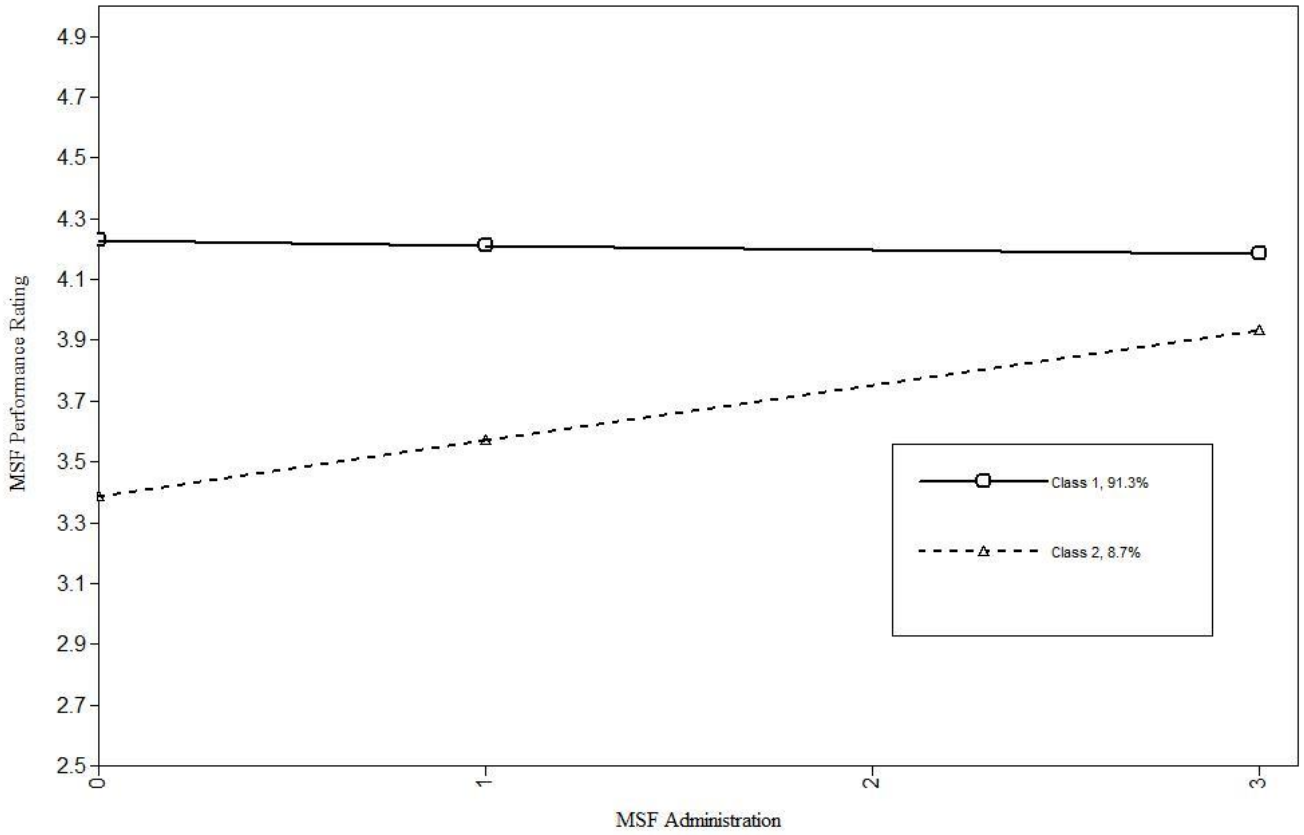


Figure 14
Supervisors: Growth Trajectories of a Two Class LCA

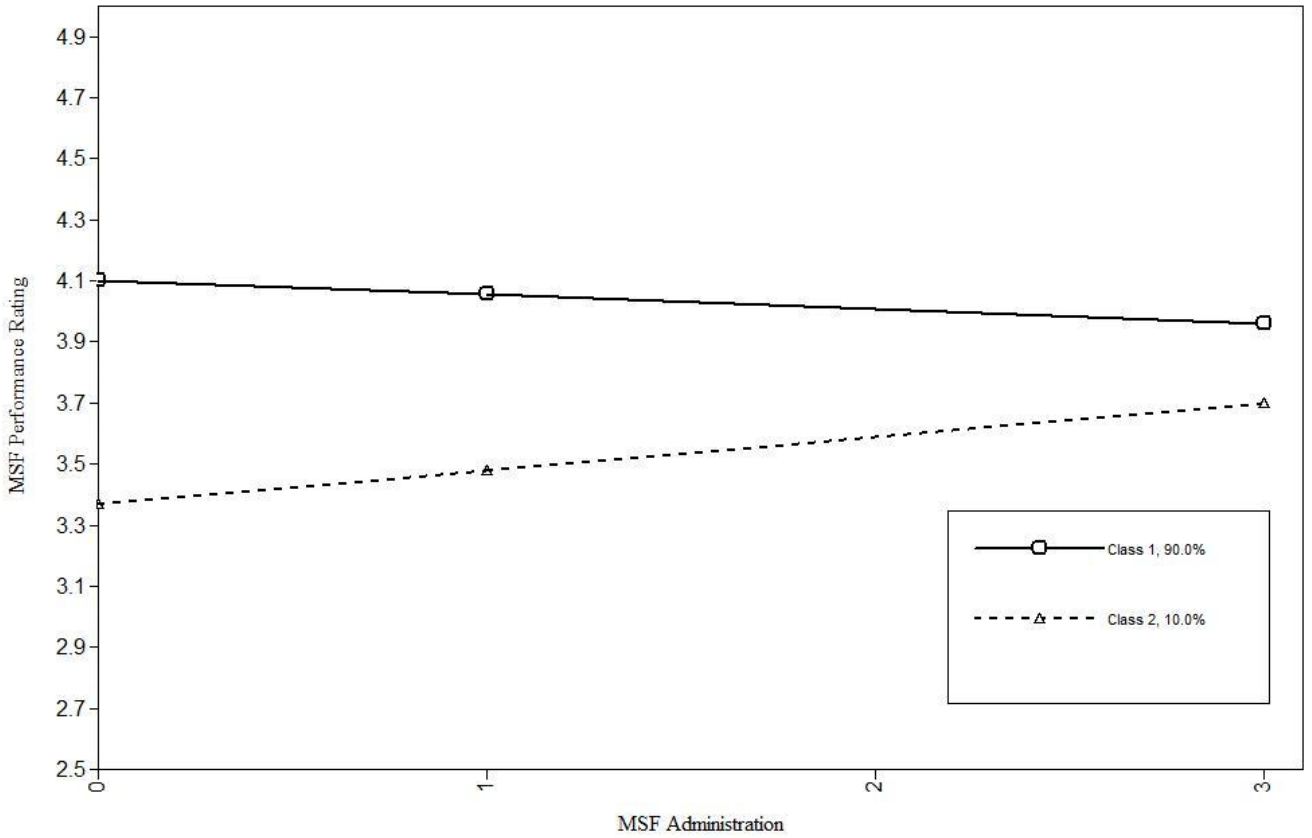


Table 1
Sample Description for 2005, 2006, and 2008

	N
2005 MSF Administration – Non aggregated	
Self Raters	5,128
Peer Raters	20,841
Supervisor Raters	5,162
Indirect Supervisor Raters	1,428
Total Raters	**32,559**
2006 MSF Administration – Non aggregated	
Self Raters	5,128
Peer Raters	21,363
Supervisor Raters	5,168
Indirect Supervisor Raters	1,422
Total Raters	**33,081**
2008 MSF Administration – Aggregated	
Self Raters	5,128
Peer Raters	5,128
Supervisor Raters	5,128
Minimum Raters	**15,384**

Note: The 2008 MSF administration data was only available in aggregate form.

Table 2
Measurement Model Fit Statistics.

	χ^2	df	RMSEA	CFI	TLI	SRMR
2005 MSF Administration – Second Order Factor Model						
1. Self, One Factor Model	18543.42	629	.075	.867	.859	.042
2. Self, Two Factor Model	18458.43	628	.074	.867	.859	.042
Model 1 vs. Model 2	(84.99**)	(1)		(0)		
3. Peers, One Factor Model	28389.43	629	.093	.892	.886	.03
4. Peers, Two Factor Model	28342.45	628	.093	.892	.886	.03
Model 3 vs. Model 4	(46.99**)	(1)		(0)		
5. Supervisors One Factor Model	29803.46	629	.095	.854	.846	.042
6. Supervisors Two Factor Model	29735.82	628	.095	.855	.846	.042
Model 5 vs. Model 6	(67.65**)	(1)		(-.001)		
2006 MSF Administration – Second Order Factor Model						
7. Self, One Factor Model	20044.25	629	.078	.861	.853	.043
8. Self, Two Factor Model	19986.17	628	.078	.861	.853	.042
Model 7 vs. Model 8	(58.08**)	(1)		(0)		
9. Peers, One Factor Model	28808.98	629	.093	.896	.89	.028
10. Peers, Two Factor Model	28750.85	628	.093	.896	.89	.028
Model 9 vs. Model 10	(58.14**)	(1)		(0)		
11. Supervisors One Factor Model	25648.67	629	.088	.839	.83	.047
12. Supervisors Two Factor Model	25597.28	628	.088	.839	.83	.046
Model 11 vs. Model 12	(51.40**)	(1)		(0)		
2008 MSF Administration – First order Factor Model						
13. Self, One Factor Model	1068.96	20	.101	.973	.962	.02
14. Self, Two Factor Model	604.92	19	.078	.985	.978	.016
Model 13 vs. Model 14	(464.04**)	(1)		(-.012)		
15. Peers, One Factor Model	2262.19	20	.148	.964	.95	.016
16. Peers, Two Factor Model	1370.31	19	.118	.978	.968	.013
Model 15 vs. Model 16	(891.87**)	(1)		(-.014)		
17. Supervisors One Factor Model	1800.44	20	.132	.955	.937	.027
18. Supervisors Two Factor Model	1219.31	19	.111	.97	.955	.023
Model 17 vs. Model 18	(581.13**)	(1)		(-.015)		

*Note: *p<.05; **p<.001; Parenthetical values represent difference values between the comparison models; RMSEA denotes Root Mean Square Error of Approximation; TLI denotes Tucker Lewis Index; CFI denotes Comparative Fit Index; SRMR denotes Standardized Root Mean Square Residual*

Table 3
Interrater Agreement Statistics

	Average $r^*_{wg(j)}$	SD
2005 MSF Administration		
Peers	.89	.15
Supervisors	.94	.10
2006 MSF Administration		
Peers	.89	.13
Supervisors	.93	.13

Note: SD denotes standard deviation.

Table 4
Correlations and Descriptive Statistics among Study Variables

	M	SD	1	2	3	4	5	6	7	8	9	10	11
1. Self 2005	4.01	0.50	-										
2. Peer 2005	4.14	0.42	.27	-									
3. Supervisor 2005	4.02	0.42	.36	.70	-								
4. Self 2006	4.03	0.49	.61	.21	.27	-							
5. Peer 2006	4.17	0.41	.19	.52	.46	.24	-						
6. Supervisor 2006	4.01	0.54	.12	.30	.46	.21	.41	-					
7. Self 2008	3.99	0.49	.51	.15	.18	.58	.16	.15	-				
8. Peer 2008	4.16	0.40	.17	.39	.32	.17	.40	.24	.26	-			
9. Supervisor 2008	3.93	0.53	.09	.26	.31	.10	.24	.34	.20	.38	-		
10. Experience 2005	8.02	9.88	-.01	.16	.13	.00	.15	.07	-.03	.08	.04	-	
11. TAP 2008	2.20	5.08	-.03	.03	.07	-.04	.00	.03	-.06	-.03	.03	-.03	-
12. Promotion 2008	.15	.37	-.02	.02	.04	-.04	.00	.00	-.09	-.03	-.01	-.03	.09

Note: |r values| >.03, p<.05; |r values| >.04, p<.01; M denotes mean; SD denotes Standard deviation; Experience denotes Managerial Experience; TAP denotes enrollment in the Tuition Assistance Program.

Table 5
Measurement Invariance Tests

	χ^2	df	RMSEA	CFI	TLI	SRMR
Self Raters						
1. Configural Invariance	3,323.45	225	.05	.978	.97	.01
2. Metric Invariance	3,815.91	239	.05	.974	.97	.06
Model 1 vs. Model 2	(492.46**)	(14)		(-.004)		
3. Scalar Invariance	8,011.55	255	.077	.944	.940	.092
Model 2 vs. Model 3	(4,195.64**)	(16)		(-.030)		
4. Partial Scalar Invariance	4,676.41	249	.059	.968	.965	.064
Model 3 vs. Model 4	(860.50**)	(10)		(-.006)		
Peer Raters						
5. Configural Invariance	5,071.00	225	0.07	.978	.97	.01
6. Metric Invariance	5,435.67	239	0.07	.976	.97	.05
Model 5 vs. Model 6	(364.67**)	(14)		(-.002)		
7. Scalar Invariance	9,235.95	255	.083	.959	.956	.060
Model 6 vs. Model 7	(3,800.28**)	(16)		(-.017)		
8. Partial Scalar Invariance	7,188.73	249	.074	.968	.965	.050
Model 7 vs. Model 8	(1,753.06**)	(10)		(-.008)		
Supervisor Raters						
9. Configural Invariance	4,592.20	225	.06	.972	.97	.01
10. Metric Invariance	5,340.48	239	.07	.967	.96	.07
Model 9 vs. Model 10	(748.28**)	(14)		(-.005)		
11. Scalar Invariance	9,325.33	255	.083	.942	.937	.082
Model 10 vs. Model 11	(3,984.85**)	(16)		(-.025)		
12. Partial Scalar Invariance	6,368.59	249	.069	.961	.956	.074
Model 11 vs. Model 12	(1,028.11**)	(10)		(-.006)		

*Note: *p<.05; **p<.001; Parenthetical values represent difference values between the comparison models; RMSEA denotes Root Mean Square Error of Approximation; TLI denotes Tucker Lewis Index; CFI denotes Comparative Fit Index; SRMR denotes Standardized Root Mean Square Residual*

Table 6

Fit Statistics and Change and Initial Status Estimates for Latent Growth Models

	χ^2	df	RMSEA	CFI	TLI	SRMR
Self Raters						
1. Linear, Heterogeneous	4,717.37	252	.059	.968	.965	.071
2. Linear, Homogeneous	8,259.58	268	.076	.943	.941	.091
Model 1 vs. Model 2	(3,542.21**)	(16)		(-.03)		
Peers						
3. Linear, Heterogeneous	7,456.74	252	.075	.967	.964	.060
4. Linear, Homogeneous	13,581.87	268	.098	.939	.938	.073
Model 3 vs. Model 4	(6,125.13**)	(16)		(-.03)		
Supervisors						
5. Linear, Heterogeneous	6,496.80	252	.070	.960	.956	.081
6. Linear, Homogeneous	23,313.15	268	.129	.852	.848	.153
Model 5 vs. Model 6	(16,816.35**)	(16)		(-.11)		

*Note: *p<.05; **p<.001; Parenthetical values represent difference values between the comparison models; RMSEA denotes Root Mean Square Error of Approximation; TLI denotes Tucker Lewis Index; CFI denotes Comparative Fit Index; SRMR denotes Standardized Root Mean Square Residual; IS denotes initial status; IS Var. denotes the variance estimate for initial status; CH denotes performance change; CH Var. denotes the variance estimate for performance change.*

Table 7
Initial Status and Change Parameter Estimates and Confidence Intervals

	IS	IS CIs	IS Var.	IS Var. CIs	CH	CH CIs	CH Var.	CH Var. CIs
Baseline LGM								
1. Self Raters	3.97	3.93 – 4.01	.153	.144 - .163	-.02**	-.027 - -.018	.009**	.007 - .012
2. Peers	4.06	4.04 – 4.09	.097	.09 - .103	.004*	.001 - .008	.004**	.002 - .006
3. Supervisors	3.98	3.97 – 4.00	.118	.109 - .128	-.002,*ns*	-.007 - .003	.014**	.011 - .017

*Note: *p<.05; **p<.001; Parenthetical values represent difference values between the comparison models; RMSEA denotes Root Mean Square Error of Approximation; TLI denotes Tucker Lewis Index; CFI denotes Comparative Fit Index; SRMR denotes Standardized Root Mean Square Residual; IS denotes initial status; IS Var. denotes the variance estimate for initial status; CH denotes performance change; CH Var. denotes the variance estimate for performance change; CI denotes Confidence Interval*

Table 8

Model Fit Statistics for Structural Latent Growth Models

	χ^2	df	RMSEA	CFI	TLI	SRMR
Self Raters						
1. Measurement	5,059.89	327	.053	.966	.964	.066
2. Structural	4,996.44	320	.053	.966	.963	.064
3. Structural - parsimonious	5,001.90	324	.053	.966	.964	.064
Model 2 vs. Model 3	(5.46, *ns*)	(4)		(0)		
Peer Raters						
4. Measurement	7,874.48	327	.067	.966	.963	.062
5. Structural	7,687.39	320	.067	.966	.963	.055
6. Structural – parsimonious	7,688.83	321	.067	.966	.963	.055
Model 5 vs. Model 6	(1.43, *ns*)	(1)		(0)		
Supervisor Raters						
7. Measurement	6,931.77	327	.063	.958	.955	.076
8. Structural	6,797.31	320	.063	.958	.955	.073
9. Structural - parsimonious	6,543.29	322	.061	.96	.957	.070
Model 8 vs. Model 9	(254.02**)	(2)		(-.002)		

*Note: *p<.05; **p<.001; Parenthetical values represent difference values between the comparison models; RMSEA denotes Root Mean Square Error of Approximation; TLI denotes Tucker Lewis Index; CFI denotes Comparative Fit Index; SRMR denotes Standardized Root Mean Square Residual;*

Table 9
Parameter Estimates for Theoretical and Supported Structural LGMs

	Exp-IS	**Exp-CH**	**TAP-IS**	**TAP-CH**	**IS-CH**	**IS-PROMO**	**CH-PROMO**
Parameter Estimates for Structural Models							
Self, Theoretical	.003, *ns*	-.03, *ns*	-.01, *ns*	-.09	-.31	-.01	-.02
Self, Supported	-	-	-	-.09	-.31	-.01	-.02
Peers, Theoretical	.19	.03, *ns*	.04	-.14	-.85	-.05	-.07
Peers, Supported	.19	-	.05	-.17	-.81	-.04	-.06
Supervisors, Theoretical	.15	-.01, *ns*	.08	-.01, *ns*	-.43	.001	-.01
Supervisors, Supported	.15	-	.07	-	-.44	-	-.01

Note: Exp denotes Managerial Experience; TAP denotes Tuition Assistance Program; PROMO denotes Promotion Rate; IS denotes Initial Status; CH denotes Performance Change.

Table 10
Model Fit Statistics for LCAs

	BIC	Adjusted BIC	Entropy
Self Raters			
2 Latent Classes	12,009.30	11,974.35	.98
3 Latent Classes	11,985.70	11,941.22	.83
4 Latent Classes	11,700.71	11,646.69	.79
5 Latent Classes	11,677.76	11,614.21	.81
6 Latent Classes	11,663.52	11,590.43	.80
Peer Raters			
2 Latent Classes	**9,443.98**	**9,409.02**	**.79**
Supervisor Raters			
2 Latent Classes	**13,609.79**	**13,574.84**	**.71**
3 Latent Classes	13,613.29	13,568.80	.79
4 Latent Classes	13,613.05	13,559.03	.68
5 Latent Classes	13,524.42	13,460.87	.73
6 Latent Classes	13,504.88	13,431.80	.72

Note: Bolded class denotes the supported class structure

Table 11
Descriptive Statistics for Latent Class Solutions

Class	N	Exp (*sd*)	*t* (*df*)	TAP (*sd*)	*t* (*df*)	Promo (*sd*)	*t* (*df*)	IS (*sd*)	*t* (*df*)	Change (*sd*)	*t* (*df*)
Peer Raters											
1	4,180	8.70 (9.98)	-	2.17 (4.98)	-	.17 (.38)	-	4.19 (.19)	-	-.01 (.01)	-
2	248	4.80 (8.60)	-	1.94 (5.28)	-	.10 (.31)	-	3.41 (.19)	-	.14 (.03)	-
Independent t-test			-6.78** (287.56)	-	-.71, *ns* (4426)	-	-3.12** (292.54)	-	-63.35** (4426)	-	73.16** (253.91)
Supervisor Raters											
1	4,147	8.68 (9.98)	-	2.18 (4.99)	-	.17 (.39)	-	4.07 (.23)	-	-.04 (.07)	-
2	232	5.69 (8.85)	-	1.31 (4.31)	-	.07 (.25)	-	3.28 (.18)	-	.09 (.08)	-
Independent t-test			-4.98** (265)	-	-2.95** (266.88)	-	-5.55** (294.53)	-	62.07** (272.55)	-	24.03** (248.28)

Note: * denotes p<.05; **denotes p<.01; Exp denotes managerial experience; TAP denotes tuition assistance program; Promo denotes promotion rate; IS denotes initial status; *t* denotes *t*-value; *sd* denotes standard deviation; *df* denotes degrees of freedom; Change denotes performance change; Exp scaled in years of managerial performance at the beginning of the study (2005); TAP scaled in 1000ths of dollars paid for tuition assistance through the end of the study; Promo scaled in number of upward job moves made during the study's duration.

Table 12
MSF Performance Rating Effect Sizes

	M	SD	*d (sd)*	*F-test*	***Significant Post Hoc Comparison***
Self Raters					
1. 2005	4.01	.50	-		
1 vs 2	-	-	.04 (.42)	-	**
2. 2006	4.03	.49	-		
2 vs 3	-	-	-.09 (.38)	-	**
3. 2008	4.00	.49	-		
3 vs 1	-	-	-.04 (.42)	-	*
1 vs. 2 vs. 3				12.62** (1.96,10040.06)	
Peer Raters					
4. 2005	4.14	.42	-		
4 vs 5	-	-	.16 (.24)	-	**
5. 2006	4.17	.41	-		
5 vs 6	-	-	-.04 (.22)	-	-
6. 2008	4.17	.40	-		
6 vs 4	-	-	.12 (.24)	-	**
4 vs 5 vs 6				20.04** (1.96, 10081.80)	
Supervisor Raters					
7. 2005	4.02	.42	-		
7 vs 8	-	-	-.03 (.43)	-	-
8. 2006	4.00	.54	-		
8 vs 9	-	-	-.13 (.46)	-	**
9. 2008	3.95	.52	-		
9 vs 7	-	-	-.21 (.35)	-	**
7 vs 8 vs 9				49.69** (1.91, 9796.31)	

*Note: * denotes p<.05; ** denotes p<.01.*