THE ROLE OF PERFORMANCE DIRECTIVES IN PREDICTING SUPERVISORY EFFECTIVENESS

by

MICHELLE L. REYNARD MINNICH

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment

of the requirements for the degree of Doctor of Philosophy,

The City University of New York

2007

UMI Number: 3283139

Copyright 2007 by
Minnich, Michelle L. Reynard

All rights reserved.

UMI®

UMI Microform 3283139
Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2007
MICHELLE L. REYNARD MINNICH
All Rights Reserved

This manuscript has been read and accepted for the
Graduate Faculty in Psychology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Judith L. Komaki, Ph.D.
Chair of Examining Committee

Joseph Glick, Ph.D.
Executive Officer

Joel Lefkowitz, Ph.D.

Walter Reichman, Ph.D.

James Smither, Ph.D.

Robert E. Long, Ph.D.
Supervision Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE ROLE OF PERFORMANCE DIRECTIVES IN PREDICTING

SUPERVISORY EFFECTIVENESS

by

Michelle L. Reynard Minnich

Adviser: Professor Judith L. Komaki

In making decisions about whom to select or promote as a supervisor, practical predictors are needed that tap interpersonal behaviors required for motivating employees to do their jobs well. An in-basket exercise based on the Operant Model of Effective Supervision (Komaki, 1998) was given to 35 investment bankers and scored for the behaviors of monitoring, providing consequences, and providing antecedents. Ratings of motivational effectiveness and technical expertise were also collected from participants' supervisors, subordinates, and peers. As predicted, monitoring, one type in particular – via work sampling, and providing warranted consequences were related to motivational effectiveness. An unexpected finding was that providing tacked-on traditional antecedents was significantly negatively related to technical expertise. With more research, the in-basket exercise could have practical applications for selecting people for managerial positions.

TABLE OF CONTENTS

Chapter 1: INTRODUCTION ... 1

Chapter 2: NONCOGNITIVE, VALIDATED PREDICTORS OF LEADERSHIP FOR SELECTING MANAGERS ... 5

- Cognitive Predictors Predominate ... 5
- Supervisory Noncognitive Predictors are Neglected ... 6
- Validated Noncognitive Supervisory Predictors are Needed ... 14

Chapter 3: NON-COGNITIVE LEADERSHIP BEHAVIORS AND EVIDENCE OF EFFECTIVENESS ... 17

- Ohio State Studies ... 17
- Path-Goal Theory ... 20
- Multiple-Linkage Model ... 23

Chapter 4: THE OPERANT MODEL OF EFFECTIVE SUPERVISION: MONITORING AND PROVIDING CONSEQUENCES AS PREDICTORS ... 26

- Two Behaviors Grounded in Theory Are Hypothesized to be Related to Supervisory Effectiveness ... 26
- Research Results Generally Support the Operant Model ... 27
- Most Research on the Model Utilized Measures that are Impractical for Applied Practice ... 38
- An In-basket Assessment Based on the Operant Model was Developed and Tested with One Sample ... 40
- The Present Study is a Replication of Previous Work, but Using an In-basket Measure of Leader Behaviors ... 43
- Additive and Multiplicative Combinations of Monitoring and Providing Consequences May be Useful for Predicting Supervisory Effectiveness ... 44

Chapter 5: DIRECTIVES AS A NONTRADITIONAL PREDICTOR ... 48

- Suppressor Variables, Though Nontraditional and Uncommon, Improve Prediction ... 48
- Another Behavior Identified in the Operant Model Fits the Definition of a Classic Suppressor Variable ... 50
- Antecedents Have a Secondary Role, and Could Be Used as a Nontraditional Predictor ... 52
- While Antecedents are Highlighted in Some Leadership Theories and Commonly Used in Organizations, Some Question Their Efficacy ... 54
- Research Evidence Supports the Secondary Role of Providing Antecedents ... 57
- Providing Antecedents Will be Utilized as a Predictor in the Present Study ... 63

Chapter 6: PROBLEM AND PROPOSED HYPOTHESES ... 65

Chapter 7: METHOD ... 67

- Subjects and Setting ... 67
- Measures ... 69
- Procedure ... 107

Chapter 8: RESULTS ... 111

- Psychometric Properties of the Predictor ... 111
- Psychometric Properties of the Criterion and Technical Skill Measures ... 135
- Hypothesis Testing ... 143

Chapter 9: DISCUSSION ... 161

- Results and their Implications ... 161
- Limitations ... 170

Future Research 172
APPENDIX A 176
APPENDIX B 204
APPENDIX C 207
APPENDIX D 209
APPENDIX E 211
BIBLIOGRAPHY 213

LIST OF TABLES

Number *Page*

Table 1. Summary of Selected Research Involving the Operant Model of Effective Supervision.... 31
Table 2. Multi-Trait Multi-Method Matrix for Monitors and Consequences Measured via Direct Observation, an In-Basket Exercise, and Scoring of E-mail Messages............................. 42
Table 3. Correlations among Selected Behavior Categories and Supervisory Effectiveness 50
Table 4. Multiple Regression Analysis for Behavior Categories and Criterion of Supervisory Effectiveness 51
Table 5. Verbatim Transcripts from Marginally Effective and Effective Leaders Aboard Sailboats 60
Table 6. Frequency of Job Title by Region for Study Participants.. 68
Table 7. Summary of Demographic Information 69
Table 8. Definitions and Examples of Major Categories and Subcategories of Behavior Measured in the Revised OSIBA.. 73
Table 9. Items in the Revised OSIBA and Points possible for Categories and Subcategories of Behavior 91
Table 10. Point Values for Various Qualities of Monitoring and Providing Positive Consequences.. 93
Table 11. Point Values for Various Qualities of Providing Negative Consequences 95
Table 12. Point Values for Various Qualities of Providing Antecedents.. 97
Table 13. Scoring Example for Monitoring and Monitoring via Work Sampling for a Single Respondent.. 99
Table 14. Scoring Calculations for Major Categories.. 100
Table 15. An OSIBA Item and Examples of Responses... 102
Table 16. Average Interrater Agreement Scores for Categories and Subcategories 115
Table 17. Intraclass Correlations for In-basket Category and Subcategory Scores 118
Table 18. Summary of In-basket Scores 122
Table 19. Intercorrelations Between Major Categories and Subcategories of Behavior in the OSIBA 125
Table 20. Summary of Reliability Analyses for Categories or Subcategories Measured in Different In-basket Items 130
Table 21. Results from ANOVAs for Job Title, Location, and Gender with Several Dependent Measures of In-basket Scores 132
Table 22. Correlations Between Selected Demographic Variables and In-basket Scores 135
Table 23. Correlations Among Motivational Effectiveness and Technical Skill Scores Provided by Various Sources.. 138
Table 24. Results from ANOVAs for Observational Context Groups with Dependent Measures of Motivational Effectiveness and Technical Skill... 140
Table 25. Results from ANOVAs for Job Title, Location, and Gender with Dependent Measures of Motivational Effectiveness and Technical Skill 141

Table 26. Correlations Between Selected Demographic Variables and Ratings of Motivational Effectiveness and Technical Skill 143

Table 27. Correlations Between In-basket Category and Subcategory Scores and Ratings of Motivational Effectiveness and Technical Skill 144

Table 28. Monitoring in Responses to an OSIBA Item from Top and Bottom Rated Motivators .. 146

Table 29. Monitoring via Work Sampling in Responses to an OSIBA Item from Top and Bottom Rated Motivators........ 146

Table 30. Providing Positive Consequences by Acknowledging the Bearer of Bad News in Responses to an OSIBA Item from Top and Bottom Rated Motivators 148

Table 31.Multiple Regression and Correlation Results Using Categories and Subcategories as Predictors and Ratings of Motivational Effectiveness and Technical Skill as Criteria ... 150

Table 32. Tacked on Traditional Antecedents in Responses to an OSIBA Item from Top and Bottom Rated Technical Experts........ 154

Table 33. Hierarchical Multiple Regression Results Using Monitoring, Consequences, and an Interactive Term as Predictors 156

Table 34. Combinations of Participant Monitoring and Consequence Scores, and Their Relationships with Effectiveness........ 159

Table 35. Hierarchical Multiple Regression Results Using Monitoring, Consequences, and Antecedents as Predictors 160

Chapter 1

INTRODUCTION

Managerial jobs are complex, requiring a variety of behaviors, skills, and abilities. In his classic study describing managerial work, Mintzberg (1975) identified ten roles or sets of behaviors that were common. These roles ranged from the interpersonal, such as figurehead, leader, and liaison, to informational -- monitor, disseminator, and spokesman, and to decisional -- entrepreneur, disturbance handler, resource allocator, and negotiator. Other researchers have developed quite lengthy, comprehensive taxonomies of managerial performance dimensions or competencies (e.g., Borman & Brush, 1993; Hemphill, 1959; Tett, Guterman, Bleier, & Murphy, 2000; Tornow & Pinto, 1976). Borman and Brush (1993), for example, identified 18 "mega-dimensions" grouped into four general areas of "interpersonal dealings and communication, leadership and supervision, technical activities and 'mechanics of management,' and useful personal behavior and skills" (p. 10). Fifty-three managerial competencies were offered by Tett et al. (2000). These included competencies associated with traditional functions (e.g., problem awareness, decision making, planning, coordinating, motivating, team building, productivity), task orientation (e.g., initiative, task focus, urgency, decisiveness), person orientation (e.g., compassion, assertiveness, customer focus), dependability (e.g., orderliness, trustworthiness, timeliness, professionalism), open mindedness (e.g., tolerance, adaptability, creative thinking, cultural appreciation), emotional control (e.g., resilience, stress management), communication (e.g., listening skills, oral communication, written communication), developing self and others (developmental goal setting, performance assessment, developmental feedback, self-development), and occupational acumen and concerns (e.g., technical proficiency, quantity concern, quality concern, financial concern, safety concern).

In examining this wide-ranging assortment of abilities and skills necessary for managers, one can observe that there are some that are cognitive in nature, or involve a good amount of

thinking, cognitive ability, or general mental ability. Many of the behaviors comprising Mintzberg's (1975) informational and decisional roles could be characterized as cognitive. For example, Mintzberg (1975) described managers as people who process information, constantly scanning the environment and disseminating certain information both within and outside their organizational units, and who make complex choices regarding how resources will be allocated and the unit structured. Competencies such as planning and organizing, decision making, and problem analysis are also examples. The following are definitions of these competencies from the published literature on assessment centers, a tool sometimes used for managerial selection:

1. Planning and organizing: establishing and implementing a course of action to achieve a business objective; ability to set goals and priorities, identify and initiate goal-relevant actions, allocate and coordinate resources, and monitor progress (Goffin, Rothstein, & Johnston, 1996)
2. Decision making: readiness to make decisions and the quality of the decision made (Bray, Campbell, & Grant, 1974)
3. Problem analysis: identifying problems, gathering information, and determining possible causes of problems (Phillips, 1985).

(Assessment centers [ACs] are well-researched programs that actually consist of multiple methods of assessment used to make multiple assessments of a wide range of dimensions of behavior relevant to the position to be filled. Any one dimension is measured through more than one method of measuring it, such as objective tests, projective tests, group and individual interviews, and individual- and group-based situational exercises, such as in-baskets and leaderless group discussions. Trained assessors observe performance during exercises, interview candidates, and/or score tests, evaluating each candidate on the identified dimensions [Guion, 1998].)

Other managerial capabilities are less cognitive in nature. These "noncognitive" skills and abilities often involve interpersonal skills, human relations, or effectively interacting with other people. The interpersonal roles of figurehead, liaison, and leader identified by Mintzberg (1975) are

noncognitive. Managers perform ceremonial duties (with little "serious communication" and no important decision making) as figureheads and build networks and contacts outside of their units as liaisons. In the leader role, the focus is more on supervising employees -- managers "motivate and encourage … employees, somehow reconciling their individual needs with the goals of the organization" (p. 10). Specific dimensions that are also examples of noncognitive capabilities include maintaining good working relationships, team building/orientation, coaching or developing others, supervising/motivating, and leadership. Borman and Brush (1993) defined maintaining good working relationships as "developing and maintaining smooth and effective working relationships with superiors, peers, and subordinates; displaying personal concern for subordinates; backing up and supporting subordinates as appropriate; and encouraging and fostering cooperation between subordinates" (p. 12). Examples of dimensions that may be measured in assessment centers include:

1. Team orientation: Being willing and able to work effectively in a work group and exhibiting interpersonal awareness, participatory decision making, a desire to help, conflict management skills, and team-building skills (Goffin, et al., 1996)
2. Leadership: ability to recognize when a group requires direction, to get others involved in solving problems, to effectively interact with a group, to guide them to the accomplishment of the task (Schmitt, Noe, Meritt, & Fitzgerald, 1984).

Both cognitive and noncognitive skills are important in managerial positions, but the interpersonal skills it takes to work with employees and the ability to motivate them to do their jobs well are critical. By virtue of the title of manager, a person is responsible for the activities and work of the people who report to him or her. Thus, skill at working and interacting with people in order to get work accomplished is fundamental. Mintzberg (1975) recognized the importance of this aspect: "The influence of the manager is most clearly seen in the leader role. Formal authority vests him with great potential power; leadership determines in large part how much of it he will realize" (p. 10). In other words, the power from the title of leader, manager, or supervisor alone will not make

work happen or determine how much will get done; what is key is how well the person interacts with employees, motivating and encouraging performance.

When it comes to selecting people for managerial positions, however, the predictors (or selection measures/instruments) used can be narrowly focused or heavily weighted on cognitive aspects (Goldstein, Zedeck, & Goldstein, 2002; Pulakos & Schmitt, 1996). For example, interviews can focus on assessing how candidates organize, plan, and prioritize, how they adapt to change, their level of knowledge regarding the technology involved in the job, and their oral communication skills and spend little time on how candidates interact with the people they supervise. A specific example is the test used for promotion of firefighters to lieutenant (first-level supervisor) in the New York City Fire Department. The test is comprised of multiple choice items focused mostly on technical aspects and procedures -- fire/emergency evaluation, management, and follow up; fire prevention laws, evaluations, inspections, and violation recognition; reports, record keeping, and general office activities; and unit inspections. Potential test takers are instructed that questions may require use of abilities such as planning and organizing, delegation, sensitivity, analysis, judgment, decisiveness, setting work standards, and technical translation (The City of New York Department of Citywide Administrative Services, 2002). Unfortunately, these selection methods tell only half the story. Persons could be selected who are intelligent, have a high level of technical knowledge (or know the policies and procedures), are good at making decisions or analyzing and solving problems, but could not supervise employees well. The “rest of the story” – the interpersonal, supervisory aspect, remains untold.

Chapter 2

NONCOGNITIVE, VALIDATED PREDICTORS OF LEADERSHIP FOR SELECTING MANAGERS

The noncognitive domain includes such variables as personality factors (e.g., conscientiousness, extraversion, agreeableness, neuroticism, openness to experience, or integrity) and other skills or abilities like tolerance for stress and flexibility, in addition to the more interpersonal skills described in the previous chapter (e.g, human relations, team building, coaching and motivating, or leadership). Several noncognitive predictors are common in the selection literature, including personality tests or inventories, integrity tests, violence scales, stress tolerance scales, and customer service scales. It is the noncognitive *leadership or supervisory* predictors that might be used to select people for managerial positions that are of interest in this research. These predictors are viewed as especially important because of the significance of interpersonal skills in managing others and the high proportion of cognitive aspects in predictors that are used. Having noncognitive predictors in a battery of selection measures in addition to cognitive ones provides a more rounded picture of applicants. It also better represents the performance domain, accounting for more variability in the criterion of managerial effectiveness. However, despite the importance of noncognitive supervisory aspects, relatively few validated predictors concerned with behavior exist in the selection literature.

Cognitive Predictors Predominate

Cognitive predictors are well-established, widely used in selection, and validation evidence in general is strong. Although the definition, structure, and conceptualization of intelligence continues to be investigated and debated (Guion, 1998), the construct of *g*, or generalized intelligence, general mental ability, or cognitive ability permeates the field of personnel selection. Standardized paper-and-pencil tests of cognitive ability are often used as predictors in selection

(Goldstein et al., 2002). Also, specific dimensions that require thinking, such as decision making and problem analysis, common in managerial selection, might be defined as cognitive abilities (or be highly related to *g*). These are often assessed in other ways than through standardized paper-and-pencil tests, such as through an interview or exercises in assessment centers.

Measures of cognitive ability have consistently been found to be the best predictor of performance across jobs in the United States (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). In a relatively recent review, Schmidt and Hunter (1998) noted that the average estimated validity (corrected for artifacts) across a wide variety of jobs and settings was .51. Partly because of this high predictive validity, they argued that general mental ability should be considered the "primary personnel measure for hiring decisions" (p. 266) and others (such as work sample tests, integrity tests, interviews, and assessment centers) supplements that might add some incremental validity if not highly correlated with it.

Supervisory Noncognitive Predictors are Neglected

While cognitive dimensions are rich and involve several facets such as reasoning skills, problem solving, judgment, planning, and thinking, it is not the same for noncognitive dimensions in general. Noncognitive variables seem like a haphazard group of very different aspects, as discussed earlier. Many are often lumped together and labeled "interpersonal skills," and hence, are vaguely defined. Goldstein et al. (2002) lamented the lack of adequate language to discuss these variables, noting that "[s]trong theoretical definitions do not exist for numerous noncognitive competencies. … [T]here is cognition, and then everything else is merely referred to as noncognitive" (p. 132).

Some specific noncognitive predictors have been well-researched and evidence of criterion-related validity has been found. Over the last 25 years, interest in personality traits and their relationship to job performance has increased exponentially, resulting in "an explosion" of studies of the Big Five personality dimensions (openness to experience, conscientiousness, extraversion,

agreeableness, emotional stability) in particular. Several meta-analyses have been conducted on this body of work (e.g., Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Mount, Barrick, & Stewart, 1998; Salgado, 2003; Tett, Jackson, & Rothstein, 1991). Average criterion-related validities of .27 for measures of conscientiousness and .13 for emotional stability were found across jobs and criterion types in a recent secondary meta-analysis (Barrick, Mount, & Judge, 2001). Other noncognitive measures have been created to predict certain criteria, such as integrity tests (designed to predict dishonest behaviors at work), violence scales (physically or psychologically violent/hostile behaviors at work), stress tolerance scales (handling work pressures well), and customer service scales (handling customers well) (Ones & Viswesvaran, 2001). These measures are similar to personality measures in that they tap applicants' predispositions to behave in particular ways (they are sometimes included in personality tests, as well). Results of meta-analyses indicated that all measures except for stress tolerance scales were related to the criteria they were designed to predict, broader measures of counterproductive criteria, and supervisory ratings of overall performance (no research could be found relating stress tolerance scales to stress reactions at work) (Ones & Viswesvaran, 2001). In the case of integrity tests, the most researched and utilized of the group of measures, mean validities of .13 for detected theft. .42 for admitted theft, .32 for counterproductive behavior, and .41 for overall performance were estimated (Ones, Viswesvaran, & Schmidt, 1993).

Noncognitive predictors that are supervisory and behavioral (rather than dispositional or trait-like), however, seem to be neglected. They are sometimes measured inadequately and validation evidence is difficult to discern in the research literature. An ideal noncognitive measure for selecting managers would tap supervisory behaviors in which supervisors interact with employees about their performance. For example, candidates could be presented with a report of a complaint from a customer that staff in a particular department did not answer the phone promptly and when calls were finally answered, the personnel were rude. One would want to see how

candidates deal with the problem and attempt to motivate the advertising staff to answer the phone and handle customer calls properly. This would be a measure of actual performance, as opposed to a self-report of performance, where candidates' say how they would respond or have responded to a similar situation in the past. It would also be important to assess how candidates would perform in a variety of contexts or situations, allowing ample opportunity to demonstrate behaviors and establish a pattern. In addition, in an ideal world, the supervisory behaviors measured would be shown to result in effectiveness, not overall, but in terms of motivating others. (Overall effectiveness or job performance includes many aspects besides supervising or motivating others.)

Unfortunately, when the literature on selection is examined, the situation appears less than ideal. Noncognitive aspects tend to make up a small proportion of the predictors used in the selection of managers. Assessment centers, for example, typically cover a wide range of skills and abilities, such as planning and organizing, decision making, problem analysis, innovation, initiative, judgment, tolerance for stress, leadership, and written and oral communication. Although these dimensions are important facets of a supervisory job, many are cognitive in nature and do not involve the interpersonal aspects of dealing with people or a supervisory component. Proficiency at analyzing problems and making decisions, for example, is necessary for success in most managerial jobs – managers are often expected to quickly identify and solve problems and make decisions. These dimensions, however, are cognitive and not interpersonal, meaning they do not capture how a person interacts with employees. Written and oral communication involve interaction with others. However, these dimensions do not often contain a supervisory component when measured. Assessments of written communication skill frequently concern technical aspects of the job and how clearly they are portrayed in writing. Oral communication is usually assessed by the candidate delivering a presentation to peers (who are also being assessed). Neither assessment focuses on communication with employees nor on how the content of the communications motivates employees. Of the few dimensions that are more clearly noncognitive (tolerance for stress and

leadership), even fewer contain interaction with others or a supervisory aspect. Tolerance for stress is a personal characteristic that does not encompass interactions with employees. Only one dimension, "leadership," is noncognitive, encompasses interactions with others, and is supervisory (depending upon the definition of the dimension and how it is measured). Thus, the noncognitive constructs involving interaction with subordinates and supervision often make up only a small aspect of the selection process.

In addition to supervisory skills playing such a small part in selection, a problem with measuring these skills is that the predictors used may not capture the "true" aspects of supervision that would be seen in behaviors during interactions with subordinates. In structured, behavioral interviews, for example, candidates are asked to describe details of particular situations and what they did in those situations. (Questions are typically a series of prompts, such as "Describe a time when someone's performance was falling short of what was expected. What were the details of the situation? What actions did you take? What was the result?") Interviewers make assessments of skills based on the "stories" the candidates tell in response to the questions. Thus, judgments of supervisory skill are made on self-reports of behaviors (which may or may not be true).

Even in assessment centers, in which observations of actual behaviors are carried out during exercises and work simulations, the measurement of noncognitive managerial skills may not be "ideal." Exercises do not always involve "subordinates," and candidates often interact with other candidates or peers. A commonly used exercise, for instance, is a leaderless group discussion. During this exercise, AC participants are given a problem to solve as a group and a time limit in which to complete the exercise. No one is assigned the role of chair or leader (Guion, 1998). In this case, the assessments of leadership or supervisory skills are made on candidates' performance interacting with peers.

Another problem with noncognitive predictors in the selection literature is that it is difficult to discern which behaviors are related to effectiveness in motivating others. In the published

research on assessment centers, evidence for the relationship of specific dimensions within them to criteria of supervisory effectiveness is somewhat limited (though a recent meta-analysis has revealed some encouraging results). Of a sampling of research studies investigating the criterion-related validity of ACs (i.e., Anderson & Thacker, 1985; Avolio, O'Connell, Martiz, & Kennedy, 1999; Binning, Adorno, & LeBreton, 1999; Bobrow & Leonards, 1997; Borman, 1982; Bray & Campbell, 1968; Bray et al., 1974; Chan, 1996; Dayan, Kasten, & Fox, 2002; Goffin et al., 1996; Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001; Hinrichs, 1969, 1978; Huck & Bray, 1976; Jones, Herriot, Long, & Drakely, 1991; Jones & Whitmore, 1995; Kraut & Scott, 1972; McEvoy & Beatty, 1982; Mitchel, 1975; Moses & Boehm, 1975; Ritchie, 1994; Ritchie & Moses, 1983; Russell & Domm, 1995; Schleicher, Day, Mayes, & Riggio, 2002; Schmitt, Noe, et al., 1984; Schmitt, Schneider, & Cohen, 1990; Shore, Shore, & Thornton, 1992; Turnage & Muchinsky, 1984; Tziner & Dolan, 1982; Tziner, Ronen, & Hacohen, 1993; Wollowick & McNamara, 1969; Worbois, 1975), only a fraction report the results and discuss the relationships between the individual dimensions and criteria (Anderson & Thacker, 1985; Avolio, et al., 1999; Bray et al., 1974; Goffin et al., 1996; Goldstein, et al., 2001; Hinrichs, 1978; Jones, et al., 1991; Jones & Whitmore, 1995; Moses & Boehm, 1975; Ritchie, 1994; Ritchie & Moses, 1983; Russell & Domm, 1995; Schmitt, Noe, et al., 1984; Shore et al., 1992; Wollowick & McNamara, 1969; Worbois, 1975).

In this group of studies, the patterns of relationships between the leadership or interpersonal dimensions and different criteria seemed to vary. (The criteria used were often overall performance measures, and not specific measures of effectiveness at motivating others.) Management levels attained by men and women were correlated with leadership dimensions (undefined) in two studies (Moses & Boehm, 1975; Ritchie & Moses, 1983). Correlations for men and women, respectively, were .38 and .26 in Moses and Boehm's (1975) study and .32 and .36 in Ritchie and Moses' (1983). Retail store managers' scores on Russell and Domm's (1995) leadership dimension (defined earlier)

were significantly related to some of the criteria measured in the study – customer relations responsibilities (r = .23), the sum of district managers' performance ratings on four dimensions (r = .20), and store profit (r = .34). Leadership surprisingly was *not* significantly related to personnel responsibility ratings, which were ratings of how well the manager selected, trained, set objectives for and communicated responsibilities to, appraised performance of, communicated with, motivated, and provided career guidance for store employees. The only interpersonal dimension in another AC, "interpersonal contact," (which was not defined), was *not* related to the criterion of change in position level after three years, r = .00 (Wollowick & McNamara, 1969).

A relatively recent meta-analysis of the criterion-related validities of AC dimensions was the first and only study to make sense of the variety in findings (Arthur, Day, McNelly, & Edens, 2003). The authors included validity coefficients in the dataset only if supervisor ratings of performance or promotion were used as criteria, which yielded 258 coefficients from only 34 studies. A total of 168 dimensions were collapsed down to six: (1) consideration/awareness of others, "the extent to which an individual's actions reflect a consideration for the feelings and needs of others as well as an awareness of the impact and implications of decisions relevant to other components both inside and outside the organization" (p.133); (2) communication, "the extent to which an individual conveys oral and written information and responds to questions and challenges" (p. 133); (3) drive, "the extent to which an individual originates and maintains a high activity level, sets high performance standards, and persists in their [sic] achievement, and expresses the desire to advance to higher job levels" (p. 134); (4) influencing others, "the extent to which an individual persuades others to do something or adopt a point of view in order to produce desired results and takes action in which the dominant influence is one's own convictions, rather than the influence of others' opinions" (p. 134); (5) organizing and planning, "the extent to which an individual systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which an individual anticipates and prepares for the

future (p. 135); and (6) problem solving, "the extent to which an individual gathers information; understands relevant technical and professional information; effectively analyzes data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; uses available resources in new ways; and generates and recognizes imaginative solutions" (p. 135).

Aspects of leadership were captured in the influencing others dimension, as evidenced from the lists of dimension labels that were included in each aggregated dimension (original dimension definitions were not provided). Interpersonal or "interactional" components were included in the consideration/awareness of others and influencing others dimensions. For example, influencing others included interpersonal/leadership dimensions such as gaining team commitment, group leadership, influencing others, negotiation, persuasion and negotiation skills, transactional leadership, and transformational leadership, as well as non-interpersonal ones such as autonomy, character/personality, inspiring trust, integrity, and self-confidence.

The meta-analysis yielded estimated true validities of .39 for the problem solving dimension, .38 for influencing others, .37 for organizing and planning, .33 for communication, .31 for drive, and .25 for consideration/awareness of others. Arthur et al. (2003) also performed a hierarchical regression analysis and found that four dimensions, namely problem solving, influencing others, organizing and planning, and communication, accounted for the criterion-related validity of AC ratings ($R = .45$, $R^2 = .20$, $p < .01$). Drive and consideration/awareness of others did not make significant contributions to the multiple correlation when the other four dimensions were included in the prediction model. Not surprisingly, those dimensions that were most closely related to cognitive ability (or had the highest cognitive load) had the highest validities. These results provide the most definitive and broad view of the relationships between common AC dimensions and managerial performance criteria.

It is also difficult to find information about the validity of specific dimensions because it is common for practitioners (and researchers) to consider the results for selection methods like interviews and assessment centers as a whole, rather than for individual dimensions or constructs. Though these methods measure a variety of dimensions, collapsing or aggregating the ratings from different dimensions for each candidate into an overall AC score or interview rating is quite common. Overall AC scores themselves, in fact, appear to predict a variety of criteria reasonably well. Several meta-analyses have examined assessment centers' average validities for different criteria. One showed the average validity coefficient for criteria of performance ratings was .43, for promotion was .41, and for scholastic achievement was .31 (Schmitt, Gooding, Noe, & Kirsch, 1984). Another indicated a mean validity of .43 for predicting promotion (Hunter & Hunter, 1984). A meta-analysis of only AC studies found mean coefficients of .33 for criteria of performance ratings on the dimensions in the AC, .35 for training performance, .36 for both performance ratings (e.g., overall or on other aspects besides the assessment center dimensions) and career advancement indices, and .53 for ratings of management potential (Gaugler, Rosenthal, Thornton, & Bentson, 1987).

Lastly, validation evidence for noncognitive predictors is not always available in the published literature. Often, selection tools and processes are proprietary, and validation research is not published (or data are not collected in such a way that leads to being published). In the author's former organization, for instance, notes from structured interviews and candidate ratings for manufacturing supervisor positions were typically maintained by hiring managers. These data were never gathered together and published, thus contributing to the scarcity of validation results for specific noncognitive predictors in the literature.

In sum, noncognitive predictors useful in selecting managers are a motley group, often playing only a small part in selection and not always being measures of actual skills in interacting with subordinates. Evidence for their validity (individually) is also difficult to find in the published

selection literature due to the proprietary nature of selection tools and the common practice of collapsing individual dimension results into an overall score. Published results have not revealed a consistent pattern, though a meta-analysis of the criterion-related validities of AC dimensions has shown promising results.

Validated Noncognitive Supervisory Predictors are Needed

Recently, a few researchers have begun to emphasize the importance of interpersonal or noncognitive dimensions in an effort to expand the "predictor domain" beyond general mental ability, increase predictive validity of selection processes, and reduce a particular problem known to plague cognitive predictors – differences in scores by members of racial subgroups (e.g., Goldstein et al., 2001; Goldstein et al., 2002; Pulakos & Schmitt, 1996; Outtz, 2002). Research has shown that Whites typically outscore Blacks by about one standard deviation on general mental ability tests (Hunter & Hunter, 1984; Schmidt, 1988). These racial differences have been estimated to be three to five times greater than for other predictors with comparable levels of validity (Outtz, 2002). If cognitive ability test results were used to make a selection decision, it could lead to a situation in which the proportion of minority applicants hired is less than 80 percent of the proportion of majority applicants hired – the legal definition of adverse impact.

A good deal of research has been conducted on the increments in validity and decrements in minority group differences in scores from utilizing other predictors in addition to general mental ability (e.g., Bobko, Roth, & Potosky, 1999; Cortina, Goldstein, Payne, Davison, & Gilliland, 2000; Goldstein et al., 1998; Schmidt & Hunter, 1998; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997) and on comparisons of alternative methods of the same test that place different levels of cognitive demand on the test taker (e.g., Chan & Schmitt, 1997; Pulakos & Schmitt, 1996). This research, however, focused mainly on predictors in terms of their methods and not content, meaning that the emphasis was not on the constructs or dimensions measured (or their level of cognitive content), but on how constructs were measured (e.g., work sample tests, job knowledge tests,

biodata measures, assessment centers). For instance, Bobko et al. (1999) presented subgroup differences for cognitive ability and “alternative predictors” of structured interviews, conscientiousness, and biodata, acknowledging that they were confounding “methods with the underlying, latent constructs” (p. 563).

Recently, however, researchers have considered the effects of utilizing noncognitive content. Using a “broad-based” assessment approach including several cognitive and noncognitive abilities measured through a biodata measure, situational judgment test, structured interview, and paper and pencil verbal ability test, Pulakos and Schmitt (1996) found large increases in criterion-related validity and decreases in racial subgroup differences, as compared to the verbal ability test alone. (Oral communication, relating to others, motivation and adapting to change were among the noncognitive abilities measured, but scores were only presented at the test, or method, level.) Goldstein et al. (2001) looked at whether ratings of managerial competencies with differing levels of cognitive load varied in terms of subgroup differences in scores. Ratings were made based on participants’ performance in seven AC activities including an in-basket, interactive role plays, and team exercises. Cognitive load was determined by the magnitudes of correlations between dimension scores and scores on a standardized, pencil-and-paper cognitive ability test. The authors found, generally, that competencies with greater cognitive load (e.g., judgment, organizing, planning) tended to have larger Black-White differences, while competencies with lesser cognitive load (e.g., human relations, effectiveness under stress) had smaller Black-White differences. In addition, all competencies (except one, which was not significantly correlated with job performance) contributed incremental validity to predicting job performance above and beyond cognitive ability. The results of these studies are promising, suggesting that noncognitive predictors are worthy of attention in prediction.

In sum, validated instruments measuring noncognitive, supervisory skills are needed to improve the selection of managers. This often neglected aspect is an important part of managerial

jobs and including predictors to measure it appears to increase selection battery validity and mitigate racial subgroup differences in scores associated with cognitive predictors.

Chapter 3

NON-COGNITIVE LEADERSHIP BEHAVIORS AND EVIDENCE OF EFFECTIVENESS

If one is interested in using noncognitive predictors to select managers, one could look to the leadership literature for guidance, since it is replete with theories that specify what constitutes effective supervisory leadership. Those models that are based in theory and for which there is some evidence of a relationship between the concepts and effectiveness offer the best suggestions for potential predictors.

In Sayles' (1979) discussion of what effective managers do, he observed that the real world of managers is hectic, time-pressured, and fragmented, and often people find it difficult to successfully get others to execute. He noted that those who consider leadership theories for guidance find that the theories appear useful and relevant, but actually focus on what managers should think or on a desired outcome, offering no real clear recommendations for what leaders should *do* to be effective. He described that the middle – the process, the how – is typically neglected. Several leadership theories that focus more on the "middle ground," in the sense that they specify the behaviors that leaders should exhibit will be examined here. The theories will be compared in terms of their origins, the suggestions for what leaders should do to motivate their subordinates and whether research has shown a link between the theories' components and supervisory effectiveness.

Ohio State Studies

One of the early investigations in what is known as the behavioral approach to leadership, which emphasizes what leaders and managers actually do on the job and the relationship of this to leader effectiveness, was the group of leadership studies conducted at Ohio State University in the 1950s. These studies sought to identify relevant aspects of leadership behavior, rather than traits of leaders.

Two Behaviors that Arose from Empirical Research are Thought to be Effective

The behaviors identified were derived from empirical research involving a questionnaire, called the Leader Behavior Description Questionnaire (LBDQ), in which subordinates of leaders were asked to indicate how frequently their leaders displayed a lengthy list of behaviors. Responses were factor analyzed and results revealed that subordinates perceived their leaders' behavior primarily in terms of two independent categories, one having to do with task-oriented behaviors, the other with people-oriented behaviors (Fleishman, 1953, 1973).

The two empirically derived factors gave rise to a taxonomy of behavior consisting of consideration and initiating structure. Consideration described the extent to which a leader displays concern for the welfare of the other members of the group. In this category, the leader is oriented toward relationships, friendship, mutual trust, and interpersonal warmth (Fleishman, 1973). The considerate leader, for example, maintains and strengthens the self-esteem of his/her subordinates by treating them as equals, makes special efforts to put subordinates at ease, and is easy to approach (Fleishman, 1973). Initiating structure described the extent to which a leader organizes activities and defines relationships in groups, establishes communication patterns and channels, and defines ways of getting work done. Leadership behaviors, such as insisting on maintaining standards and meeting deadlines and deciding in detail what will be done and how it should be done, are indications of initiation of structure (Fleishman, 1973). The underlying assumptions of the taxonomy are that all leader behaviors can be classified into the two separate and mutually exclusive categories and that extensive use of both sets of behaviors is essential for a leader to be effective (Yukl & VanFleet, 1990).

An offshoot of this work was the notion of the "high-high" leader – or that effective leaders use both person- and task-oriented behaviors extensively. For example, in the managerial grid theory, Blake and Mouton (1964) hypothesized that effective managers have a high concern for both people (consideration) and production (initiating structure). According to this view, the two

behaviors do not compensate for each other, but actually interact with one another and are mutually facilitative in their effects. In other words, one behavior is more beneficial when the other is high than when it is low.

Meta-analytic Evidence of Validity

The two-behavior taxonomy dominated the thinking of researchers for many years before being abandoned almost completely. Most of the studies from the 1950s to the 1970s relied on different versions of the LBDQ and several other questionnaires measuring consideration and initiating structure (Kerr & Schriesheim, 1974). The typical approach of hundreds of studies was to examine the correlations of leader initiating structure and consideration, with subordinate satisfaction and performance (also measured via questionnaire). Few studies were conducted with criteria that were behavioral and responsive to leaders' efforts. An example of this approach was found in a relatively recent study (Pool, 1997). The researcher administered Form XII of the LBDQ, and questionnaires measuring work motivation, job satisfaction, and substitutes of leadership (e.g., subordinates' characteristics, tasks, and organizational characteristics) to 125 respondents from several levels and settings. (The Pearson correlation between job satisfaction and the consideration leadership style was .50 [$p < .001$] and -.24 [$p < .01$] for the initiating structure style.)

Reviews of this research literature are mixed. Noting both methodological and conceptual limitations of this body of work, several reviewers concluded that results were, in general, weak and inconsistent (e.g., House & Aditya, 1997; Kerr & Schriesheim, 1974; Yukl & VanFleet, 1990, Yukl, 1998). With the exception of the finding that consideration was often positively correlated with subordinate satisfaction, reviewers observed no pattern of behavior consistently related to satisfaction or any criteria of supervisory or managerial effectiveness (House & Aditya, 1997; Yukl & VanFleet, 1990; Yukl, 1998). A recent meta-analytic review of the work found more optimistic results on the behaviors' validities than past qualitative reviews, however (Judge, Piccolo, & Ilies, 2004). Moderately strong relationships between the two behaviors and several leadership outcomes

were found. The average correlation between consideration and all criteria was .48 ($p < .01$) and initiating structure was .29 ($p < .01$). In general, consideration showed stronger correlations with various criteria than did initiating structure, and was more strongly related to follower satisfaction, motivation, and leader effectiveness. These results raise questions about the historically weak and inconsistent validity results, and even further, the abandonment of the behaviors in the research literature.

In addition, the high-high model of interaction between the behaviors was tested in only a small number of studies and in general, little evidence was found (Yukl, 1998). The results of one published study involving an instrument used by training participants to place themselves in the grid did not support the basic assumptions of the Blake and Mouton (1964) theory (Bernardin & Alvares, 1976). Also, two studies that assessed the high-high model through multiple regression found no significant gains in predicting subordinate satisfaction by adding an interactive term of consideration times initiating structure to models of consideration alone and consideration and initiating structure (Larson, Hunt, & Osborn, 1976; Schriesheim, 1982). Schreisheim (1982) concluded that “the superiority of the high-high leadership style is indeed a myth” (p. 221).

Path-Goal Theory

Path-Goal Theory (House, 1971; House & Mitchell, 1974) was considered to be a situational or contingency theory, specifying a number of situational moderators of relationships between the behaviors and their effects. In other words, a basic assumption was that different leader behavior patterns will be effective in different situations and the same behavior pattern will not be optimal in all situations.

With Both Theoretical and Empirical Origins, Four Behaviors are Hypothesized to be Effective Depending on Situational Factors

Path-Goal Theory was an extension of the results of the Ohio State studies (and was intended to address mixed results and conflicting findings regarding consideration and initiating

structure). A fairly complex model, four leader behaviors (initiating structure, consideration, participative, achievement-oriented); intervening variables (follower expectancies and valences); a number of situational or environmental moderators (task structure, task routinization, formal authority system, job autonomy, job scope) and follower trait moderators (dependence, authoritarianism, ability, locus of control); and two dependent variables (follower satisfaction and performance) were specified.

House and Mitchell (1974) added participative and achievement-oriented leader behaviors to House's (1971) model. Participative leadership was characterized by consulting with subordinates and taking into account their opinions and suggestions, while achievement-oriented leadership meant setting challenging goals, expecting excellence in performance, continually seeking performance improvements, and showing confidence that subordinates will attain high standards. The intervening variables, taken from the expectancy theory of motivation, explain how a leader's behavior affects the dependent variables. Follower expectancy (or path instrumentality) is the perceived probability of a behavior leading to an outcome, and valence is the desirability of the outcome. The force on an individual to engage in a specific behavior or motivation to work is a nonlinear, monotonically increasing product of expectancies and valences (House, 1971).

According to House (1971), the motivational functions of a leader include, "… increasing personal payoffs to subordinates for work-goal attainment and making the path to these payoffs easier to travel by clarifying it, reducing roadblocks and pitfalls, and increasing the opportunities for personal satisfaction en route" (p. 324). Leaders motivate higher performance and satisfaction in their subordinates by behaving in ways that influence workers to believe valued outcomes can be attained by making a serious effort. A successful leader not only shows his/her followers the rewards that are available, but also the paths (behaviors) through which the rewards may be obtained, if that is what is missing in the environment for the subordinate. The clarification enhances the "psychological state" of the subordinates and arouses them to boost their efforts to

perform well (Bass, 1990). Given clear tasks and roles, however, an effective leader will focus more on consideration behaviors, supporting subordinates and paying attention to their personal and interpersonal needs for satisfying relationships (Fiedler & House, 1988).

In 1996, House published a "reformulation" of the theory, which was even more complex than the theory developed in the 1970s. He broadened the scope of the model to address the effect of leaders on groups or work units, and offered 26 propositions involving additional classes of leader behavior (work facilitation, interaction facilitation, group oriented decision process, representation and networking, value based leader behavior, and shared leadership), individual differences of subordinates, and task moderators. The essence of the reformulated theory was characterized as the parsimonious meta-proposition that "… leaders, to be effective, engage in behaviors that complement subordinates' environments and abilities in a manner that compensates for deficiencies and is instrumental to subordinate satisfaction and individual and work unit performance" (p.323).

Mixed Research Results

Because of the complexity of the theory, it is difficult to empirically test the postulated relationships as a "whole." As a result, a great deal of research has been conducted on different "bits and pieces" or various propositions. These empirical tests have netted mixed results. Methodological problems, such as poor criterion measures, reliance on subordinate questionnaires for assessing leaders' behaviors and as criterion measures have made it difficult to interpret the results from much of the research (Bass, 1990, House, 1996; House & Aditya, 1997; Schriesheim & Neider, 1996; Yukl & VanFleet, 1990). Furthermore, most of the theory's propositions about situational moderators of directive/structuring behavior have not been supported (Podsakoff, MacKenzie, Ahearne, & Bommer, 1995; Wofford & Liska, 1993; Yukl, 1998). Meta-analyses conducted to make sense of the bewildering array of research have shown disappointing results. In one meta-analysis, for example, only 6 moderators of 16 tested met the criteria of moderator in the

hypothesized direction (Wofford & Liska, 1993). Based on these limitations, reviewers have suggested that the theory has not been adequately tested (Evans, 1996; Yukl, 1998; Yukl & VanFleet, 1990).

Multiple-Linkage Model

Yukl proposed a taxonomy of managerial behavior categories in conjunction with another contingency theory of leadership, known as the multiple-linkage model (Yukl, 1994; 1998). This leadership model is more complex and comprehensive than many other theories, including Path-Goal theory, since it includes a wider range of intervening and situational variables and leader behaviors.

Empirically Based Categories and Complex, Theoretically Based Relationships with Effectiveness

The 14 taxonomic categories of leadership behavior were derived from research and factor analyses of data from early versions of a survey designed to measure leadership behavior, called the Managerial Practices Survey (Yukl, Wall, & Lepsinger, 1990). Also considered were the findings from other taxonomies of managerial or leader behavior and observed managerial activities. The number of categories postulated had been as high as 22, but several factor analyses and testing of the categorization scheme resulted in collapsing some narrow categories into more general ones. The goal was to keep the taxonomy parsimonious and at a “middle level of generality.” The 14 categories include: planning/organizing, problem solving, monitoring, networking, informing, clarifying, motivating/inspiring, conflict management/team-building, supporting, consulting, recognizing, developing, rewarding, and delegating (Kim & Yukl, 1995).

The behaviors in the taxonomy are all assumed to be relevant for managerial effectiveness, but their relative importance can differ across situations, and they can be enacted in various ways in different situations (Kim & Yukl, 1995; Yukl & VanFleet, 1990; Yukl et al., 1990). Generally speaking, the model states that the effects of leader behavior on work group performance are mediated by both individual-level intervening variables, such as subordinate effort, role clarity, and

ability, and group-level intervening variables, such as work organization, teamwork, resources, and external coordination. Situational variables like the nature of the task, the characteristics of subordinates, and the external environment are believed to have an effect on the intervening variables and to determine which leadership behaviors are most relevant for a particular manager. A basic hypothesis of the theory is that group performance will improve when a leader takes action to correct any deficiencies in the intervening variables. These actions include both the short-term, reactive variety, as well as the long-term, proactive variety, which could result in delayed effects on the variables (Kim & Yukl, 1995; Yukl, 1998).

The model also predicts that the overall pattern of leader behavior and its relevance to the situation are what determine effectiveness (Kim & Yukl, 1995). Since there is no one-to-one correspondence between leader behaviors and intervening variables or performance, this seems to imply that certain behaviors are effective in conjunction with each other or that certain sequences of behaviors are effective, depending upon the situation. No specific hypotheses about which behavior patterns “work” in which situations have been made and very little has been written in the published literature regarding this aspect of the model.

Few Studies Conducted

Little empirical research has been conducted on Yukl’s (1994; 1998) model. Yukl et al. (1990) investigated the criterion-related validity of a survey of the 14 behavioral categories in six studies utilizing various samples (i.e., beauty salon managers, first sergeant military cadets, insurance managers, elementary school principals, program leaders in a cooperative extension association, high school department heads) and criteria (e.g., profit margin, observed unit performance in a marching competition, effectiveness ratings based on sales and profit performance, ratings of managerial effectiveness). The pattern of significant correlations differed across the six studies. The category of clarifying was significantly related to different criteria in four out of six of the studies, the most of any of the other categories. (The highest correlation was

with profit margin in a sample of beauty salon managers [$r = .47$, $p < .01$]). In another study involving middle-level managers from a variety of organizations, all 14 subordinate-rated behaviors were significantly correlated with overall effectiveness ratings provided by "bosses" and peers – correlations ranged from .16 to .38 ($p < .01$). Several self-rated behaviors were related to the effectiveness criterion, with the exception of supporting, consulting, rewarding, and delegating. Planning/organizing, problem solving, networking, clarifying, and motivating/inspiring behaviors as rated by both subordinates and the leader him/herself were significantly related to another criterion of rate of advancement. Correlations ranged from .12 ($p < .05$) to .21 ($p < .01$) (Kim & Yukl, 1995). These few studies investigated the behaviors in the taxonomy, but not the hypothesized relationships in the multiple-linkage model. A search of the literature yielded nothing published on the behavior patterns and differences in effectiveness depending on the situation.

In sum, the three theories presented fall short when considering their concepts as potential predictors of supervisory effectiveness. The behaviors of initiating structure and consideration in the Ohio State Studies have recently been found to have moderately strong relationships with various criteria in a meta-analysis of an often-criticized body of work. Path-Goal Theory, building on the results of the Ohio State work and including elements of the expectancy theory of motivation, has not seen strongly supportive research results. The Multiple Linkage Model, integrating concepts from several leadership theories, has not been tested, and several studies of the validity of behaviors in a proposed taxonomy showed different patterns of results.

Chapter 4

THE OPERANT MODEL OF EFFECTIVE SUPERVISION: MONITORING AND PROVIDING CONSEQUENCES AS PREDICTORS

The Operant Model of Effective Supervision (Komaki, 1998), in contrast, is based in theory and research is generally supportive. This model may be of help to practitioners interested in selecting good supervisors based on non-cognitive, interpersonal skills.

Two Behaviors Grounded in Theory Are Hypothesized to be Related to Supervisory Effectiveness

Providing consequences and monitoring performance, derived from operant conditioning theory, are the two major cornerstones of the model. Providing consequences, defined as communicating an evaluation or indicating knowledge of another's performance, was chosen as a component because of the emphasis on a similar behavior in operant conditioning theory (Komaki, 1998). In operant theory, the things that happen after performance (feedback, comments, etc.), known also as consequences, have a powerful impact on what people do. It is believed that consequences affect the probability that a behavior or performance will reoccur – they cause the behavior to happen more or less often in the future (Daniels, 1994; Scott & Podsakoff, 1985; Skinner, 1974). Thousands of experiments at all levels of the phylogenetic scale, in fact, have demonstrated the efficacy of consequences (Honig, 1966; Honig & Stadden, 1977). Significant improvements in performance have resulted (and been sustained over time) when consequences have been arranged to be frequent and contingent on performance. In effect, consequences are considered the major motivating factor of behavior, and were thus highlighted in the Operant Model as well.

Monitoring performance, the second major behavior in the Operant Model, is defined as the collection or gathering of information about followers' performance. Collecting information on behaviors is part of a three-step process used in operant motivational programs. For example, in a

program to improve occupational safety, the operant approach would be to *specify* the desired safety practices (by reviewing previous accidents and identifying what could be done to avoid recurrences), *collect information* by observing the actual safety practices of workers, and then *provide feedback* (or consequences) by posting the information collected on a graph. Though there is no related construct in operant conditioning theory, the gathering of information is a logical prerequisite to providing contingent consequences. It is for these reasons that monitoring performance is central to the supervisory model. It was reasoned that managers who monitor will more likely possess accurate and timely information that they can use to provide meaningful consequences (Jensen & Komaki, 1993; Komaki, 1986, 1998). Without this information about performance, no manager could provide feedback or consequences that are directly linked to employee performance. The central hypothesis of the model is that monitoring and providing consequences distinguish between effective and marginally effective managers, in that effective managers are likely to spend more time on these activities (Komaki, 1998). In other words, managers judged to be more effective are expected to monitor and provide consequences much more than their less effective counterparts.

Research Results Generally Support the Operant Model

Unlike some of the other leadership theories presented previously, a fair amount of data exists that shows relationships between the behaviors in the Operant Model and supervisory effectiveness.

Monitoring and/or Providing Consequences Were Found to be Important in Field Studies

Research has consistently supported the tenet that monitoring and/or providing consequences distinguish between effective and marginally effective supervisors (Komaki, 1998). The time leaders spent monitoring or providing consequences, or both, were related to criteria in every one of seven field studies in which the relationship between supervisory effectiveness and the behaviors were assessed. These studies were conducted in different settings with different samples

– insurance and newspaper managers, sailboat skippers, government agency supervisors, construction site managers, and police sergeants -- and across lines of nationality – American, Finnish, and Australian (i.e., Brewer, Wilson, & Beck, 1994; Jensen & Komaki, 1993; Komaki, 1986; Komaki, Desselles, & Bowman, 1989; Komaki, Hyttinen, & Immonen, 1991; Komaki, Reynard Minnich, Lee, & Wallace, 1997). (Refer to Table 1 for a summary of these studies.) Leaders' behaviors were measured primarily through an observational coding instrument (or a version of it), known as the Operant Supervisory Taxonomy and Index (OSTI), in which observers watch leaders while they work in their natural environment and record what they say and do. Measures of supervisory effectiveness have included manager rankings and ratings of performance, finishing order in sailboat races, and length of time for a sailing crew to hoist a sail.

Four of the six field studies showed that effective leaders monitored more than lackluster ones, with correlations between monitoring and effectiveness ranging from .33 to .51 (i.e., Brewer, et al., 1994; Komaki, 1986; Komaki, et al., 1989; Komaki, et al., 1997). In four of the six studies, significant correlations between consequences and effectiveness ranged from .37 to .51 (i.e., Brewer et al., 1994; Jensen & Komaki, 1993; Komaki, et al., 1989; Komaki et al., 1991). For example, the percent of time spent monitoring distinguished between groups of 12 effective and marginally effective managers (as rated by their managers) in an insurance company ($M = 2.9$ percent, effective group; $M = 2.0$ percent, marginal group; $t = 2.59$, $p < .01$) (Komaki, 1986). (This extreme groups analysis technique with small sample sizes, which was used in two of the studies, may have exaggerated results.) In a sailboat regatta, skippers' monitoring was related significantly to series standings ($r = -.51$, $p < .05$) and providing consequences was related to both series standings ($r = -.47$, $p < .05$) and a judgmental measure of team motivation calculated from coaches' ratings and rankings of skippers ($r = -.42$, $p < .05$) (Komaki et al., 1989).

What can be observed from the group of research studies as a whole is that effective managers spent similar amounts of time performing some of the same kinds of activities as marginal

managers, such as interacting, discussing work, and talking about performance-related matters other than providing consequences. The key differences between the groups were in the amount of time they spent monitoring and/or providing consequences. Effective managers sought out information about their subordinates' work performance, in some cases through observing their workers in action and examining their work. They also provided feedback by indicating what they liked and disliked or making casual comments showing they knew what was happening.

Monitoring and Providing Consequences Made a Difference in Laboratory and Field Experiments

The preceding studies captured leaders' behaviors during the day-to-day happenings of the "real" work world. No attempts were made to alter or manipulate the situation or leaders' behaviors. The obvious strength of conducting research in environments to which the results are applicable can also be a disadvantage in that the manipulation necessary to make conclusions about cause and effect relationships may not be possible. Four studies of behaviors in the Operant Model, however, have been conducted in the laboratory, with random assignment of subjects to different experimental conditions. (Refer to Table 1 for a summary of these studies.)

Monitoring, the focus of several of the lab studies, improved subordinate performance. Larson and Callahan (1990) used a 3 X 2 factorial design with subjects working at proofreading and alphabetizing tasks for two hours. Subjects were exposed to one of three patterns of supervisory behavior (conditions): (1) monitoring only, in which the experimenter came into the room approximately every 20 minutes to collect and examine the quantity of work on one of the tasks (half of the subjects were monitored on alphabetizing, half on proofreading); (2) monitoring plus consequences, in which the experimenter monitored as above, but also gave subjects information about their performance on the monitored task; and (3) control, in which subjects worked continuously for two hours without interruption. The researchers found that subjects in the monitoring-only condition (M = .33) completed significantly more work on the monitored task than those in the control condition (M = -.12), t = 2.15, p < .05. Using the same tasks, two more complex

experiments were conducted by Brewer and Ridgway (1998). One study used a 4 (supervisory behavior) X 2 (output identifiability) X 2 (task [within subjects factor]) mixed design with supervisory behavior conditions of monitoring the alphabetizing task, monitoring the proofreading task, monitoring both, and control. For all monitoring conditions, the researchers found that subordinates' performance quantity increased on the task that was perceived to be monitored more frequently. Similar results were found in the second experiment, which used a 5 (monitoring type) X 2 (presence of monitoring [within subjects factor]) X 2 (task monitored) mixed design. Increased quantity on the monitored task coupled with reduced quality on the monitored task and reduced quantity and quality on the unmonitored task were found in monitoring conditions that differed in terms of the method of monitoring. Effects were more pronounced in conditions in which the monitoring specifically emphasized performance quantity.

Furthermore, two studies showed that providing consequences in addition to monitoring improved subordinate performance. Work output was higher in Larson and Callahan's (1990) monitoring-plus-consequences group as compared to the group who was only monitored ($t = 2.41$, $p < .01$). Monitoring and providing consequences together were found to be important in both performance and attitudes in a study by Komaki, Desselles, & Schepman, (1988). Subordinates who were both monitored by and provided with feedback from their supervisor on a bulk mail sorting task performed better than those who received only feedback ($M = 94.2$ percent, $M = 82.8$ percent, respectively). The subordinates in the monitor-and-consequence condition also had the most favorable reactions to the manager, spending significantly more time talking positively ($M = 8.5$ percent) and less time talking negatively about him/her ($M = 8.4$ percent), as compared to both the monitor-only ($M = 4.3$ percent positively, $M = 14.5$ percent negatively) and consequence-only conditions ($M = 2.1$ percent positively, $M = 16.6$ percent negatively). (Probability levels for the Komaki et al. [1988] study were unavailable.)

Table 1

Summary of Selected Research Involving the Operant Model of Effective Supervision

Reference	Subjects/Setting	Measures	Results
		Field Studies	
Brewer, Wilson, & Beck (1994)	20 Australian police patrol sergeants	• Observation of supervisory behaviors • Performance rating on 13 dimensions	• Patrol sergeants of higher performing teams spent more time monitoring the performance of their officers ($r = .40$, $p < .05$). • Highly rated sergeants used the self-report method of monitoring ($r = .47$, $p < .05$). • Neutral consequences were related to effectiveness ($r = .51$, $p < .05$), but consequences in general were not ($r = .35$, n.s.).
Jensen & Komaki (1993)	33 newspaper managers	• Observation of supervisory behaviors • Judgmental measure of motivational effectiveness calculated from ratings and rankings provided by superiors, collected twice	• Consequences were related to effectiveness ($r = .37$, $p < .05$) and distinguished between effective ($M = 6.4\%$, N = 8) and marginal ($M = 4.4\%$, N = 8) groups of managers, $t = 3.42$, $p < .05$. • Antecedents fit the definition of a suppressor variable, i.e., were not related to effectiveness ($r = .01$, n.s.), had a negative beta weight (ß = -.47, $p < .05$) in a multiple regression analysis, and were related to consequences ($r = .69$, $p < .05$).
Komaki (1986)	24 insurance managers	• Observation of supervisory behaviors • Judgmental measure of motivational effectiveness calculated from ratings and rankings provided by superiors • Judgmental measure of job knowledge calculated from ratings and rankings provided by superiors	• Monitoring distinguished between effective ($M = 2.9\%$) and marginal ($M = 2.0\%$) groups of managers ($t = 2.59$, $p < .01$), and did not between knowledgeable ($M = 2.5\%$) and less-knowledgeable ($M = 2.4\%$) groups ($t = 0.46$, n.s.) • Work sampling was the only method of monitoring to distinguish between effective ($M = 0.46\%$) and marginal ($M = 0.09\%$) groups, $t = 3.32$, $p < .01$.

Table 1, cont.

Reference	Subjects/Setting	Measures	Results
Komaki, Desselles, & Bowman (1989)	19 sailing crews in a round-robin regatta	• Observation of supervisory behaviors • Series standings • Judgmental measure of crew handling (team motivation) calculated from ratings and rankings provided by coaches	• Series standings were related to providing consequences ($r = -.47$, $p < .05$) and monitoring ($r = -.51$, $p < .05$). • Providing consequences was related to measure of crew handling ($r = -.60$, $p < .05$).
Komaki, Hyttinen, & Immonen (1991)	31 Finnish construction crew managers, 16 supervisors in government agency	• Observation of supervisory behaviors • Subordinate ratings of how well construction crew supervisors met the demands of the job • Ratings of job satisfaction of government workers • Ratings of mental well-being of government workers	• The effective group of construction managers spent more time providing consequences (*M* = 5.3%, N = 8) than the lackluster managers (*M* = 3.6%, N = 8), $t = 1.7$, $p = .05$. • The product of monitor and consequence scores distinguished between effective and marginal construction managers (*M* = 59.8% and *M* = 33.2%, respectively), $t = 1.9$, $p < .05$. • A specific aspect of worker satisfaction (job security) in the Finnish government agency was associated with supervisor monitoring ($r = .82$, $p < .05$) and the product of monitoring and providing consequences ($r = 87$, $p < .01$). • Workers' positive feelings about their mental effort, sociability, self-confidence, and state of mind were related to providing consequences (*r*s ranged from .73 [$p < .05$] to .89 [$p < .01$]). • Workers positive feelings about their mental effort, self-confidence, state of mind, and tolerance were associated with the product of monitoring and providing consequences (*r*s ranged from .71 to .80, $p < .05$).

Table 1, cont.

Reference	Subjects/Setting	Measures	Results
Komaki, Reynard Minnich, Lee & Wallace (1997)	28 practice sailboat races	• Observation of videotaped supervisory behaviors • Seconds to chute hoist as measure of sailing team effectiveness • Seconds after start as measure of sailing strategy	• Monitoring performance was related to quicker chute hoist times ($r = -.33$, $p = .05$). • Monitoring performance was related to the seconds after start strategy measure in the opposite direction ($r = .44$, $p < .05$).
		Laboratory Experiments	
Brewer (1995)	124 undergraduate students	• Number of lines read as measure of performance quantity on proofreading task • Number of misspelled words identified minus number of correctly spelled words identified inappropriately, divided by the total number of lines read as a measure of performance quality on proofreading task • Total number of words alphabetized as measure of performance quantity on alphabetizing task • Overall percentage of words placed in correct order as measure of performance quality on alphabetizing task	• Participants in conditions in which they worked in a group setting and were monitored individually increased quantity on the monitored task at the expense of the unmonitored task ($F(1,120) = 10.16$, $p < .01$ for interaction of presence of monitoring and performance measure; $t(118) = 2.62$, $p < .01$ for quantity in individual monitoring condition). • Participants in conditions in which outputs of the group as a whole were monitored made no performance adjustments ($t(118) = -0.35$, n.s. for quantity in group monitoring condition).

Table 1, cont.

Reference	Subjects/Setting	Measures	Results
Brewer & Ridgway (1998)	192 undergraduate students 150 undergraduate students	• Number of lines read as measure of performance quantity on proofreading task • Number of misspelled words identified minus number of correctly spelled words identified inappropriately, divided by the total number of lines read as a measure of performance quality on proofreading task • Total number of words alphabetized as measure of performance quantity on alphabetizing task • Overall percentage of words placed in correct order as measure of performance quality on alphabetizing task	• For all three monitoring conditions (either one of the tasks monitored or both), performance quantity increased on the task that was perceived to be monitored more frequently, as compared to unmonitored controls ($F(3,134) = 6.58$, $p < .001$ for interaction of supervisory behavior and task; univariate test statistics not presented). • In two of the monitoring conditions, the quantity gains on the monitored task were outweighed by decrements in quantity and quality of the unmonitored task. • Similar results were found in the second experiment. Increased quantity on the monitored task with reduced quality on the monitored task and reduced quantity and quality on the unmonitored task were found in monitoring conditions that differed in terms of the method of monitoring ($F(4,140) = 4.12$, $p < .01$ for interaction of supervisory behavior and presence of monitoring). Effects were more pronounced in conditions in which the monitoring specifically emphasized performance quantity.

Table 1, cont.

Reference	Subjects/Setting	Measures	Results
Komaki, Desselles, & Schepman (1988)[a]	73 undergraduate students	• Performance quality on a sorting task • Observation of subordinate attitudes during videotaped discussions	• Performance was significantly better in the monitoring and providing consequences condition (M = 94.2%) than in the consequence-only condition (M = 82.8%), but not in the monitoring-only condition (M = 86.0%). • Employees in the monitoring and consequences condition exhibited more positive (M = 8.5%) and fewer negative (M = 8.4%) attitudes about the manager as compared to monitor- (M = 4.3% pos., M = 14.5% neg.) and consequence-only (M = 2.1% pos., M = 16.6% neg.) conditions.
Larson & Callahan (1990)	90 undergraduate students	• Number of lines read as measure of performance quantity on proofreading task • Number of misspelled words minus number of correctly spelled words circled (inappropriately), divided by total number of lines read as a measure of quality on proofreading task • Total number of word lists alphabetized as measure of quantity on alphabetizing task • Overall percentage of words correctly sorted and placed in proper alphabetical order as measure of quality on alphabetizing task	• Subjects in the monitoring-only condition (M = .33) completed significantly more work on the monitored task than those in the control condition (M = -.12), $t = 2.15$, $p < .05$. • Subjects in the monitoring-plus-consequences condition (M = .82) completed significantly more work on the monitored task than did those in the monitoring only condition, $t = 2.41$, $p < .01$. • The monitored task received significantly higher importance ratings from those in the monitoring-only condition (M = 4.79) versus those in the control condition (M = 4.07), $t = 2.34$, $p < .05$. • The monitored task received significantly higher importance ratings from subjects in the monitoring-plus-consequences condition (M = 5.87) than from those in the monitoring-only condition, $t = 3.45$, $p < .01$.

Table 1, cont.

Reference	Subjects/Setting	Measures	Results
		Field Experiment	
Methot, Williams, Cummings, & Bradshaw (1996)	1 manager, 4 supervisors, 7 direct care staff, and 16 clients in residential human service facility	• Observation of supervisory behaviors • Observation of client target behaviors	• After receiving training, increases in the use of contingent performance consequences were seen for the manager and all supervisors. • Six of 7 staff members also increased their use of contingent consequences in interactions with clients. • Desired decreases in target behaviors occurred for 9 or 16 clients, and desired increases for 8 of 13 clients.

[a] Probability levels were not available for this study.

These laboratory experiments, though conducted with undergraduate students for the most part, suggest that managers can have a meaningful impact on the performance and attitudes of their employees. They indicate monitoring how employees are doing can improve performance, especially if employees perceive that their performance is being examined. Workers whose supervisors monitored and provided consequences performed better than those whose supervisors monitored alone or provided only consequences. There is also some evidence that workers whose supervisors monitored and provided consequences had more positive attitudes about their supervisors.

In addition, a field experiment was conducted in a residential facility for persons with developmental disabilities (Methot, Williams, Cummings, & Bradshaw, 1996). (Refer to Table 1.) One manager and four supervisors of direct care staff members were trained in negotiating goals with staff, monitoring performance, and providing contingent feedback. Using a multiple baseline design, the researchers collected information on supervisory behaviors with a version of the OSTI both before and after the delivery of the training session. After training, increases in the use of contingent performance consequences were observed for the manager and all supervisors in interactions with their subordinates. In addition, 6 of 7 direct care staff members increased their use of contingent consequences in interactions with their clients. Observational measures of client behaviors showed that some changes in clients' behavior also occurred after training -- desired decreases in target behaviors were observed for 9 or 16 clients and desired increases for 8 of 13 clients. Thus, the study showed that supervisors providing consequences resulted in desired outcomes (changes in clients' behaviors).

In sum, the Operant Model of Effective Supervision suggests that two non-cognitive behaviors are key to effectively managing employees – monitoring performance and providing consequences. Research, conducted both in the field and laboratory, has shown that the quantity and quality of these behaviors does indeed make a difference.

Most Research on the Model Utilized Measures that are Impractical for Applied Practice

The Operant Model seems well-suited to address the gap in non-cognitive predictors used in practice because the behaviors are interpersonal in nature, derived from theory, clearly defined (as indicated by interrater agreement checks, which have been conducted in all studies), and describe how supervisors can motivate their employees. There is also some empirical evidence of their validity. Leaders' behaviors, however, in most of the previous studies of the Operant Model were measured through an observational instrument (called the Operant Supervisory Taxonomy and Index, or OSTI) requiring 10 hours of watching leaders while they work, recording verbatim what they say and describing their actions. Although research has shown that this measure is a rich, sensitive, and reliable measure of leader behaviors (e.g., Komaki, Zlotnick, & Jensen, 1986), it is not well-suited to be used as an operational predictor. The great amount of time necessary to observe leaders in action is a significant detractor to using the observational measure in selection situations.

When selecting from among a group of candidates for a job, measures of knowledge, skills, or abilities need to be of short duration, both in terms of completion by candidates and for obtaining results. For most positions, neither candidates nor organizations can afford to spend several days going through a lengthy battery of selection processes. Also, for non-cognitive interpersonal skills especially, it is beneficial to assess these directly in somewhat complex situations representative of the work environment. For example, better data could be attained by assessing candidates' supervisory behaviors in an exercise simulating the work environment than by asking candidates how long they had supervised others.

Many selection tools/methods, such as biographical data measures, paper and pencil tests, interviews, situational judgment tests, and assessment centers, are available to practitioners. Of these tools (which have all been shown to be valid predictors of job performance), some are better than others for taking a short amount of time to complete and for assessing skills directly. The

collection of background information or biographical data (biodata), though non-cognitive in nature, is not usually about supervision. Information about past experiences, preferences, interests, and opinions is collected from candidates via questionnaire and can be completed and scored relatively quickly (depending on the number of questions) (Guion, 1998; Lefkowitz, Gebbia, Balsam & Dunn, 1999; Robertson & Smith, 2001). However, typical biodata questions, such as "Were you ever class president?", "How old were you when you got your first paying job?", "Compared to most of your friends, how self-confident are you?", or "How nervous are you if you have to give a speech to a community group?" are not measures of actual skills. Paper and pencil tests, similarly, are often of short duration and can be scored quickly. They are not measures of actual skills, but commonly measures of knowledge, cognitive ability, personality, or integrity. Interviews, probably the most widely used of all selection tools, are also relatively short experiences for candidates, though scoring may take a bit longer than paper-and-pencil tests depending on such factors as the number of interviewers, the number of questions asked, the structure of the interview, and the rigor of the scoring/rating and compilation process. The data collected in interviews do not capture the actual, demonstrated skills of candidates. At best, when interviews are behavioral (asking interviewees to describe what they did and how they did it in specific situations), they yield applicants' descriptions of their behaviors, or self-reports of skills (which could have varying levels of veracity). Though considered simulations of the work environment, situational judgment tests are also *not* measures of the actual skills of applicants. In this method of assessment, candidates are presented with written or video-based hypothetical scenarios or vignettes and from a list of possible responses, asked to rate the effectiveness of each or choose which would be how they would most and/or least likely respond (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001; Motowidlo, Dunnette, & Carter, 1990). Since they are similar to multiple choice pencil and paper tests, situational judgment tests do not require a large amount of time for respondents to take them or for them to be scored. Assessment centers, on the other hand,

assess actual skills through group- and/or individual-based situational exercises that require respondents to react to rich and sometimes complicated work-related stimuli in whatever way they see fit. While candidates participate in a leaderless group discussion or complete tasks in a simulation of the work environment, for example, assessors rate their level of skill on various dimensions. Comprised of a variety of such activities and others such as objective or projective tests and interviews, entire assessment centers require large time commitments for both candidates and assessors. They are often full- or multi-day events, and with the large number of dimensions assessed, require significant effort to score and produce results.

An in-basket exercise is an alternative method that does not require an overly large time commitment and is also an assessment of candidates' demonstrated skills. In-baskets are simulations in which respondents are given fictional memos, phone messages, letters, etc. – things typically found in managers' "IN" bins. Respondents take on the role of a manager and respond to the items, make decisions, or attempt to resolve the issues created by the items (Brass & Oldham, 1976; Gill, 1979; Schippmann, Prien, & Katz, 1990). (They are frequently included as an exercise in assessment centers and not typically used or researched on their own. In fact, Schmidt and Hunter's [1998] review of selection method validity and utility did not include data for in-baskets.) In-basket exercises themselves typically do not take more than a few hours for candidates to complete. They do, however, require more time for scoring than a pencil-and-paper test because respondents can answer in a seemingly infinite number of ways. Scoring must be done "by hand," reading all responses and scoring on relevant dimensions.

An In-basket Assessment Based on the Operant Model was Developed and Tested with One Sample

In response to these practical measurement concerns in selection, an in-basket measure of the behaviors in the Operant Model, the Operant Supervisory In-Basket Assessment (OSIBA), was developed (Komaki, Newlin, & Desselles, 1990). The OSIBA is well-suited for real selection situations – it captures actual behaviors to a common set of stimuli in a relatively short amount of

time (one to one-and-a-half hours), as compared to the 10 hours required for direct observation in situ. The exercise contains 22 items, ranging in content from telephone messages from clients complaining about poor service to reports from accounting on production costs. Some items require the respondent to "directly" monitor and provide consequences to his/her employees, while others involve situations where the respondent makes recommendations to others.

One study has investigated the instrument to date (Komaki et al., 1990), utilizing a small sample of computer managers. Inter-scorer agreement, calculated as a percentage, averaged 83.5 percent. Construct validity of the measure was investigated in a multi-trait, multi-method matrix of correlations between scores on monitoring and providing consequences on the OSIBA; direct observations of behavior measured via the OSTI; and coded electronic mail messages.

Convergent validity evidence, as indicated by the correlations between the same traits (behaviors) measured with different methods, was fairly good. On average, convergent validity coefficients were about .49. Scores for monitoring and providing consequences on the in-basket were significantly correlated with the amount of time spent on the same observed categories (monitors, $r = .57$, $p < .05$; consequences, $r = .60$, $p < .05$). (The relationship between the in-basket and the on-the-job display of behaviors was the focus of the study. Coding the behaviors in e-mail messages was an exploratory measure and not expected to yield any significant results.) Providing consequences in e-mail was significantly correlated with providing consequences as measured through direct observation, $r = .54$, $p < .05$. Correlations between e-mail and in-basket scores on providing consequences and e-mail, in-basket, and observed scores on monitoring were of fairly high magnitude and in the expected direction, but did not reach significance.

Discriminant validity was also fairly good. Low, nonsignificant correlations between measures of *different* traits using the same instrument is an indication of discriminant validity, and these correlations should be smaller than the correlations among measures of the same trait across

different measures. Correlations between different behaviors measured with the same measure were .19, .17, and -.18. Refer to Table 2 for the results.

Table 2

Multi-Trait Multi-Method Matrix for Monitors and Consequences Measured via Direct Observation, an In-Basket Exercise, and Scoring of E-mail Messages

		Observation		In-basket		E-mail	
		M	C	M	C	M	C
Observation	M		.19	.57*	-.26	.50	-.35
	C			.46	.60*	.21	.54*
In-basket	M				.17	.35	-.30
	C					-.23	.36
E-mail	M						-.18
	C						

Note. From “Walking on the Wild Side: Criterion-related Validation of an In-basket Exercise of Supervisory Behaviors,” by J.L. Komaki, M.H. Newlin, & M.L Desselles, 1990, In M.H. Newlin (Chair), *Simulated performance assessment: Fact or fantasy?*, Symposium presented at the conference of the Society for Industrial and Organizational Psychology, Miami. Reprinted with permission of the author.

M = Monitoring. C = Consequences. $N = 12$. * $p < .05$

In sum, the research conducted on the OSIBA was geared to understanding whether the behaviors measured in the in-basket are reflective of observations of what a person would do on the job. The OSIBA does seem to be a reflection of on-the-job behaviors, given the high correlations of monitoring and providing consequences on the in-basket with the same behaviors measured via observation. The strong relationships found between the two behaviors measured via on-the-job observation and various effectiveness measures in the research on the Operant Model indirectly indicate that the OSIBA, too, might be good predictor of supervisory effectiveness. However, no

research has been done to directly link monitoring and providing consequences measured via the in-basket to effectiveness.

The Present Study is a Replication of Previous Work, but Using an In-basket Measure of Leader Behaviors

The present study is the first attempt to link monitoring and providing consequences measured via the OSIBA, a measure suitable for use in selection situations, to supervisory effectiveness. Thus, the research is a replication of previous field studies, but instead of measuring the behaviors via direct observation, an in-basket assessment will be used. **Hypothesis 1a**, consistent with previous work, is that **monitoring and providing consequences will each be positively and significantly related to supervisory (motivational) effectiveness**.

While one would expect that monitoring and providing consequences will be related to supervisory effectiveness, one would also expect that the behaviors will be *unrelated* to other measures *not* having to do with motivating employees (e.g., technical savvy, knowledge of work processes and jobs). In other words, monitoring and providing consequences would be expected to be orthogonal to technical, industry, or job knowledge.

Supervisors' technical knowledge was measured in only one study of the Operant Model (Komaki, 1986). In this research, managers in an insurance company were rated and ranked by the vice presidents in their divisions in terms of their "technical expertise." Results showed that monitoring did not make a difference – extreme groups of managers formed on the basis of ratings and rankings spent essentially the same percent of time monitoring on average. Thus, supervisors who have high levels of technical knowledge might monitor little or a lot and might provide few or many consequences. **Hypothesis 1b** of the present study, therefore, is that **monitoring and providing consequences will each be unrelated to technical expertise**.

Additive and Multiplicative Combinations of Monitoring and Providing Consequences May be Useful for Predicting Supervisory Effectiveness

While monitoring and providing consequences individually have been found to be related to effectiveness and are predicted to be in the current study, one could ask whether there might be some combination of the two behaviors that might also be effective. Komaki (1998) discussed various options for combining monitoring and consequence scores based on aggregation rules in the field of economics and their implications for the relationship between the behaviors. One possibility is to take the minimum score of the two. In this case, the lowest score on either monitoring or providing consequences is the best indication of the behavior of a manager. This method implies that a person is only as strong as his/her weakest aspect. Another possibility is to take the maximum score of the two, which infers the more monitoring or consequences the better. (This, however, is incompatible with the notion that effectiveness declines above a certain limit of monitoring or providing consequences.) Adding the two behaviors is another option. With this combination, one behavior of a manager can make up for a lack of the other. For example, providing many consequences could compensate for little monitoring. Multiplying the scores on the behaviors is yet another possible combination. In this case, low scores on one cannot be compensated for by high scores on the other -- the behaviors actually interact to produce a “synergistic reaction.” In other words, some level of monitoring makes providing consequences more effective, and vice versa.

In selection situations, additive combinations of predictors are quite common. Though overall additive scores are not directly calculated, information from several predictors is summed when analyzed via multiple regression (Guion, 1998). This commonly used statistical technique utilizes an additive composite of predictor scores in the equation that is the basis of the analysis – predictor scores are weighted optimally and summed to form a composite. The analysis yields a multiple correlation coefficient, indicating the strength of the relationship between the criterion and

the composite. (The squared multiple correlation is an index of how much criterion variance is accounted for by the set of predictors.)

In the present study, the additive relationship between monitoring and providing consequences will be tested using the traditional multiple regression analysis. Monitoring and providing consequences will be predictors and a measure of supervisory effectiveness will be the criterion in the analysis. Thus, **Hypothesis 2a** is that **an additive, weighted linear combination of monitoring and providing consequences will be positively and significantly related to supervisory effectiveness**.

A somewhat different feature of this research study is that the multiplicative relationship between monitoring and providing consequences will also be tested. In selection research, a multiplicative relationship between two predictors/variables, called an interaction, is rarely taken into account. Interaction (or moderation) occurs when a variable influences the relationship between two other variables (Cortina, 1993; Ganzach, 1998; Guion, 1998). In the selection arena, researchers have looked for demographic and other variables that might have an effect on the relationship between various predictors and criteria. These searches have been relatively fruitless, leading many to give up on finding interactions (Guion, 1998). Though conceding that interactions (or moderators) in personnel decisions are “rarely reported and rarely replicated” (p. 345), Guion (1998) offers hope that interactions might be more likely to be found “after serious thinking, hypothesizing, and theory formulation than after searches among variables for which there is no useful rationale” (p.346).

The Operant Model, however, suggests that monitoring and providing consequences interact. A second proposition of the model, known as the conjunctive hypothesis, is that monitoring and providing consequences “work together” for superior performance (Komaki, 1998). The nature of the interaction may be such that when monitoring is high, the relationship between providing consequences and supervisory effectiveness is positive, but when monitoring is low,

providing consequences will be unrelated to effectiveness. With this theoretical rationale, the interaction or multiplicative combination of monitoring and providing consequences might be useful to take into account in predicting supervisory effectiveness.

This hypothesis has not been addressed much in the literature on the model, nor been adequately tested. A multiplicative combination of monitoring and providing consequences was computed in only one study (Komaki et al., 1991). This combination was found to distinguish between effective (M = 59.8%) and marginal (M = 33.2%) construction managers ($t = 1.9$, $p < .05$) and to be correlated with government workers' satisfaction with job security and several aspects of mental well-being (rs ranged from .71 to .87). The most common (and appropriate) way of testing for an interaction, adding a multiplicative term to a multiple regression equation as a predictor in addition to each of its components (Cortina, 1993; Guion, 1998), was not carried out.

In the present study, a hierarchical multiple regression including monitoring, providing consequences, and the multiplicative combination of the two, will be performed. If the interaction is present, the total amount of variance explained when the multiplicative term is part of the regression equation will be significantly more than that explained when the regression equation consists of the component parts of the interaction alone. Therefore, **Hypothesis 2b** of the present study is that **the interaction of monitoring and providing consequences will add significantly to the prediction of supervisory effectiveness, over and above monitoring and providing consequences alone**.

Additive and multiplicative combinations of monitoring and providing consequences will also be examined in a second way. This will be an exploratory analysis of relationships between effectiveness and scores computed for each respondent. Individuals' scores for monitoring and providing consequences will be added together to form a single overall "additive score" and multiplied to form a single "multiplicative score," and these will be correlated with effectiveness. No formal hypotheses are offered, but it is expected that both scores will be significantly related to

effectiveness. Because multiple regression weights the predictors optimally and statistically controls for the spurious associations between the predictors, the correlations will not be as high as those found in the multiple regression analyses. Also, if the interaction between monitoring and consequences is found, the relationship between the multiplicative score and effectiveness may be stronger than for that of the additive score.

In sum, the Operant Model of Effective Supervision suggests two behaviors, monitoring and providing consequences, could be useful in predicting who among a group of candidates would be successful as a supervisor. The model is based in theory and research, both in the field and laboratory, has generally been supportive of the hypothesis that the behaviors distinguish among effective and marginal managers. Little work has been done, however, with measures that practitioners could use in real selection situations. In addition, the model hypothesizes that monitoring and providing consequences may interact, and this hypothesis has not been tested thoroughly. It is possible that the multiplicative combination of the two behaviors could aid in the prediction of effectiveness.

Chapter 5

DIRECTIVES AS A NONTRADITIONAL PREDICTOR

We turn now to another behavior identified, though not highlighted, in the Operant Model of Effective Supervision -- providing directives (or performance antecedents). This behavior could also potentially be used as a predictor, though not in a traditional way, in combination with monitoring and providing consequences. Providing antecedents, though controversial, may be useful in improving the prediction of supervisory effectiveness.

Suppressor Variables, Though Nontraditional and Uncommon, Improve Prediction

Traditional predictors in a selection battery are related positively to the criterion (are valid) and, ideally, have low correlations with other variables and predictors. Prediction (in terms of a squared multiple correlation coefficient) can be improved by adding valid predictors that are not correlated highly with each other, which account for more unique variance in the criterion. However, valid predictors may contain invalid, contaminating variance components, and a predictor that is *not* related to the criterion but is correlated with this contamination in other predictors may improve prediction (Guion, 1998). This invalid predictor actually accounts for (or “suppresses”) unwanted sources of variance from other predictors to which it is related, and is known as a suppressor variable (Conger & Jackson, 1972; Horst, 1941; Lubin, 1957; Smith, Ager, & Williams, 1992; Tzelgov & Henik, 1991). Prediction is improved by the statistical removal of the irrelevant variance from the predictors in the regression equation.

Horst (1941) identified a suppressor variable in his work on selecting test pilots during World War II and noted how the multiple correlation increased when verbal scores were included in the regression equation with mechanical, spatial, and numerical ability test scores. He found that verbal scores were not related to the criterion, but were correlated with mechanical, spatial, and numerical ability test scores, and had a negative beta weight in the multiple regression analysis, the

three elements of a classic suppressor. Horst (1966) explained the results by suggesting that "verbal ability of a high order is not essential for success. It *was* necessary, however, in order to perform successfully on the paper and pencil tests used to measure mechanical, spatial, and numerical ability" (p. 355). He concluded, "to include the verbal score with a negative weight served to suppress or subtract the irrelevant ability, and to discount the scores of those who did well on the test simply because of their verbal ability" (p. 355).

There is some controversy over the use of suppressors in prediction. Suppressors are not commonly found, and when they are, it is often difficult to interpret them meaningfully, or replicate their occurrence (Cascio, 1987; Guion, 1998; Licht, 1995). Also, because it is traditional to use valid predictors in selection, to select someone on the basis of a score that has no relationship with how well they may do on the job seems counterintuitive. Some have argued further that efforts would be better spent on identifying and developing valid predictors because the predictive gains from suppressors are minimal (Cascio, 1987; J. Lefkowitz, personal communication, January 30, 2007). Despite the controversy, some contend that suppressors are still useful in improving prediction (Tzelgov & Henik, 1991). Put simply, the inclusion of suppression relationships in the prediction equation usually leads to a higher multiple correlation. When meaningfully interpreted, as in the case of Horst's (1941) verbal ability, the suppressor can help practitioners make more accurate selection decisions. (The addition of verbal ability increased the amount of variability in test pilot performance explained by the set of predictors. Had verbal ability *not* been included in Horst's battery, people who scored well on mechanical, spatial, and numeric ability tests but actually did not possess high levels of these qualities might have been selected as pilots.)

Another Behavior Identified in the Operant Model Fits the Definition of a Classic Suppressor Variable

Two behaviors, monitoring and providing consequences, were hypothesized to distinguish between effective and marginally effective supervisors in the Operant Model. Results from a number of field studies generally support this hypothesis. However, one study has shown that with monitoring and providing consequences, a third behavior identified in the Model – communicating expectations of performance or providing "antecedents" – fit the definition of a classic suppressor variable (Jensen & Komaki, 1993).

In the study, which was summarized in Table 1, Jensen and Komaki (1993) performed correlational and multiple regression analyses on data from a sample of newspaper managers from several departments at three organizational levels. Observations of monitoring, providing consequences, and delivering antecedents were predictors and the criterion was average overall supervisory effectiveness (based on ratings and rankings of effectiveness by superiors). (See Tables 3 and 4 for the results.)

Table 3

Correlations among Selected Behavior Categories and Supervisory Effectiveness

Category	Supervisory Effectiveness	Antecedents	Monitors
Antecedents	.01		
Monitors	.03	.15	
Consequences	.37*	.69*	.25

Note. From "Beware of Too Many Directives: Spotting a Suppressor Variable in the Operant Model of Effective Supervision" by M. Jensen and J.L. Komaki, 1993, A paper presented at the meeting of the Society for Industrial and Organizational Psychologists, San Francisco, CA, p. 25.

[a]Degrees of freedom = 31. $N = 33$. *$p < .05$.

Table 4

Multiple Regression Analysis for Behavior Categories and Criterion of Supervisory Effectiveness

Category	β[a]	*F*
Antecedents	- .47	4.45*
Monitors	- .08	0.23
Consequences	.71	9.76*

Note. From "Beware of Too Many Directives: Spotting a Suppressor Variable in the Operant Model of Supervisory Effectiveness" by M. Jensen and J.L. Komaki, 1993, A paper presented at the Meeting of the Society for Industrial and Organizational Psychologists, San Francisco, CA, p. 26.

[a]Standardized regression coefficient

R = .50*. *p < .05

Jensen and Komaki (1993) found a multiple correlation coefficient of .50, indicating that the linear combination of monitoring, providing consequences, and directives accounted for 25% of the variability in effectiveness. Similar to the pattern Horst found, providing directives was not related to motivational effectiveness ratings, but was highly correlated with providing consequences, and had a negative beta weight in the multiple regression. Providing antecedents had little or no relationship with average ratings/rankings of supervisory effectiveness (r = .01) and was highly related to delivering consequences (r = .69, p < .05). Antecedents also received a significant, negative beta weight in the multiple regression analysis, beta = -.47, p = < .05. This interesting pattern of results suggested that providing antecedents was acting as a classic suppressor variable. (Analyses comparing multiple correlation coefficients with and without antecedents were not performed, however. Thus, there was no computation of how much antecedents improved the prediction of effectiveness.)

Antecedents are specifically defined within the model as instructions, commands, or reminders that convey an expectation of performance (Komaki, 1998). These could be explicit

instructions, such as “Make the directions clearer and include an estimate of time for travel.” or “Tell Lisa to create the project plan.”; cues, such as “This is a regular reminder for all staff to refer to the service standard memo. All dept. heads should remind their subordinates to read this memo again carefully and follow in their working environment.”; calls for training, such as “If this problem exists as the client indicates, we will need to train your people better.”; or general exhortations, such as “We need to be focused on our mission to be the best in the business.” or “Remember that quality is the key to our success.” Jensen and Komaki (1993) were able to offer a logically appealing explanation for why this behavior may have functioned as a suppressor. They thought that the irrelevant variance being removed might be the propensity to make declarative statements about performance. When managers deliver both directives and consequences, they are telling others about performance. In the case of directives, they tell workers what they should do and how they should perform. In providing consequences, they acknowledge good performance or note when errors have been made – in essence, they tell workers what they think about their performance. Thus, Jensen and Komaki (1993) thought this inclination to talk about performance could explain the high correlation between the two behaviors, but was probably not essential for being an effective manager. If it is assumed that the inclination to talk about performance is the irrelevant variance in the consequence score, then including the antecedent score serves to discount the scores of those who provide consequences simply because they talk about performance.

Antecedents Have a Secondary Role, and Could Be Used as a Nontraditional Predictor

In comparison to monitoring and providing consequences, the role of providing antecedents in the Operant Model is a secondary one. This is for theoretical reasons. Operant conditioning theory is concerned with the things that happen both *before* and *after* the behavior of interest. Consequences, which occur *after* behavior, are believed to have a strong effect on future behavior, providing the fuel for motivation. However, the things that happen *before* behavior, called antecedents, are thought to operate in an informational or cuing fashion. Antecedents indicate what

persons should do or signal that a certain response will or will not be reinforced. They are needed to initiate performance, but are not sufficient to sustain performance over time (Daniels, 1994; Scott & Podsakoff, 1985; Skinner, 1974). To many operant conditioners, antecedents such as "instructions… [are] effective only if the specified responses made in their presence are reinforced, and only if the indicated response goes unreinforced when the instructions are absent" (Whaley & Malott, 1971, p.248). In other words, changes in performance are not expected to be sustained unless antecedents are accompanied by consequences. (A good deal of empirical evidence supporting the question of whether consequences improve performance over antecedents alone has, in fact, been noted in a qualitative review of applied behavior analysis research in work settings [Komaki, Coombs, Redding, & Schepman, 2000]. Of 17 studies addressing the question, 13 showed support, three were inconclusive, and one did not find support.) Thus, consistent with operant conditioning theory, providing antecedents in the Operant Model was not hypothesized to be an effective supervisory behavior for motivating employees to perform their jobs well.

Because behaviors categorized as antecedents have not been hypothesized to distinguish top-performing managers from marginal managers, it seems as if it would *not* be useful to employ them as a predictor. If the goal were to find out which people among a group are effective supervisors, for example, measuring for antecedents would not yield any relevant information for identifying those who are effective. Monitoring and providing consequences, on the other hand, have been shown to have a positive relationship with supervisory effectiveness, and it would seem that they would be useful for predicting effectiveness. As has been discussed, though, adding traditional, valid predictors (that are uncorrelated with each other) to a selection battery is not the only way to improve prediction. Suppressors, invalid predictors that are correlated with another predictor and receive a negative beta weight in a multiple regression analysis, also improve prediction. Giving performance antecedents has been found to function as a suppressor variable in a sample of newspaper managers. It may be that measuring antecedents actually yields information

that can help weed out people who would not be good supervisors. For instance, people with high scores for monitoring and providing consequences, who might "look" like they would be good supervisors, might not be so effective if they give an abundance of antecedents. Thus, providing antecedents might still be useful to consider as a nontraditional predictor in conjunction with more traditional predictors of monitoring and providing consequences.

While Antecedents are Commonly Used in Organizations and Highlighted in Some Leadership Theories, Some Question Their Efficacy

The suggestion of utilizing providing antecedents as a nontraditional predictor (suppressor) is controversial not only because of being a suppressor variable, but also because the Operant Model's stance on antecedents is contrary to management practices that many people in organizations use and to the hypotheses of several leadership theories.

Directives are very common and sometimes celebrated behaviors in organizations. Daniels (1994) observed that businesses heavily invest in "antecedent activities" such as memos, training, policies, mission statements, slogans, posters, and buttons and reported that in an informal survey, he discovered that managers spent about 85% of their time "either telling people what to do, figuring out what to tell them to do, or deciding what to do because employees didn't do what they told them to do" (p. 18). Data have shown that providing antecedents is a fairly common part of normal, everyday supervisory behavior (e.g., Komaki, et al., 1989). For example, managers across several settings spent on average 5.8% of the time they were observed giving instructions or clarifying performance (Brewer et al., 1994; Komaki, 1986; Komaki, et al., 1986; Komaki et al., 1989). For most samples, the time spent giving antecedents was greater than the time spent monitoring and providing consequences. In addition, providing antecedents is frequently looked upon favorably. A former IBM executive, for example, was heralded as "a great leader, the ultimate motivator." He was described as being skilled at explaining what employees needed to do tackle

problems and having such an impact that "people not only understood what they needed to do, but were ready to charge a machine-gun nest to do it" (Rifkin, 1994, p. 8).

Several leadership theories also posit that concepts and behaviors similar to antecedents are effective. For example, initiating structure behaviors described in the Ohio State Studies would likely be categorized as delivering performance antecedents according to the Operant Model. As discussed earlier, a leader high in initiating structure "organizes and defines the relationships in the group, tends to establish well defined patterns and channels of communication and ways of getting the job done (e.g., he or she assigns people to particular tasks, emphasizes deadlines, etc." [Fleishman, 1973, p. 8]). Leadership behaviors, such as insisting on maintaining standards and meeting deadlines and deciding in detail what will be done and how it should be done, are indications of initiating structure (Bass, 1990), as well as performance antecedents. Charismatic/ transformational leadership theory also recommends "antecedent-like" behaviors to arouse followers emotionally and inspire them to extra effort and greater accomplishment, with followers in turn becoming leaders themselves. The literature describes inspirational speeches, communications and reminders of vision, exhortations, and confidently worded expectations of subordinate performance, all of which are providing antecedents (Bass, 1985; Burns, 1984; House, 1977; Tichy & Ulrich, 1984).

Some, however, have questioned the appropriateness of the transformational/charismatic approach in all situations (e.g., Bass, 1985; House, Spangler, & Woycke, 1991, Inkson & Moss, 1993). House et al. (1991) suggested that charismatic leadership might be "dysfunctional" in situations "requiring routine but reliable performance in the pursuit of pragmatic goals" (p. 391). In a similar vein, Inkson and Moss (1993) raised several questions about transformational leadership, including whether it is necessary when tasks are simple and everyone understands and achieves them. In addition, Bass (1985), a major proponent of the theory, admitted that "there are situations in which the transformational approach may not be appropriate" (p.40). These authors, thus, have

questioned whether antecedent-like behaviors -- inspirational speeches, communications of vision, and reminders about performance – are universally useful.

Operant conditioners have questioned the efficacy of antecedents for motivating performance over time, especially if not accompanied by consequences. Antecedents occur before behavior and are thought to signal or cue the behavior. Alone, or without subsequent reinforcement of the behavior, antecedents will not sustain or increase the occurrence of the behavior (Scott & Podsakoff, 1985; Skinner, 1974). The "peril" of relying on antecedents in the world of organizations was simply put by a management consultant with an operant conditioning background:

> In most organizations we attempt to manage performance by telling people what to do.… We send memos, have meetings, write policies, hold classes, and make informational and inspirational speeches. Interestingly, when these methods don't get the desired response or level of performance we want, we tell the same people again, usually in the same ways.… We send new memos around (with bolder type, capital letters, and even exclamation marks) about old memos that were ignored, or have meetings about why meetings don't seem to be productive. If we train people about the importance of doing it "right the first time, every time," and they don't, we bring them in and we train them again. If we inspire them to reach for the heavens and they barely raise their hands, we make a more impassioned plea (Daniels, 1994, pp. 17-18).

In sum, antecedents are common and often well-regarded behaviors in organizations when attempting to motivate employees. Several leadership theories also specify hypotheses in which antecedent-like behaviors are effective. However, some have questioned whether these behaviors are really a good leadership or managerial behavior in all situations.

Research Evidence Supports the Secondary Role of Providing Antecedents

Besides theoretical rationale, the results from research support the secondary role of providing antecedents in the Operant Model. Three pieces of empirical evidence exist: one study showed that delivering antecedents may have functioned as a suppressor variable (Jensen & Komaki, 1993), no relationships were found between providing antecedents and effectiveness in field tests (Brewer, et al., 1994; Jensen & Komaki, 1993; Komaki, 1986; Komaki, et al.; 1989; Komaki, et al., 1991; Komaki, et al., 1997), and the timing of the delivery of antecedents made a significant difference in effectiveness in one study (Reynard & Komaki, 1995).

Suppressing Effect Suggested

In the study by Jensen and Komaki (1993), described in Table 1, antecedents fit the definition of a classic suppressor variable. Providing antecedents was not related to effectiveness ($r = .01$), was highly related to delivering consequences ($r = .69$, $p < .05$), and received a negative beta weight in the multiple regression equation ($\beta = -.47$, $p < .05$). Though unrelated to effectiveness, providing antecedents may have served to remove irrelevant variance and improve the prediction of effectiveness.

Consistently Lacking Relationship with Effectiveness

Little evidence exists to show that providing antecedents has a direct role in effectiveness, as do monitoring and providing consequences. In six studies in which monitoring or providing consequences were significantly related to effectiveness, giving antecedents was *not* (Brewer, et al., 1994; Jensen & Komaki, 1993; Komaki, 1986; Komaki, et al.; 1989; Komaki, et al., 1991; Komaki, et al., 1997). For example, monitoring distinguished between effective (M = 2.9%) and marginal (M = 2.0%) groups of insurance managers ($t = 2.59$, $p < .01$) (Komaki, 1986) and the time skippers spent providing consequences ($r = -.47$, $p < .05$) and monitoring ($r = -.51$, $p < .05$) were related to round robin regatta standings (Komaki et al., 1989). The frequency of providing antecedents (measured as the percent of time spent) did not distinguish between groups of managers categorized

as effective and marginally effective in two studies (Jensen & Komaki; 1993; Komaki, 1986). In Jensen and Komaki (1993), the correlation between providing antecedents (for all subjects in the sample) and a judgmental measure of newspaper managers' ability to motivate others to work was .01, ns. In addition, a study conducted in a sailboat racing setting found that although leaders spent an average of 40.1% (*SD* = 9.8) of their time delivering antecedents, the correlation between antecedents and a measure of crew coordination was -.03, ns (Komaki et al., 1997).

In the above studies, the quantity or frequency of providing antecedents was discussed. But the category of antecedents has also been divided into qualitative subcategories according to: (a) delivery: those that are delivered directly versus indirectly; (b) responsibility: whether it is obvious or not obvious who is responsible for doing the directive; (c) planning: whether the directives are given in anticipation of the future or to deal with the current situation; and (d) teammanship: whether coordination is or is not a part of the directive. Neither the percent of time giving antecedents overall, nor the percent of time for the qualities of antecedents in regard to their delivery, clarity, and planning have distinguished between effective and marginally effective newspaper managers (Jensen & Komaki, 1993). Both groups essentially spent the same amount of time giving antecedents that were: delivered directly to the person for whom they were intended (*M* = 4.4%, 4.0% for effective and marginally effective managers, respectively); delivered indirectly for someone else (*M* = .9%, .7%); clear in regard to who was responsible to perform them (*M* = 4.7%, 4.3%); unclear in responsibility (*M* = .5%, .4%); pertinent to current objectives or problems (*M* = 5.2%, 4.6%); and intended to prevent an anticipated problem (*M* = .1%, .1%) (Jensen & Komaki, 1993). In fact, in none of the studies have the qualities of providing antecedents been found to be predictive of effective leaders (e.g., Jensen & Komaki, 1993; Komaki, 1986; Komaki et al., 1997).

As can be seen from the results presented here, neither the quantity -- the amount of time spent giving antecedents -- nor the quality – the time spent giving different types of antecedents -- seem to be important in motivating employees.

Relationship Shown Between Timing of Antecedents and Effectiveness

Another study provides the third piece of evidence for the secondary role of delivering antecedents in the Operant Model. Curious about the role of this behavior category, Reynard and Komaki (1995) thought that perhaps the placement of performance antecedents in relation to other behaviors might be important. In other words, they speculated that it may not be the frequency or how often antecedents are given, but *when* they are provided that has an effect. They hypothesized that marginally effective managers would give more antecedents before monitoring performance. To test the hypothesis, they examined the sequencing or temporal order of delivering antecedents, monitoring, and providing consequences in teams of leaders aboard sailboats (Reynard & Komaki, 1995). Leaders' behaviors were measured through observation of videotapes taken aboard sailboats and coded in five-second intervals. The focus of the study was a particular sequence of behaviors: an antecedent, followed by a "monitor," followed by a consequence. Sequences were identified by progressing through the transcripts and video until the interval in which the first antecedent was given (marking the time), and then continuing until monitoring occurred (noting all other behaviors occurring and time passing in between), and then continuing until a consequence was given (noting all other behaviors occurring and time passing in between). This process was repeated until the race ended. A measure of the time it took the leaders' crews to hoist a sail was used to assess how effectively the crews aboard the boats worked together.

Reynard and Komaki (1995) found support for the hypothesis that marginally effective leaders would give more antecedents before monitoring within the sequences they deliver. Those leaders who gave many antecedents before monitoring, were not as effective as those who gave antecedents less. Results showed that the greater the number of antecedents the leader delivered *before* monitoring, the worse the crew performed – it took the crew longer to raise a sail known as the chute ($r = .44$, $p < .01$). In other words, giving more commands or directives *before* finding out how people were doing on their tasks was associated with being *less* effective as a leader.

To illustrate these differences, Table 5 contains transcripts from the study. The left column of the table shows conversation of an ineffective leadership team (comprised of a skipper, tactician, and crew boss) and the right, an effective team of leaders. The transcripts are from a few minutes during a race during the period in which the boat rounds a mark in the race course. As can be seen in Table 5, the marginal leader team spent many intervals giving antecedents before monitoring. The first antecedent occurred at 5:19.00 and these leaders provided antecedents to the crew for 15 intervals, before monitoring the sail preparation at 5:23.30.

In contrast, the effective leaders' sequences were much shorter. The first one begins at 6:02.40 and ends within 10 seconds. By 6:02.45, these leaders gave one antecedent and monitored immediately. Consistently throughout the effective leader's transcript, few antecedents were given before monitoring.

Table 5

Verbatim Transcripts from Marginally Effective and Effective Leaders Aboard Sailboats

Marginally Effective Leader			**Effective Leader**		
Time	Statement/Action	Cat.	Time	Statement/Action	Cat.
5:19.00	(Steering boat) Guys, you gotta build up speed.	**A**	6:02.40	Pull forward a _(?)_, pull forward, yeah.	**A**
	Keep it fat until we get speed.	A		(Looking up at sail)	**M**
.05	Break 'em down a little bit and we'll go over faster.	A	.45	Now bring it back a little bit.	**C**
.10	…(?)… cross this …(?)…	CD[1]	.50	We're supposed to jibing again right here.	
.25	(Looking out over water)		.55	Square back the pole.	**A**
				Bring the pole back.	A
				Back, back, back, back, back.	C
				Square back the pole. Square back the pole.	C
				Who's on the foreguy?	**M**
5:19.30	(Looking out over water)		6:03.00	Square back the pole more.	**C**
				That's why we're having trouble trimming chute.	
.35	(Looking out over water)		.05	Trim that chute! You gotta come down.	**A**
				Coming down. Go ahead centerline.	A
			.10	Whew.	

Table 5, cont.

Marginally Effective Leader			Effective Leader		
Time	Statement/Action	Cat.	Time	Statement/Action	Cat.
			6:03.15	Trip.	A
			.25	Looks great, guys.	C
				I know I don't.	
			.30	Wait, wait, wait.	C
5:20.00	(Looking around over the water)		.45	(Looking at sail)	**M**
.05	(Looking around over the water)				
	We're _(?)_ that red mark.				
.10	(Looking around over the water)				
	…(?)… up.	CD[1]			
.15	(Looking around over the water)		6:04.05	…(?)… Okay, go ahead and jibe.	A
.20	Allright. …(?)… what I was thinking …(?)… jibe.	CD[1]			
.25	You gotta go high by Privateer so you don't get their gas.		.15	…(?)… Coming down.	A
.30	As long as you're coming in late, you might as well come in late fast. Ah shit we are moving.	C	.30	Easy. What's going on?	M
				Come on, guys. Come on. Hard! Now. Hard!! Easy.	C
			.35	Watch the boom. Watch your head. Watch your head.	A
				Somebody push out on that boom. Hard now, hard_(?)_ like.	C
.40	Ready about.	A			
.45	Helm's over! (whistles) All right guys, come on let's move.	C	.45	Hey Jeff, trim that chute, trim that chute.	A
.50	Come on. Come on. Come on. Let's go	C			
.55	Let's go Ernie. Let's go Dave.	C			
5:21.00	Skirt. (Whistles)	A			
.05	(Looking out over the water)				
.10	(Looking out over the water)				
.15	(Looking out over the water)				
.20	…(?)… you guys …(?)	CD[1]			
.25	Everybody's going forward, forward.				
.30	(Looking out over water)				
.35	(Looking out over water)				
.45	The mark's up high, …(?)…				
5:22.00	Allright, we'll do a jibe set then.				
.05	Allright. Everybody understand that? When we come around this mark, we're going to jibe.	A			

Table 5, cont.

Marginally Effective Leader			Effective Leader		
Time	Statement/Action	Cat.	Time	Statement/Action	Cat.
5:22.10	(Looking around over the water, steering)				
	Dave, we're going to jibe set.	A			
.15	(Looking around over the water, steering)				
.20	We're gonna tack, then we're gonna jibe				
.25	Hold this, Rick. Rick, hold this. Come right in behind Skyway.	C			
.30	(Looking around over water)				
.40	Tack.	A			
	Ready about. Turn the boat. Tacking.	A			
5:23.00	Allright, prepare to jibe.	A			
.05	Down.	A			
.10	Jibing.	A			
	Jibing.	A			
	Pat, put a little main on.	A			
.15	Haul up some chute.	A			
	Come on Andy. Come on.	A			
.20	Come on Eric, let's go. Good, good, good. Ease the main. Going down.	C			
.30	(Watching crew get sail ready)	M			
	Come on, guys.	C2			
.45	Come on guys.	C			
	Come on Eric. Come on Dave.	C			
.50	Easy.	C			
.55	Watch the sail! Watch the _(?)_ sail. Hey the guy's not in the jaws!	C			

Note. Adapted from "The Music of Management: Rhythms of Effective Leaders," by M. Reynard & J.L. Komaki, 1995, Poster presented at the conference of the Society for Industrial and Organizational Psychology, Orlando.

[1] CD= Can't Decipher, the whole statement made by the leader could not be heard, and categorization could not be accomplished.

[2] The picture and the leader's tone of voice indicated that the leader was not pleased with the progress of the sail preparation. He was thus indicating his evaluation of the crew's performance.

This study indicates that the timing or placement of antecedents in relation to other behaviors may have an impact on performance. The potentially negative impact of delivering antecedents is an important finding because managers frequently use these behaviors.

In sum, based on the occurrence of antecedents and the patterns of relationships between monitoring, consequences, antecedents, and effectiveness that have been found in the research on the Operant Model, the role of antecedents in the model seems to be fitting.

Providing Antecedents Will be Utilized as a Predictor in the Present Study

In all the research on the Operant Model, providing antecedents has been measured via observation. For the first time, in the present study, information about providing antecedents will be collected in the in-basket and utilized as a predictor, albeit a nontraditional one. It is expected that if information is collected about monitoring, providing consequences, *and* providing antecedents, better predictions of supervisory effectiveness can be made. Thus, **Hypothesis 3** is that **directives will improve the prediction of motivational effectiveness over and above monitoring and providing consequences**. In other words, more variability in effectiveness should be predicted by providing antecedents, monitoring, and delivering consequences than by monitoring and providing consequences alone.

One possibility is that providing antecedents would function as a suppressor variable, similar to the Jensen & Komaki (1993) study. In this situation, directives themselves would *not* be related to effectiveness (which is consistent with other previous field research on the model), would be significantly related to another predictor (likely providing consequences), and have a negative beta weight in a multiple regression with monitoring and providing consequences. However, the results from the Reynard and Komaki (1995) study showing that lackluster leaders provided more antecedents before monitoring performance indicate that directives could have a *negative* relationship with effectiveness, particularly without monitoring first. In this case, directives would be a valid predictor, but have a negative sign (and a negative beta weight in the multiple regression).

(Traditional predictors have a positive relationship with effectiveness.) Regardless of the relationship between providing antecedents and effectiveness, the behavior could still improve the prediction of effectiveness with monitoring and providing consequences. The in-basket actually affords an opportunity to investigate both possibilities because the context in which respondents deliver antecedents is known.

In sum, a third behavior identified in the Operant Model of Effective Supervision, providing antecedents, may be useful in the prediction of effectiveness in conjunction with two other major behaviors in the theory, monitoring and providing consequences. Based on operant conditioning theory, providing antecedents has been relegated to a secondary role in the model. Though contrary to recommendations in several leadership theories and despite their popularity, research results in regard to providing antecedents seem to support their place within the model. The category of antecedents has not been found to be related to effectiveness in any setting, and in fact, acted as a suppressor variable in predicting effectiveness of a group of newspaper managers. In addition, one study has suggested that marginal managers provide many antecedents before monitoring performance. Thus, providing antecedents is expected to aid prediction in a nontraditional way.

Chapter 6

PROBLEM AND PROPOSED HYPOTHESES

The Operant Model seems well-suited to address the lack of non-cognitive predictors for selecting managers because the behaviors are theoretically based, clearly defined, interpersonal in nature, and describe how supervisors can motivate their employees. There is also some empirical evidence of their validity. Little work has been done, however, with measures of the behaviors that practitioners could use in real selection situations. The present study is the first to attempt to link the behaviors measured via an in-basket to supervisory effectiveness.

Based on previous research on the model involving monitoring and providing consequences, the two major behaviors hypothesized to be related to effectiveness, the following is the first set of hypotheses:

Hypothesis 1a: Monitoring and providing consequences will each be positively and significantly related to supervisory (motivational) effectiveness.

Hypothesis 1b: Monitoring and providing consequences will each be unrelated to technical expertise.

The second set of hypotheses of the study concerns combining the predictors of monitoring and providing consequences. Additive combinations are common in selection and imply that scores on the predictors can compensate for one another, while multiplicative combinations, in contrast, are not very common and imply that the predictors interact. There is some theoretical rationale for the multiplicative combination, in that one proposition of the Operant Model (which has not received much attention) is that monitoring and consequences interact to produce an effect. Both additive and multiplicative combinations will be tested, and the following hypotheses are offered:

Hypothesis 2a: An additive, weighted linear combination of monitoring and providing consequences will be positively and significantly related to supervisory effectiveness.

Hypothesis 2b: The interaction of monitoring and providing consequences will add significantly to the prediction of supervisory effectiveness, over and above monitoring and providing consequences alone.

The last hypothesis is related to including a third behavior identified in the Operant Model as a predictor in addition to monitoring and providing consequences. Providing antecedents or directives has a secondary role in the model, and has never been found to be positively related to effectiveness. It is expected to aid in prediction in a nontraditional way, possibly as a suppressor variable.

Hypothesis 3: Directives will improve the prediction of motivational effectiveness over and above monitoring and providing consequences.

Chapter 7

METHOD

Subjects and Setting

The research took place with 35 bankers in a group within a major, multinational investment bank. The group that participated was considered a product group, responsible for structuring various "deals" -- in essence providing advisory services to corporate customers for a fee. Deals were usually large, very complex, and custom-designed transactions to help customers meet tax or accounting objectives, ranged in value from one hundred million to one billion dollars, and took generally six to twelve months to complete from the time of hire. An example of a typical deal involved arranging the one hundred million dollar transaction for a Los Angeles company that was financing two ships being built in Germany (and "flagged" in Liberia) to be owned by United Kingdom subsidiaries.

The environment of the group was very fast-paced and high-pressure. Employees were expected to work long hours and be available nearly 24 hours a day, seven days a week. Travel was frequent. Also, deadlines were extremely important for deals to be viable. For example, the customer buying ships decided to do the deal in September and needed it to be closed by the end of the year in order to get tax and accounting benefits. Had the work schedule not been accelerated to three months (which was half of what was normal), the bank would not have had a customer, nor made any of the two million dollars in revenue on the deal.

The work of closing deals involved employees creating or participating in "deal teams," ad hoc groups composed of the people from both within and outside the group needed to bring particular deals to fruition. A deal team could include a customer representative from within the bank (but not the group), outside investors, legal staff, and other people from the group who create pricing structures, for example. The lead person on the deal drew on people with the expertise

he/she felt she needed to move the deal to closure, and might "invite" members of the group to participate. (Invited group members could get involved, or not, as they wished.) Closing deals required the employees to negotiate with, coordinate the tasks of, and exchange information with many people, often in different countries or across several time zones.

Some members of the group directly and formally supervised the work of others, and often worked on their own deals. Others who were essentially leaders of deal teams, but did not formally supervise were known as transactors. These transactors acted as leaders in that they coordinated the work of others and made sure that tasks were being completed so that their deals would be closed.

The group's Human Resources representatives and leaders compiled a list of 52 potential participants in the research. This list was created so that only those who were integral in creating deals and who had knowledge of others' work were included in the study. This potential participant list was comprised of people from three geographic regions--the United States, Europe, and Asia/Pacific--and from various levels, spanning from Associate to Managing Director.

From the potential list of 52, 35 (67.3%) volunteered to participate and submitted the required information. Of this sample, 45.7% was located in Europe, 40.0% in the United States, and 14.3% in Asia/Pacific. In ascending organizational level, 11.4% of the sample held the job title of Associate, 2.9% Assistant Vice President, 40.0% Vice President, 25.7% Director, and 20.0% Managing Director. Refer to Table 6 for the breakdown of job title by region for the participants in the study.

Table 6

Frequency of Job Title by Region for Study Participants

Region	Associate	Asst. Vice President	Vice President	Director	Managing Director	Total
Asia/Pacific	0	1	2	0	2	5
US	1	0	7	4	2	14
Europe	3	0	5	5	3	16
Total	4	1	14	9	7	35

Other demographic information collected included gender, undergraduate and graduate degrees, age, years of education, years in the organization, years in the current position, years of supervisory experience (including prior jobs), and days of supervisory training (including those in other organizations). The majority of the sample was male (82.9%, $n = 35$). Most participants had both undergraduate (97.0%, $n = 33$) and graduate degrees (80.6%, $n = 31$). On average, participants were 38.6 years of age, had 17.9 years of education, 6 years at the organization, 3.2 years in their present position, 7.3 years of supervisory experience, and 8.9 days of supervisory training. See Table 7 for more detail on the demographics of the sample.

Table 7

Summary of Demographic Information

Demographic variable	*N*	Minimum	Maximum	*M*	*SD*
Age	34	25	57	38.6	7.9
Education	35	12	22	17.9	1.7
Tenure in organization	35	1	33	6.0	6.0
Tenure in position	35	1	14	3.2	2.5
Supervisory experience	34	0	29	7.3	7.2
Supervisory training[a]	34	0	40	8.9	12.6

Note. Except where indicated, the unit of measurement is years.

[a] Unit of measurement is days.

Measures

Predictors: Supervisory Behaviors via Revised Operant Supervisory In-Basket Assessment (OSIBA)

To assess supervisory behavior, a revised version of the Operant Supervisory In-Basket Assessment (OSIBA) was used. As discussed earlier, the OSIBA is a type of simulation exercise in which respondents take on a supervisory role while responding to items they find in their "in-baskets." During the exercise, respondents are asked to take on the role of a publisher of a monthly food magazine, with a staff of managers from several departments, including production, human resources, advertising, marketing/circulation, art, and accounting. Participants must read through

the items in the form of memos, written phone messages, and e-mail messages and write responses to “characters” depicted in the scenarios.

The measure was developed as an alternative to the observational OSTI, because of the great amount of time spent collecting data in behavioral observation (Komaki, 1998). Due to time considerations, the OSIBA was developed to be administered in a one- to one-and-one-half-hour period. The in-basket measure was also seen as advantageous over observation in that the in-basket provides the same context for all respondents, allowing for better comparisons. In other words, each participant’s behaviors are being captured while in the same circumstances. Direct observation of participants carried out in the natural work environments, on the other hand, could be very different for each participant depending upon the number of subordinates or the amount of activity occurring during observation times, for example. In addition, the in-basket measure allows for the assessment of some subcategories of behavior that it would not be possible to assess via observation in the field. For instance, in the in-basket, it can be determined whether feedback (consequences) is contingent on performance. Without knowledge of the full context of the situation, or of the performance upon which the consequence was based, an observer in the work environment would not be able to tell if feedback given was deserved or not. Thus, the OSIBA was viewed as a viable alternative to observation, with benefits for use in selection situations. (One issue with the measurement of supervisory behaviors with the OSIBA in comparison to live observation, however, is that there is a need to assign values when behaviors are demonstrated in the in-basket. In observation, the scale is a simpler, straightforward percentage of time.)

The OSIBA is currently administered via paper (consideration has been given, however, to developing a computer-based version). An introductory page describes the situation and the role the respondent will be playing, and gives tips for responding to the items. Respondents are instructed to apportion their time to deal with all items, respond as they normally would, and not to be formal. After reading through the instructions (and memos if they wish), respondents then write their replies

for each of the 22 items in the exercise in spaces provided. (Refer to Appendix A for a copy of the instrument, which, for this sample, was named the "Motivational Effectiveness Exercise.") There are five versions of the in-basket with the same 22 items. Each version begins with the same first item, which contains a list of people's names and departments that report directly to the publisher. This item is followed by the remaining 21 items in different random orders in each of the five versions.

Categories

As discussed previously, the OSIBA was scored for the categories of monitoring, providing consequences, and providing antecedents. Refer to Table 8 for definitions and examples of these behaviors and their subcategories, as assessed in the in-basket.

Monitoring. Monitoring was defined as in the Operant Model (Komaki, 1998), as gathering information about performance. Two methods of monitoring also identified in the model were measured in the in-basket. Hypothesized in the model to be the most effective form of monitoring, *work sampling* is defined as observing workers in action and/or examining the products of their work. Because it is not possible for in-basket respondents to actually work sample as they could during live observation, the definition was modified to requesting to observe workers in action and/or examine the products of their work. Monitoring via *self-report* was defined as asking for performance information.

Providing consequences. Providing consequences was simply defined as communicating knowledge of performance. The major subcategory of this behavior had to do with the evaluation expressed in the communication – either *positive/neutral* (referred to as positive consequences throughout the remainder of the paper) or *negative*. Similar to previous research on the Model, providing positive consequences was defined as recognizing effort, excellence, or improvement, or expressing neither approval nor disapproval of performance and negative consequences was letting people know when they had erred or needed to improve.

Positive consequences were further broken down in terms of their deservedness and directness. Positive consequences that were *not warranted*, or non-contingent, were defined as occurring when respondents gave positive evaluations when they were not merited. This is illustrated in Table 8 -- the respondent wrote that the artwork sent to him/her by a graphic designer looked "terrific," when it actually had a spelling error in it. Positive consequences that were *warranted*, on the other hand, were defined as being merited. For example, a respondent wrote, "Understand from Bruce that you found a great place for our awards banquet at 80% of what he had estimated as the probable cost. Great job!" to a person whose work was mentioned in a report about the preparations for the banquet. The communication of the performance evaluation was also categorized according to its directness, which could be either direct or indirect. *Direct* consequences were defined as those delivered directly to the person for whom the consequence was intended and *indirect* ones were those delivered through someone else.

In addition, six different special types of positive consequences were specified: (1) giving *simple*, short acknowledgements, (2) *acknowledging the bearer of bad news*, (3) *informing someone out of the contact loop*, (4) *broadcasting to multiple parties*, (5) *letting* the memo *sender know of action*, and (6) *relaying a problem non-judgmentally*. Acknowledging the bearer of bad news, for example, was defined as recognizing a person for drawing attention to unwanted or undesirable information. Refer to Table 8 for definitions of each and examples.

Another aspect of positive consequences, the form in which they were given, was coded. If the recognition of effort/excellence (regardless of whether it was warranted/not, direct/indirect, or one of the special cases) was in the form of a directive or a directive was given in association with it, it was coded as having an *antecedent-like* form. For example, the comment, "Keep up the excellent work." was considered antecedent-like because the positive consequence that the recipient did excellent work was presented in the form of a directive.

Table 8

Definitions and Examples of Major Categories and Subcategories of Behavior Measured in the Revised OSIBA

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
Monitoring (M)		**Gathering information about performance**	(Responding to note that person needs help and is having problems with budget projections) *What stage is your budget projection in? How could we help you with it?*
Work Sample		*Requesting to observe workers in action and/or examine the products of their work*	(After receiving error-laden e-mail about forwarding material) *Please show me the final BEFORE it goes to print.*
Self-Report		*Asking for performance information*	(After getting information that a new form is difficult to complete) *Have you discussed the form with Accounting? Is this an isolated problem or is it widespread?* Thank you for sorting out the payment.
Providing Consequences (C)		**Communicating knowledge of performance**	(To graphic designer about artwork with a misspelling in it) *Looks good, but make sure the typo is corrected (Foodstuff).*
Positive or Neutral		*Recognizing effort, excellence, or improvement; or expressing neutral, neither approval nor disapproval of performance*	(To Production Dir. After receiving report of production costs over last five months) *Very nice improvement from August to December – keep up the good work*!
	Deservedness		
	When not warranted	When not merited	(To graphic designer about artwork with a misspelling in it) *Thank you for the prototype artwork for the new column. It looks terrific*! Please send along to the Production Dept.
	When warranted	Only when merited	(To person mentioned in status report on banquet preparations) *Understand from Bruce that you found a great place for our awards banquet at 80% of what we had estimated as the probable cost. Great job*!

[a] The portion of the verbatim example that is in italics is an illustration of the category or subcategory.

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C	When warranted (cont.)		
Positive or Neutral (cont.)			
	Directness		
	Direct	Delivered directly to person for whom the consequence was intended	(To Anne after receiving a compliment about her) *Heard good feedback from the auditor on your job performance. GREAT JOB!*
	Indirect	Delivered through someone else	(About Anne after receiving a compliment about her) *Please congratulate Anne on our behalf. Auditors were impressed with her. (P.S. Make sure they don't steal her away with a job offer!)*
	Special types		
	Simple	Shorter, often one or two words acknowledging routine actions	(Upon receiving report of advertising accounts gained or lost) *Thanks*. How does this compare with competition?
	Acknowledging bearer of bad news	Acknowledging person for drawing attention to unwanted or undesirable information	(To person concerned with quality) *Thanks for the heads-up. You are right that the quality of those images is critical in achieving our vision of being the top gourmet industry mouthpiece in the country*. I will talk with Bobby.
	Informing someone out of contact loop	Acknowledging a person who was mentioned in a memo but was not an obvious candidate for a comment	(To outside photographer, after getting information that she had difficulty completing a new reimbursement form) *I heard from J. that we had a number of problems sorting out your fee for the shoot. Apologies for that, and I trust it is now sorted. If not, let me know and I will see to it.*

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C	When warranted (cont.)		
Positive or Neutral (cont.)	Special types (cont.)		
	Broadcasting to multiple parties	Publicizing the recognition of effort/excellence to other people or groups	(To all direct reports upon receiving report of advertising accounts gained or lost) *[Forward results] Great work by the Ad team, but also a testament to all of the company in improving the quality of our magazine. Thanks to you all.*
	Letting sender know of action	Informing sender of intended response or course of action and/or of receipt of message	(To HR Director when notified of Marketing group missing a training program) *I'll talk to Marketing*. or *Noted*.
	Relaying problem non-judgmentally	(In case of a problem or potential problem) Notifying appropriate person about the facts, without blaming or making judgments on fault	(To Accounting after receiving information that contracted photographer had difficulty completing a new reimbursement form) *I understand that some people are having difficulty filling in the new form*. Could you investigate and change it if necessary. Please talk to some of the people who use it regularly to get their feedback.
	Form		
	Antecedent-like or accompanied by antecedent	In the form of a directive or adding a directive	(To Anne after receiving a compliment about her from someone else) Good job Anne. *Keep up the excellent work.*
Negative		*Letting persons know when they have erred or need to improve*	(To graphic designer about artwork with a misspelling in it) *Would prefer another F at the end of Foodstuf.*

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C			
Negative (cont.)	Deservedness		
	When warranted	Based on first-hand knowledge	(To Advertising about tabulation mistake) *I think there is a calculation error on the accounts at the end of 2000, it should be 83, not 53.* In any case, good progress.
	When not warranted	Not related to performance	(To entire staff when only two had not submitted work) As you are aware, we need to submit the departmental budget projection by tomorrow. *To achieve this, I need everyone's input for my comments and consolidation. As some of you may have difficulties in the last few days, I would appreciate if you can finish it by 4 pm today and send to me as early as possible.* If you have any questions please do not hesitate to give me a call.
	When questionably warranted	Based on report of performance by secondary source	(To Marketing Director when notified by HR that Marketing group missing a training program) I understand that the trainees were late arriving to a session, requiring a re-schedule. *These are important; otherwise we would not arrange them. Please ensure prompt attendance next time.*
	Target/Reason		
	About targeted person without disconfirming information	Assigning blame to targeted person (or group), based only on a report by a secondary source, and without disconfirming information	(To Accounting after receiving information that contracted photographer had difficulty completing a new reimbursement form) *M., we need to improve the form to make it more user friendly. Give it a try and work with J.R. to get his input.*

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C	When questionably warranted		
Negative (cont.)	<u>Target/Reason</u> (cont.)		
	About targeted person with mixed or conflicting information	Assigning blame to targeted person (or group) based on only a report by a secondary source, and with mixed or conflicting information	(To person concerned with quality of images and production workers' lack of attention to quality) I'll speak with them. (To Production Manager who believes problem lies with equipment) *We need to improve the quality on Press #1. Please make sure Flay's images are top notch.*
	About supervisor (or other person)	Assigning blame, not to targeted person (or group), but to supervisor (or other person), without any indication that supervisor was responsible	(To supervisor of graphic designer who sent artwork with a misspelling in it) *Can you please make sure you review work before it gets sent to me? I don't expect to be correcting typos from your graphic designers*.
	<u>Severity</u>		
	Diplomatic	Delivered in a way that is not sarcastic, abrupt, or demeaning	(To Advertising Director after receiving customer complaint regarding phone answering) *Let's have a meeting to discuss the issues we are having with some of the large accounts. I want us to discuss how we can be more responsive and deal with accounts more efficiently.*
	Harsh	Delivered in a way that is non-denigrating, non-punitive, possibly sarcastic or abrupt	(To graphic designer about artwork with a misspelling in it) *Perhaps you should check the spelling of Foodstuff. Really this kind of schoolboy error is pretty poor. Please correct before sending to Production*. But I like the rest of the design.

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C			
Negative (cont.)	Severity (cont.)		
	Injudicious	Delivered in a way that is denigrating or overly punitive	(To editorial assistant, after receiving mistake-ridden e-mail about forwarding material) *Please take more time reviewing issue. If it's anything like this email, you're fired.*
	Frequency		
	Once	Stated one time to one recipient	(To Accounting after receiving information that contracted photographer had difficulty completing a new reimbursement form) *There seems to be a problem with filling out the new forms.* [1] Do you think we could simplify it?
	Twice	Stated twice or to two recipients	(To Accounting after receiving information that contracted photographer had difficulty completing a new reimbursement form) *Looks like someone needs to review this form!* [1] *You know my pet project: Q.U.A.L.I.T.Y.* [2]
	Three or more times	Stated three or more times or to three or more recipients, or some combination of these	(To Accounting after receiving information that contracted photographer had difficulty completing a new reimbursement form) *We need to improve the form to make it more user friendly.* [1] *Give it a try* [2] *and work with J. Rosso to get his input.* [3]

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
C			
Negative (cont.)	<u>Form</u>		
	Classic	Conveying disapproval or doubt about another's work or noting work that has not been completed	(To HR Director when notified of Marketing group missing a training program) *Text and map directions are inconsistent. You're instructing people to go West on Concord, the map shows heading East. I'm not surprised people were late if they followed your written directions*.
	Antecedent-like or accompanied by antecedent	Same as above but stated as a directive, instruction, or reminder	(To editorial assistant, after receiving mistake-ridden e-mail about forwarding material) *Please ensure quality! This is very important and I do not want any mistakes. Re-doing increases the cost! Please remember the "Do it right the first time" concept.* Thanks.
	Pseudo-monitor	Same as above but stated as rhetorical question	(To graphic designer about artwork with a misspelling in it) *Have you checked this for accuracy? (Or have we changed the spelling of our company's name and I just didn't know about it?)* I expect you to be more careful with your work. Quality is the key to our success. Please show me the corrected version before it goes out.
Providing Antecedents (A)		**Conveying expectations of performance (e.g., delivering directives, instructions, or reminders) in order to motivate others**	(To Advertising Dept. upon receiving description of complaint from customer) *This is a regular reminder for all staff to refer to the service standard memo, last circulated and revised in Dec. 2001. All dept. heads should remind their subordinates to read this memo again carefully and follow in their working environment*. Thanks.

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
A			
Tacking on traditional A		Giving directives, reminders, instructions when neither warranted nor preferred (e.g., no directives were needed or requested)	(To person providing a status report on banquet preparations) Thanks for the update, I have the following comments: 1. *We need to focus on a keynote speaker*. The individual we select will need time to prepare for the banquet and at this point we are running out of time. *Touch base with Ethel* and get on my calendar for this week. I want to know who she has spoken with and who is on our short list. *Let's decide who we should invite* and I will put a call in to them. 2. *Talk with Nobu*, I believe he knows the person who we used to engrave the plaques from last year's event. 3. Nobu, excellent job on finding a facility. You too, Bruce, great job. *Let's keep up the momentum.*
	<u>Frequency</u>		
	Once	Delivered one time to one recipient	(To person describing complaint received from customer) Give me his details and I will call him myself later this morning. *We must get on top of this quickly*. [1] Thanks.
	Twice	Delivered twice or to two recipients	(Upon receiving cost report) The cost of last minute alterations has been getting rather high in recent months. *Please keep an eye on this.* [1] *I would like to see it much lower than it currently is.* [2]
	Three or more times	Delivered three or more times or to three or more recipients, or some combination of these	(To person providing a status report on banquet preparations) Bruce, great news on #3 – let Nobu know. *May I suggest helping Ethel* [1] *(let's not be too picky – time is starting to slip away* [2]*) and Gokhan (maybe an alternative engraver?*). [3]
Relying on A		Using (or recommending) directives, instructions, or reminders alone to promote performance	(To manager describing difficulty his group is having in completing a project done well before) *Perhaps a staff meeting – go back and re-examine the timeline, but be very clear to individual responsibilities and get each person's buy-in.*

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
A			
Relying on A (cont.)	Condition		
	When the situation is unclear and respondent does not monitor	When there is a question (about inconsistencies in behavior or reasons for difficulties) and respondent does not inquire or recommend inquiry	(To manager reporting that a good employee is often late to meetings in the afternoon) *It's time for you to reinforce with Dan that we expect him to be on time*. If he cannot provide a commitment, bring me in.
	Frequency		
	Once	Delivered one time to one recipient	(To employee whose manager reported that he is a good employee, but often late to meetings in the afternoon) I've received complaints re your tardiness to afternoon meetings. *Please be more attentive to timing as it will reflect on our decision to keep you as a manager*. [1]
	Twice	Delivered twice or to two recipients	(To employee whose manager reported that he is a good employee, but often late to meetings in the afternoon) Appreciate your hard work and commitment. I'm well aware of the hours put in in the morning, *but it's important to the team to start on time in pm.* [1] *Please be punctual*. [2]
	Three or more times	Delivered three or more times or to three or more recipients, or some combination of these	(To manager reporting that a good employee is often late to meetings in the afternoon) These creative types sometimes need more slack. … *Remind him of the value of other people's times.* [1] *Next meeting start without him.* [2] He will catch on. (To the employee) *Would like you to try to be on time for department meetings.* [3] We are too busy to keep 6-8 folks tied up. *If the time set does not work for you, perhaps you can get the time re-set*. [4]

Table 8, cont.

Category/ *Subcategory 1*	Subcategory 2/ Subcategory 3	Definition	Example[a]
A			
Relying on A (cont.)	Expanse		
	Sweeping	Delivered to multiple people or group (e.g., memo to all, training program)	(Upon receiving memo describing complaint from customer) *To Everyone in the Department: This is a regular reminder for all staff to refer to the Service Standard memo – last circulated and revised in December 2001. All dept. heads should remind their subordinates to read this memo again carefully and follow in their working environment*. Thanks.
	Condition (cont.)		
	When respondent asked to choose between antecedents and consequences	When given a choice, opts to provide antecedents, rather than feedback	(To person presenting visionary/ transformational leadership training or program for giving and receiving feedback to build managers' skills in motivating others) *I would vote for the former* [transformational]. *I think communication and shared vision is the way to go. We also need to encourage people to evolve and grow and not to get too comfy or stagnate.* Let's do it.

Note. Subcategory 1 is nested under category, subcategory 2 is nested under subcategory 1, and subcategory 3 is nested under subcategory 2.

Providing negative consequences was also broken down in terms of deservedness, similar to positive consequences. Giving a negative consequence that was *not* warranted was defined as letting persons know when they have erred or need to improve, but not related to performance, and was basically construed as giving negative feedback that was not merited. For example, a negative consequence that was *not warranted* was coded when a respondent sent a memo to his/her entire staff requesting that budget projections be submitted, when records indicated that only two people had not sent them in. In this scenario, those staff members who sent in their budget projections would have received a negative consequence that was not deserved and not based on their performance. In contrast, negative *when warranted* consequences were defined as being deserved (i.e., seen first hand). In certain in-basket items, the respondent was presented with a sample of work with a problem or error in it. Letting the person who wrote the memo know that they had erred or needed to improve was considered a negative consequence based on a sample of performance that was seen first hand, and thus, was considered warranted. In other items, the respondent was presented with reports of poor or substandard performance by others, and negative consequences in response to them were categorized as a third type of deservedness. Delivering negative *when questionably warranted* consequences was defined as letting persons know when they have erred or need to improve, based on a report of performance by a secondary source (not seen first hand). An illustration of this is that when notified by the Human Resources Director that a group of Marketing employees missed a training program, a respondent wrote the following to the Marketing Director: "I understand that the trainees were late arriving to a session, requiring a re-schedule. These are important; otherwise we would not arrange them. Please ensure prompt attendance next time."

Questionably warranted negative consequences were further subdivided into three types based on the target of or reason for the negative evaluation. The first type was *about the targeted person without disconfirming information*, which was defined as assigning blame to the person (or group) targeted in the report of a secondary source, and no information was present to disconfirm

the report. For example, in response to a memo from an employee complaining about a new reimbursement form and describing how a contracted photographer had difficulty completing it, a respondent wrote to Accounting that "… we need to improve the form to make it more user friendly. Give it a try and work with J.R. to get his input." This negative consequence was considered questionably warranted because there was no information to explain why the photographer had difficulty, nor any indication of other complaints. The second type was *about the targeted person with mixed or conflicting information*, which was defined as assigning blame to the person (or group) targeted in the report of a secondary source, and information is present that casts question or doubt on the report. An example of this behavior is a note written to an employee whose manager reported he was often late to afternoon meetings, but arrived early in the morning, "… I'm well aware of the hours put in in the morning, but it's important to the team to start on time in the pm. Please be punctual." The third type was *about the supervisor* (or other person), which was defined as assigning blame to the supervisor (or someone else) for a person's or group's poor performance, without any indication that the supervisor was responsible. A respondent, for example, did not provide a negative consequence to the graphic designer who sent him/her artwork with a misspelling in it, but laid the blame on his supervisor by writing the following to the supervisor: "Can you please make sure you review work before it gets sent to me? I don't expect to be correcting typos from your graphic designers."

All types of negative consequences were also subcategorized according to their severity, frequency, and form. (Refer to Table 8 for examples of each.) Three "levels" of severity were specified – diplomatic, harsh, and injudicious. Letting persons know they have erred or need to improve in a way that is not sarcastic, abrupt, or demeaning was the definition of *diplomatic*, doing so in a way that is non-denigrating, non-punitive, but possibly sarcastic or abrupt was *harsh*, and doing so in a way that is denigrating or overly punitive was considered *injudicious*.

The amount of times negative consequences were given in the response to any one item was assessed. Letting persons know they had erred or needed to improve, one time to one recipient was defined as a frequency of *once*. *Twice* was defined similarly, except the negative consequences were stated twice or to two recipients, and *three or more times* was defined as being stated three or more times or to three or more recipients, or some combination of these (e.g., delivered twice to one person and once to another).

Negative consequences were also differentiated in terms of the form in which they were delivered. *Classic* negative consequences were defined as conveying disapproval or doubt about another's work, noting that work had not been completed, or mentioning an error. Negative consequences that were categorized as *antecedent-like or accompanied by an antecedent* were similar to the classic form, except that the consequence was stated as a directive, instruction, or reminder. Negative consequences stated in the form of rhetorical questions were called *pseudo-monitors*. A simplified example shows the contrast between these three forms: "There is an error in your calculations." is a classic negative consequence, whereas "Fix the error in your calculations." is antecedent-like or accompanied by an antecedent; and "Didn't you see the error in your calculations?" is a pseudo-monitor. Refer to Table 8 for more complex examples from the OSIBA.

Providing antecedents. New to the OSIBA, the category of providing antecedents was defined consistently with the Operant Model as conveying expectations of performance, commonly known as directives, reminders, or instructions. However, in the in-basket, the behavior was defined as being limited to those instances in which it was used to motivate employees. For instance, a manager might use an exhortation such as, "This is really important; we need to do this." to motivate his/her employees to work on a particular project. This is an example of providing an antecedent. In one item in the in-basket, a marketing manager describes that his group is having problems getting a project completed that has been done many times previously. If a respondent were to suggest giving the marketing group an exhortation such as the one mentioned, giving an

inspirational speech on the importance of teamwork, reviewing deadlines, or assigning or clarifying team member responsibilities, these would be considered providing antecedents. If a respondent were to suggest that the marketing manager talk to people in the group about the nature of the issues occurring and what makes the current project different that before, it would be considered as monitoring. The *content* of the suggestion, or how to motivate the group to perform was what was important. The suggestions themselves would *not* be considered as providing antecedents because they were not being used to motivate the marketing manager.

Two types of antecedents were measured. The first type, giving *tacked-on traditional antecedents* was formally defined as giving directives when they were neither called for nor preferred (e.g., no directives were needed or requested). In several items, for example, respondents were presented with reports or status updates, and any instructions or reminders given were considered unessential. For example, in response to a message containing a status report for an upcoming banquet, one person gave instructions for each point in the report: "… I have the following comments: 1. We need to focus on a keynote speaker. … Touch base with Ethel and get on my calendar for this week. … Let's decide who we should invite and I will put a call in to them. 2. Talk with Nobu, I believe he knows the person who we used to engrave the plaques from last year's event. 3. Nobu, excellent job on finding a facility. You too, Bruce, great job. Let's keep up the momentum." (It was thought that this type of antecedent would not be related to effectiveness, and similar to the Jensen and Komaki [1993] study, might function as a suppressor variable in a multiple regression.)

The second type, *relying on antecedents*, was defined as using or recommending directives, instructions, or reminders alone to promote performance (when other strategies would be preferable). This type was measured under two conditions. In items in which respondents were presented with a problem and mixed or conflicting information about it such that questions might be raised (*unclear situation*), responses were categorized as antecedents when respondents did not

monitor (or recommend monitoring) to find out more about the problem or performance, and gave directives, instructions, or reminders to solve the problem. As an illustration, when a respondent was presented with a memo describing that a good employee who came to work on time in the morning was often late to meetings in the afternoon, he/she did not try to find out the reason for the lateness, but instead suggested providing directives or reminders: "It's time for you to reinforce with Dan that we expect him to be on time." (This type of antecedent was thought to be similar to the conditions in the Reynard and Komaki [1995] study in which the amount of antecedents provided before monitoring was negatively related to effectiveness.) In yet other items, respondents were asked to make a choice between training programs emphasizing lectures or instructions (antecedents) and feedback or practice (consequences) (*choose between antecedents and consequences*). Responses that indicated the person preferred the lectures and instructions were categorized as antecedents. This type was formally defined as choosing directives, instructions, or reminders over providing consequences in the form of feedback, when providing feedback is preferred.

Two additional subcategories were assessed for the tacked-on traditional antecedents and relying on antecedents in unclear situations. Similar to negative consequences, the amount of times antecedents were given in the response(s) to any item, or the frequency, was measured. Giving a directive, instruction, or reminder one time to one recipient was defined as a frequency of *once*. *Twice* was defined as directives, instructions, or reminders that were given twice or to two recipients, and *three or more times* was defined as three or more times or to three or more recipients, or some combination of these (e.g., delivered twice to one person and once to another). Refer to Table 8 for examples of responses for each frequency level. Another aspect, the "expanse" of the antecedent, was coded. If the directives, reminders, or instructions applied to multiple people or a group, it was noted as being *sweeping*. Writing memos to all, publicizing policies, or

suggesting training programs were examples of this because the antecedents were more far-reaching than those for a single person.

Revisions and Refinement

Addition of category and subcategories. The OSIBA used in the present research study was a revision of the original exercise. Items designed to elicit the behaviors of monitoring and providing consequences only were used in the original OSIBA (Komaki et al., 1990). Scoring in the original version was for the behaviors of monitoring, providing positive consequences, and providing negative consequences only. The original exercise itself and the scoring system were modified by the addition of a category of behavior and several subcategories before being utilized in the current study.

The major way that the in-basket was revised was to assess the category of providing antecedents. This was accomplished by adding five new items to the exercise and enhancing the scoring system of the original items to include providing antecedents. Finer distinctions were made among different types of positive consequences than in the original version of the in-basket. It was observed in responses to items during test administrations, for example, that some respondents thanked people for drawing attention to unwanted or undesirable information ("acknowledging the bearer of bad news") and others replied to memos by simply acknowledging that they had received the messages or letting the sender know what actions were being taken ("letting sender know of action"). These more subtle types of positive consequences were identified, defined, and consistently scored across both new and original items. The OSIBA was also revised to consistently capture the deservedness of both positive and negative consequences. Scoring on the original version for these behaviors was incomplete, in that it was noted on the score sheet of only one or two items. In addition, the form of consequences was a new aspect that was added to the coding system. This was developed for two reasons: (1) a research study indicated that subordinates had negative reactions when supervisors delivered negative consequences as rhetorical questions (called

pseudo-monitors because of the appearance of gathering information) (Bayazit, Komaki, Redding, Haimovich, & Kalt, 2000) and (2) it was thought that people who have a propensity to give antecedents might actually deliver consequences as directives or accompany them with antecedents. Another subcategory of consequences, and of antecedents as well, that was incorporated into the revised version of the in-basket was the frequency of the delivery of these behaviors. It was thought that it would not be effective for managers to go on and on with negative evaluations in particular, or with directives, so the score sheets of the revised OSIBA also included provisions for coding the frequency of the behaviors in responses.

Field testing. With a focus on those items that were being added to the exercise, the OSIBA was pilot-tested before being utilized in the present study. Seven potential antecedent-eliciting items were initially developed by brainstorming real-world situations in which it would be likely for a person to give commands, directives, and instructions to employees or to show that they prefer to give antecedents over other categories of behavior. For example, it was thought that persons who "believe in" the efficacy of providing antecedents would be expected to choose a training program for employees comprised of lengthy lectures over one with practice and receiving feedback. Situations were also crafted in which there was ambiguity about an underlying problem and monitoring would be prudent before proceeding with any action. It was thought that persons who preferred to give directives would give speeches or pep talks, or issuing warnings, for example, without monitoring. The potential items were integrated into the OSIBA and the exercise was then tried out on a sample of 19 people, including college students and a few acquaintances of the author who worked full-time in organizations and volunteered to take the in-basket.

The responses to each of the items were carefully scrutinized for evidence that the scenarios presented were *not* clear, or that the respondents misunderstood what was presented. As a result, three of the seven items were revised. For example, one item contained a message from the editor indicating that she would not be able to submit ideas for an upcoming issue by the deadline

(the current day) because of having problems completing changes for the issue coming out next. A few respondents were focused on a phrase used in the item--“something came up”--and did not conclude that what had “come up” for the editor was difficulty dealing with the changes for the next issue. (One response was, “Bee, your ‘something’ had better be meaningful.”) The item was reworded so the reason the editor could not meet the deadline would be clearer.

The items were tried out on 98 more college students from the Introductory Psychology course subject pool, and more attention was paid to the content and categorization of the responses. In fact, the previously discussed item was dropped from further refining and testing because there was so little variability in responses. As a result of this testing process, five items were added to the original OSIBA, for a total of 22 items. Six items tapped other behaviors besides monitoring, providing consequences, and providing antecedents. Responses to these items were not scored, so a total of 16 items were used for the purposes of this research. See Table 9 for a list of the items in the revised version and the categories/subcategories that were scored on each. The antecedent items that were added were numbers 2, 13, 14, 15, and 16.

The data collected during field testing were also used to create score sheets for new antecedent items and to modify those for the original monitoring and consequence items. The sheets were initially developed by examining the responses from the people in the test populations (i.e., college students and colleagues of the author), determining appropriate behavior categories (e.g., monitoring if it was an item developed to elicit that behavior and giving positive consequences if good work performance was described in the item), formulating the scoring and what was important in differentiating point values (e.g. monitoring via work sampling received more monitoring points than self-report), and fitting examples into the appropriate category and point values. Each score sheet contained definitions of all behaviors that could reasonably be expected to be exhibited in responses, examples, and point values. (Refer to Appendix B for an example score sheet.)

Table 9

Items in the Revised OSIBA and Points possible for Categories and Subcategories of Behavior

		Monitoring		Providing Consequences					Providing Antecedents	
		Self-report	Work Sample	Positive		Negative			Tacking on Unessential Ones	Relying on Them
Item	Name[a]			Warranted	Not Warranted[b]	Warranted	Questionably Warranted[b]	Not Warranted[b]		
1	Tickler	4	12	4	-2			-6		
2	Resistance	4		4			-6			6
3	Old Form	2	6	4			-6		5	
4	Rings	2	6	4			-6		5	
5	Sloppy Joe	2	6		-4	6	-6		5	
6	Accounts Gained	4		5		6	-6		5	
7	Foodstuf		6	2	-4	6	-6		5	
8	Map			3	-4	6	-6		5	
9	Costs	4		5			-6		5	
10	Impress	6		3						
11	Banquet			3					5	
12	Room	2		3						
13	Transformational			2						2
14	Training Schedule			2						2
15	Press	2	6	4			-6			6
16	We Warned	4		4			-12	-6		6
	Total	36	42	52	-14	24	-66	-12	40	22

Note. The table shows points possible designated on the score sheets for each item. Respondents could receive points for behaviors that were *not* designated on the score sheet. Points shown, unless otherwise noted, are the maximum number for the item.

[a] Six items were not scored for monitoring, providing consequences, or providing antecedents: Comp Time, Playwright, Postal, Bollinger's Score, Lost Copy, and Izmir. [b] Because the subcategory was negatively scored, the maximum number of *negative* points is shown (i.e., the minimum score).

The author and her adviser's graduate student research team practice-scored the responses from the test populations and calculated inter-rater agreement as a percentage score for each behavioral category (on each of the items). All disagreements were discussed and resolved. The score sheets were modified to reflect the decisions made, and revisions included any or all of the following: adding behavioral categories, honing the definitions, or adding examples or non-examples to better clarify the definitions. This practice scoring and revision process was carried out until all members of the team scored above 90% on each behavior in each item.

The in-basket was also administered to samples in three work settings prior to the bank – a group of 20 university administrators, 8 university library staff, and 51 university security sergeants. Inter-rater agreement was formally assessed for the sergeants, and the average across all the items was 95% (with at least 86% on each item). Agreement scores for the categories of monitoring and providing consequences were 86% and 82%, respectively (Cohen's Kappas were .78 for both). (The author and her advisor were considered primary raters for half of the data each, and two graduate students were reliability raters.) Responses were so rich and varied that modifications to score sheets were necessary after each administration. The scoring, thus, evolved into a very rich and detailed system.

Scoring

Scoring of responses to OSIBA items, in a general sense, was done by comparing responses against a "preset" standard of behaviors expected from effective managers for each item. Points for monitoring and providing consequences, for example, were awarded to responses when the respondent monitored and provided consequences in the scenarios where it was appropriate to do so. Specific scores were assigned for the behavior categories demonstrated in responses to each OSIBA item, with different values depending upon the presence of various qualities or subcategories. Higher points were awarded for responses that exhibited the desired supervisory behavior given in an appropriate manner for the situation.

Point allocation. As shown in Table 10, six points were given for the highest quality of monitoring. A response given six points meant that the respondent requested to observe employees in action or examine the products of their work – a "work sample." Five points for monitoring meant that the response was a work sample, but the respondent might have requested someone else do it, and four points meant that the respondent asked for performance information (a "self-report").

For the category of consequences, points were allocated differently for positive and negative evaluations. Scores from five to one were given for different types of positive when warranted consequences (see Table 10). A score of five was given to responses in which accolades and recognition for good work were expressed to multiple people (or whole departments). Four points were awarded to responses in which someone out of the normal contact loop was directly acknowledged. If a respondent thanked someone for bringing up unwanted or undesirable information or relayed the facts of a potential problem without casting blame, he/she received three points. Two points were given when respondents acknowledge that they had received senders' messages and/or informed them of their intended actions and one point when they gave a simple "Thanks" in their reply.

Table 10

Point Values for Various Qualities of Monitoring and Providing Positive Consequences

# of points	Category or Subcategory	
	Monitoring	Positive When Warranted Consequences
6	Work sample	
5	Work sample through someone else	Broadcast recognition to multiple parties
4	Self-report (asking about both problems)	Inform or acknowledge someone out of the normal contact loop
3	Partial self report (asking about only 1 problem)	Acknowledge the bearer of bad news Relay a potential problem non-judgmentally
2		Let the sender know of action
1		Simple 1 or 2 words (e.g., "Thanks")

Note. Higher quality = higher number of points (e.g., a monitor score of 6 is a higher quality than a score of 4). Behaviors that did not fit the specified descriptions were coded as "not identified" and assigned an appropriate point value.

The valuation for negative consequences was done differently than for other categories – points were subtracted from a "base score" for problematic aspects. As shown in Table 11, respondents received six points for delivering a negative consequence only if it was warranted (contingent on performance), delivered once in a diplomatic manner, and of "classic consequence" form (e.g., "You're missing an 'F' on FOODSTUFF."). Any deviation from this in terms of severity, frequency, or form resulted in point deductions from the base of six. With all possible deductions, a score of -3 could result. The following response was considered harsh, had a frequency of three or more negative consequences, and contained both classic and antecedent-like forms: "Perhaps you should check the spelling of Foodstuff [1, antecedent-like]. Really this kind of schoolboy error is pretty poor [2, classic]. Please correct before sending to Production [3, accompanied by an antecedent]. But I like the rest of the design." It was awarded two points for negative when warranted consequences (i.e., 6 [base] + -1 [harsh] + -2 [three or more] + -1 [antecedent-like]).

Consequences that were questionably or not warranted were scored separately from those that were warranted and given negative point values because they were thought to be undesired, ineffective, or inappropriate behaviors for the situation. (The rationale for this was based both in theory and research – contingent consequences are emphasized by operant conditioners and I/O psychologists and research has suggested there may be negative effects from the use of non-contingent consequences [e.g., Bass, 1990; House & Mitchell, 1974; Podsakoff & Schreisheim, 1985; Scott & Podsakoff, 1985; Skinner, 1974].) For these negatively scored behaviors, the lowest point value given was capped at -6 (and the highest was zero). Point valuation for negative consequences that were questionably and not warranted was similar to that for the warranted type discussed above. (Refer to Table 11.) Scores were assigned starting with a base score and then points taken off for the same three sub-subcategories of severity, frequency, and form. However, the "base" point value for each was -1 and was defined as the particular subcategory delivered once,

diplomatically, and in classic negative consequence form. Thus, when a respondent let a supervisor of a graphic designer know *his* performance was being negatively evaluated by writing, "Can you please make sure you review work before it gets sent to me [1, antecedent-like]? I don't expect to be correcting typos from your graphic designers [2, classic]." he/she received a score of -3 for questionably warranted negative consequences (i.e., -1 [base] + 0 [diplomatic] +-1 [twice] + -1 [antecedent-like]). Positive consequences that were not warranted were given either -4 or -2 points depending on the circumstances of the item.

Table 11

Point Values for Various Qualities of Providing Negative Consequences

Subcategory	# of points	Sub-subcategory	Quality
When warranted	6		Negative when warranted (contingent) "classic consequence," with diplomacy, and only once
When questionably or not warranted	-1		Negative when questionably or not warranted consequence, delivered once, diplomatically, and in classic negative consequence form
	0	Severity	Diplomatic
	-1		Harsh
	-4		Threaten to fire/demote
	-5		Zinger (fire/demote)
	0	Frequency	Once
	-1		Twice
	-2		Three or more
	0	Form	Classic
	-1		Antecedent-like negative
	-1		Pseudo-monitor

Note. Score = sum of "base score" (6 or -1) and any sub-subcategory quality point values coded in the second column. More than one form can be coded. Behaviors that did not fit the specified descriptions were coded as "not identified" and assigned an appropriate point value.

For the category of providing antecedents, the highest score was six points. The highest points were given to responses in which the only behavior exhibited was giving several directives, instructions, or reminders to a wide ranging group. As shown in Table 12, a "base score" of one

point was given to a single directive, instruction, or reminder given to a single individual, if the other desired behaviors for the item were shown. One point would be given to the respondent who, after receiving a memo describing an advertising customer's complaint about answering the phones, for instance, wrote to the sender, "Let's get together with the Advertising Director and discuss this. Either we have some real problems or Pheasant Peasant is just a little unreasonable. As an important client they deserve our best attention, so let's get to the bottom of it." This respondent requested monitoring the Advertising Department's performance via self-report by getting together with the Director to discuss it, but added the directive, "Let's get to the bottom of it." If other desired behaviors had not been shown, more than one antecedent had been given, or if the antecedent applied to many people or a group, points would have been added to the antecedent score. For example, a respondent would have received five points for providing antecedents if after receiving the customer complaint memo, he/she wrote, "To Everyone in the department [sweeping]: This is a regular reminder for all staff to refer to the Service Standard memo – last circulated and revised in December 2001 [frequency of 1]. All dept. heads should remind their subordinates to read this memo again carefully [frequency 2] and follow in their working environment [frequency 3]. Thanks." (i.e., 1 + 1 [no monitoring of advertising performance] + 1 [sweeping] + 2 [three or more]).

An option, called "not identified," (or "ni") was available to scorers for assigning point values to responses that did not seem to fit the pre-existing definitions of the subcategories or qualities outlined above. In these situations, scorers suggested a point value on the score sheet for the category under "not identified." (Points could be added or subtracted.) Scorers could also write in a category coded "ni" that was not even listed on a score sheet of a particular item if a response could be coded as that behavior. This scoring option was included so that categories that had never been or were very rarely displayed in responses to particular items would be captured.

Table 12

Point Values for Various Qualities of Providing Antecedents

# of points	Sub-subcategory	Quality
1		Directive, instruction, or reminder to a single individual in addition to other desired behaviors for item
1		No other desired behaviors exhibited
1	Sweeping	Applies to multiple people or entire group
0	Frequency	Once
1		Twice
2		Three or more

Note. Score = sum of "base score" of 1 and any sub-subcategory quality point values coded in the first column. A base score of 2 was used for the subcategory of relying on antecedents. Behaviors that did not fit the specified descriptions were coded as "not identified" and assigned an appropriate point value.

Points possible. The points designated on the finalized scores sheets for the major categories of monitoring, providing consequences, and providing antecedents, and for subcategories within them were displayed in Table 9. For each of the 16 items scored, the points possible for the category/subcategory are shown -- the maximum for those behaviors scored positively and minimum for those scored negatively. (The maximum score for negatively scored behaviors is zero.) It is, however, extremely unlikely that a respondent could get all the points possible (either negative or positive).

Monitoring via any method was present on 12 of the 16 score sheets. Monitoring via self-report was scored in 11 items with a total of 36 points possible and monitoring via work sampling in 6 items for a total of 42 points. Positive consequences that were warranted were possible on nearly every sheet – 15 – for a total of 52 points. Respondents could also have a total of 14 points subtracted for positive consequences that were not warranted, which were scored in four items. Providing negative consequences that were warranted was possible in four items with a total of 24

points. Sixty-six points could potentially be subtracted for questionably warranted negative consequences, scored in 10 items, and 12 points for not warranted negative consequences in two items. Tacking on unessential antecedents was scored in eight items, with a total of 40 points possible and relying on antecedents in five items, with a total of 22 points possible.

Category score calculations. Category scores were calculated by dividing raw scores for categories by the total points possible, and then multiplying by 10. Raw scores for each category were calculated by summing the scores for the categories over all the items for each person. See Table 13 for an example of scoring for monitoring for a respondent. Monitoring was possible on 12 items in the in-basket, for a total of 66 points possible. As shown, Respondent A monitored on seven items, earning 28 points. Respondent A's score for monitoring, thus, was 4.2 (equal to 28 divided by 66, multiplied by 10). Scores for major subcategories were calculated in the same fashion. Respondent A displayed in Table 13 monitored via work sampling on three out of six items in which it was possible. He/she received 17 points out of the possible 42, and the score for work sampling was 4.0. Category and subcategory scores were generally on a scale of zero to 10, though lower or higher scores could potentially be possible if participants had points subtracted or added from responses scored as "not identified."

In effect, overall raw category scores were calculated by adding all subcategory scores, as shown in Table 14. The raw score for monitoring overall was the sum of points for the methods of monitoring, work sampling and self-report. If a respondent received 17 points for monitoring via work sampling and 11 points for self-report, the overall monitoring score would be 28 points. In the same way, providing antecedents was calculated as the sum of the two types of antecedents measured – tacking on traditional and relying on antecedents.

Table 13

Scoring Example for Monitoring and Monitoring via Work Sampling for a Single Respondent

Item #	Total Monitoring Points Possible[a]	Work Sampling Points Possible	Respondent A Monitoring Points Earned	Respondent A Work Sampling Points Earned
1	12	12	0	0
2	4		3	
3	6	6	6	6
4	6	6	6	6
5	6	6	5	5
6	4		2	
7	6	6	0	0
9	4		2	
10	6		0	
12	2		0	
15	6	6	0	0
16	4		4	
Total	66	42	28	17

[a] Total monitoring points possible is equal to the greater of points possible for monitoring via work sampling or self-report.

Providing consequences was calculated in two ways, called “warranted only” and “questionably and not warranted subtracted.” Providing “warranted only” consequences was calculated simply as the addition of points for positive and negative consequences that were warranted. The second calculation, essentially an overall or total score, was defined the sum of warranted positive and negative consequences minus the points for positives that were not warranted and negatives that were questionably and not warranted. These definitions were developed because previous research had not included measurement of non-contingent (not warranted) and questionably warranted consequences and so that a measure of consequences in which nothing was subtracted for these types of deservedness could be tested.

Table 14

Scoring Calculations for Major Categories

Category	Scoring Calculation
Monitoring (overall)	monitoring via work sampling + self-report
Providing antecedents (overall)	tacking on traditional antecedents + relying on antecedents
Providing consequences (warranted only)	warranted positive + warranted negative consequences
Providing consequences (questionably and not warranted subtracted)	warranted positive + warranted negative – (unwarranted positive + questionably warranted negative + unwarranted negative)[a] consequences

[a] These subcategories were negatively valued.

Additive and multiplicative score calculations. Two single, overall scores on the in-basket were also calculated for each participant. The first was an additive score, in which scores for monitoring and providing consequences overall were summed. To compute this score, the raw scores for both categories were divided by the points possible for each, multiplied by ten, and then added together. For example, a respondent's raw score for monitoring was 34 and consequences (questionably and not warranted subtracted) was 21. The points possible for monitoring was 62 and 82 for providing consequences, thus yielding scores of 5.48 and 2.56, respectively. The overall additive score was thus 8.04. The second overall score was multiplicative, in which scores for monitoring and providing consequences were multiplied. To calculate this score, the raw scores for monitoring and providing consequences were first standardized by converting them into z scores. The constant of five was then added to the standardized scores for both categories so that no score would be negative, and then the resulting monitoring and providing consequence scores were multiplied together. For the example respondent above, the z score conversions with five points added were 6.34 for monitoring and 5.92 for consequences, resulting in an overall multiplicative score of 37.57. (The same procedure was followed using the alternative warranted only definition of consequences in place of the one that included all types of deservedness.)

Responses

Scoring was detailed and fairly complex because respondents' replies to the OSIBA items typically varied and were sometimes quite complicated. Table 15 shows one item, a sample of responses, how those responses were categorized, and the point values awarded. The item shown in the table contained a report, sent by the Director of Advertising, displaying the number of advertising accounts gained and lost each year for the last four years. Three of the four years showed gains, with the highest and most outstanding gain in 2002. One year, 2001, showed a loss of three accounts. There was also a typographical error in the number of accounts shown for the end of the year in 2001. The focus or intent of the item was to generate both positive and negative consequences (for the large gain and typographical error, respectively).

In the sample of responses shown in Table 15, one can observe that some respondents saw the error and mentioned it to the writer (i.e., example numbers 1 through 9), receiving points for providing negative consequences. One example, number 2, is less diplomatic in pointing it out, and thus received fewer points for negative consequences than those who did it more diplomatically (i.e., numbers 1, 3 through 9). Some did not indicate they noticed the error (i.e., numbers 10 through 14). Respondents may have acknowledged the accounts gained, which was awarded points for providing positive consequences (i.e., numbers 3, 4, 6, 7, 8, 9, 11, 12). Higher point values were given when the congratulations were broadcast to multiple parties, as in example number 11. In addition, thanking the sender for the report (i.e., numbers 3, 9, 10, 13) was considered providing a positive consequence (but of small point value). Some respondents asked questions about the data and wanted more information on the patterns shown, which was considered monitoring via self-report (i.e., numbers 5 through 10). Some focused on the gains only (i.e., numbers 6, 9), others on the loss only (i.e., numbers 5, 7), and others on both (i.e., number 10). Also, some respondents set goals or expectations for the number of accounts for the following year, a behavior that was defined as providing antecedents (i.e., numbers 9, 11, 14). This variety of tactics in responding can,

furthermore, occur in different combinations and through responses to multiple people or groups. Table 15 shows several examples of this, as well (i.e., numbers 3 through 11). Responses 8 and 11 show replies to different people or groups besides the memo writer.

Table 15

An OSIBA Item and Examples of Responses

Item 6

TO: You
FROM: J. Child, Dir., Advertising
SUBJECT: Advertising accounts by year

Here's the tabulation of the number of accounts as gained and lost by Advertising Dept.:

	# Accounts at beginning of year	Changes in account status during year	# Accounts at end of year
1999	70	+7	77
2000	77	+9	86
2001	86	-3	53
2002	83	+12	95

Responses

Ex. No.	Example[a]	Category(ies)
1	Year 2001: either the numbers are wrong or I'm missing something …	N(6)
2	[Circled 53 and wrote "?" next to it] Please let's discuss! Quality!!!	N(3)
3	Thanks. This shows your hard work turned things around after the decline in 2001. By the way, I believe the FYE 2001 number should be 83, not 53.	P(4), N(6)
4	Good job. You guys are impressive. Please note typo (83 instead of 53). Remember: QUALITY!	P(4), N(4)
5	Please check your numbers for any typos or similar mistakes and revert to me by tomorrow morning. I'd also like more detail on the accounts lost in 2001 (e.g., names, annual revenue, where we believe they are advertising now, date of last attempt to get them to advertise in our magazine again),	N(5), M(2)
6	Congratulations on the 2002 result in particular. I'd be interested to hear some background on the reasons for the growth, especially after account losses in 2001. Is there a typo for the 2001 "# Accounts at YE"?	P(4), M(3), N(6)
7	Looks pretty impressive, J. Do you have any ideas on why 2001 went down? Who did we lose out too? Any lessons to be learned there? Also, how does this compare to our competition? P.S. Note the typo for year end 2001, I guess it should be 83, not 53.	P(4), M(2), N(4)

Table 15, cont.

Ex. No.	Example[a]	Category(ies)
Responses		
8	Well done on gaining 12 new accounts in such a difficult economic environment! However, I noticed that there is an error in the 2001 figure. The total accounts should be 83 rather than 53! To Accounting Director: This year we gained 12 new advertising accounts, with the total number of accounts now at 95. Please can you let me know what is the revenue increased this year?	P(5), N(5), M(1)
9	Thanks for the information. Could you please help me to better understand these numbers by providing your opinion on: (a) a reasonable target for 2003 and why, (b) why the huge impressive turnaround in 2001-2002. You might want to correct the typo in 2001 before distributing the table elsewhere.	M(3), A(1), P(3), N(5)
10	Thank you. Could you give me an explanation for the decrease in 2001 to 83 and then increase in 2001 to 95. What motivated our advertisers and what was the impact on earnings?	P(3), M(4)
11	Great work—let's try to get over 100 by midyear 2003. That would certainly be a landmark. To All Direct Reports: [Forward results at the top.] Great work by the Ad team, but also a testament to all of the company improving the quality of our magazine. Thanks to you all.	P(5), A(1)
12	Great reversal of last year's falloff in new accounts!	P(4)
13	Thank you!	P(1)
14	Prepare in-house posters promoting going over 100 next year.	A(2)

Note. N = providing a negative consequence, P = providing a positive consequence, M = monitoring, A = giving antecedent(s). Numbers in parentheses following the categories indicate the score.

[a] All responses, unless noted, were written to the sender of the original item, J. Child.

Scoring of Responses

The author and her adviser scored all responses received. The author is knowledgeable about the categories and definitions in the Operant Model of Effective Supervision and was trained in scoring the OSIBA. Training included instruction and practice in identifying the categories of behavior shown in respondents' answers to items from previous administrations of the in-basket. When she had three consecutive overall inter-rater agreement scores of 90% or better with her

adviser, she was considered "qualified." Coding the responses for the 35 people in the sample took an average of 2 hours per item. Thus, for the 16 items, it took 32 hours for one person to code the responses, yielding a total of 64 hours.

All categorizations made by the two scorers for all participants on all items were compared to see whether they matched exactly. Every disagreement, regardless of whether it was regarding category, subcategory, or sub-subcategory, was reviewed and discussed. Final categorizations and scores were determined by consensus agreement. In addition, anything scored as "not identified" by either scorer was reviewed so that a final point value could be decided. Reconciliation of disagreements and discussion of not identified responses took anywhere from 1.5 to 3 hours per item depending upon the complexity of the item (and amount of disagreements). Approximately 35 hours was spent in this activity. Thus, for the two scorers, a total of 134 hours (64 hours for scoring and 70 for reconciliation) was spent in scoring the responses from the group (which averaged 3.8 hours per respondent).

Criterion: Motivational Effectiveness

A judgmental measure of effectiveness similar to those used in several previous studies (e.g., Jensen & Komaki, 1993; Komaki, 1986; Komaki et al., 1989) was utilized in this research study. Participants were rated by other study participants (and the Managing Director) in terms of motivating others, on a graphic rating scale with "grade-related" anchors of *A* through *F* (*A* being the best, *F* the worst), and then ranked. (Jensen and Komaki [1993] found acceptable levels of test-retest reliability for both the rating and ranking measures [.76 for ratings and .93 for rankings].)

From a list of all potential participants (52 names), participants were instructed to rate and rank all people they had directly observed, or seen in action. They checked off the context(s) in which they had observed each person they were rating (i.e., in the role of manager or supervisor, creating or helping to create deal-specific teams, and supervising informally or coaching), noted a letter grade rating for motivating others, and marked if they were currently or had been a supervisee

of the person. Participants then completed rankings based on the grade ratings they had assigned, which was called "ranking within rating." If a participant gave three people As for motivating others, for instance, he/she ranked the three people from one to three (with one being the highest rank). If five Bs were assigned, the five people given that rating were ranked from one to five, and so on for all grade ratings given.

The actual rating form was set up as a table, with a column containing the list of 52 names. Other columns were provided to list the ratings and rankings for effectively motivating others, mark the context(s) in which the rater had observed each person rated, and note whether the rater was or had been a supervisee of the person rated. Refer to Appendix C for a blank copy of the form and an example showing hypothetical data, which was presented in the instructions to the participants.

A single effectiveness score was calculated by substituting the numbers of five through one for the letter ratings of A through F, respectively, and then calculating the average for each participant. (If a participant, for example, received 10 ratings consisting of 5 As and 5 Bs, his/her effectiveness score would be equal to 4.5 [5 points for each of the 5 As plus 4 points for each of the 5 Bs, or 45 points, divided by 10 ratings].) Also, means for ratings provided by supervisees and others were calculated separately. Because not all participants formally supervised others, supervisee ratings were obtained for 20 participants.

Two other exploratory effectiveness scores were calculated as composites of the ratings and rankings, in the event that the ratings alone did not adequately differentiate between the participants. One took the rank within the letter grade into account and in essence, added "gradations" to the ratings. On each rating form submitted, those rated were placed in order by grade rating and then ranking, and the numerical points assigned to each person depended upon the rating, the number of people given that same rating, and the position within the group receiving the rating. If there were five As assigned, for example, the first person ranked A received 5 points, the second received 4.8 points, the third 4.6 points, the fourth 4.4 points, and the fifth 4.2 points. If

there were then two Bs assigned, the first person received 4 points and the second 3.5 points. Similarly, if 3 Cs were given, the first-ranked person received 3 points, the second 2.67 points, and the third 2.33 points. Means were then calculated by participant.

The second exploratory motivational effectiveness score utilized the overall placement of rated participants. To calculate this score, the participants rated on each form were ordered by grade rating and ranking, yielding an overall ranking. Percentiles were then calculated for each person rated. (The person ranked last was assigned a percentile of .01, instead of zero.) Then the percentile was multiplied with the numeric version of the grade rating of 1 through 5. (Take as an example if one person rated 11 others and gave everyone As: the score for participant 1 [the highest ranked] was 5.0, participant 2 [at the 90th percentile] was 4.5, participant 3 [at the 80th percentile] was 4.0, and so on to participant 11 [the lowest ranked], whose score was 0.05.) Averages were then calculated for each participant. The score was calculated in this way so as to give more "weight" to the ranking information. (Participants consistently ranked near the top of all raters' lists would get high scores and those ranked often at the bottom would get very low scores.)

In addition, participants provided ratings of their own motivational effectiveness at the end of the in-basket. These data were not included in the calculation of effectiveness scores for each participant, but used as a comparison with other rating sources.

Orthogonal Measure: Ratings of Technical Knowledge

To determine whether monitoring and providing antecedents and consequences were *not* related to an aspect of supervision they were not supposed to be related to, an item regarding the participants' technical knowledge was included in the rating form. This technical knowledge question was also included so that raters might not be as hesitant to rate someone poorly on motivation--they are able to give a good rating on technical knowledge to "make up" for a bad motivational rating. A similar rating was used in a previous study, as well (i.e., Komaki, 1986).

The same A through F rating scale used for the motivational item was also used for the technical knowledge question. See Table 15 for how this was presented to the participants. The technical knowledge score was calculated similarly to the motivational effectiveness score, i.e., substituting the numbers 5 through 1 for A through F, and calculating the mean for each participant. Means for supervisees' and others' ratings were calculated, as well. Self-ratings of technical expertise were also gathered.

Procedure

The study was conducted as an organization development initiative, designed to give the participants a personal growth and development opportunity. To "kick off" the project, the Managing Director of the group sent an e-mail to all potential participants introducing the project, expressing that participation is voluntary, and encouraging all to take advantage of the opportunity to receive individual coaching and increase their managerial skills. Refer to Appendix D for a copy of the email message.

Informed consent

An informed consent document and introductory letter from the author's adviser were attached to the e-mail. The informed consent form explained that the information collected was being used for research purposes, individual responses would not be shared with anyone, and consent could be withdrawn at any time without penalty. The potential participants were also informed that no one at the bank would know who participated and who did not, and all materials would be sent directly to the researchers.

Data collection

The effectiveness rating form was included with the "kick-off" e-mail (along with the letter from the author's adviser and the informed consent document). Respondents were given a number of ways to return the forms to the researchers, such as via fax, e-mail, or regular mail. See Appendix D for a copy of the correspondence potential participants received.

A conference call was scheduled for one week after the kick-off e-mail. This meeting was conducted by one of the organization's leaders to introduce the author and her adviser, explain the project and put it in context, discuss the procedures, and answer any questions the potential participants had. In the several weeks following the conference call, 30 informed consent and 33 effectiveness rating forms were received. (Three ratings were returned without consent forms or any identification, and were not used in the research.)

The author then randomly assigned a code number and one of the five versions of the in-basket to each potential participant. The code numbers were assigned to ensure confidentiality and guard against biased scoring. Printed copies of the in-baskets (marked with the code numbers) were mailed to those who had returned informed consent and effectiveness ratings. This document was comprised of (a) a title page, (b) a letter giving the deadline for completing the exercise and instructions for returning it, (c) a license agreement, (d) instructions for responding to the in-basket items and describing the fictitious situation, (e) the 22 in-basket items followed by space to respond, and (f) a demographic information page. Packages containing a letter offering a second chance to participate and the informed consent document, effectiveness rating form, and in-basket were sent to each of the other potential participants. The instructions indicated that the in-basket should take approximately 90 minutes to complete and all items were to be returned to the researchers in three weeks time.

Twenty-five in-baskets were returned at the time of the deadline (and shortly thereafter). The relatively small number of returns was discussed with the leaders of the group, and it was agreed to extend the deadline for a few weeks, especially to accommodate those who had consented but had not returned the in-basket. During this time, the author, another graduate student, and her adviser placed phone calls to all who were on the original list of potential participants. The people who had returned all of the materials were informed that their materials had been received, thanked for participating, and informed of the revised project timeline. Those who had consented and

provided effectiveness ratings were informed of the extended deadline and asked if they had any questions on the in-basket. Those who had not responded at all were asked if they had any questions and assured that participation was voluntary. Everyone was told that calls were made to all potential participants so that the bank would not be able to tell who did and did not participate, if phone records were inspected. By the time the last portion of the project began about two months later, 10 more completed in-baskets were returned, for a total of 35.

Feedback

The last phase of the project involved presenting group-level results and giving individuals face-to-face feedback on how they scored on the in-basket. Two video-conference sessions were held at the bank, one for the United States and Europe and the other for the Asia-Pacific region. The presentation included definitions of the behaviors in the Operant Model of Effective Supervision (Komaki, 1998), highlights from the research done on the model to date, verbatim examples of the behaviors from the in-basket, results for the group (means for supervisory behaviors and effectiveness and correlations between the behaviors and effectiveness), and a description of how the individual feedback would be presented. The leaders of the group were given a preview of the presentation and received their individual feedback before the rest of the group.

One-on-one feedback sessions were scheduled with the author and her adviser during a four-day period over three weeks. Arrangements were made to meet in person for participants located in the United States and via video-conference for those in Europe and Asia-Pacific. The sessions were approximately 20 minutes in length and consisted of a presentation of the participants' scores on each of the behaviors measured, comparisons to the means for the group, and discussion of examples of responses that supported the scoring. If, for example, someone had a high score for monitoring via work sampling, he/she was shown the items and responses in which he/she monitored via work sampling. Other items in which the person did *not* monitor via work sampling, but for which this was a preferred response were also shown. The person's responses to these items

were reviewed and a dialogue was conducted about how monitoring via work sampling might be possible in the context of the items. Average scores on motivational effectiveness and technical knowledge were presented upon request. Participants left the sessions with a printed copy of a short summary of the behaviors that were measured in the in-basket and their scores in chart form. See Appendix E for an example of this feedback sheet.

Chapter 8

RESULTS

Psychometric Properties of the Predictor

Scoring Consistency

The consistency of scores on the in-basket was assessed in two ways: through inter-scorer percentage agreement and intraclass correlations.

Interrater Percentage Agreement

Interrater or inter-scorer agreement was assessed by directly comparing scores of two scorers for categories or subcategories coded in participants' responses to the in-basket items. Agreement was assessed for 100% of the participants' responses (not for just a sample). Comparisons for agreement were made stringently, on an item-by-item basis, for each respondent. Inter-scorer agreement was calculated as a percentage score. For the purposes of computing scoring consistency only, percentages were computed for scoring units and sub-units in each item. One scorer, the author, was designated as primary (PO, for primary observer) and her adviser was reliability (RO, for reliability observer). When the PO coded a unit for a respondent on an item (e.g., monitoring), an agreement was noted when the RO coded the same unit. The counts of the units the PO coded and the agreements were summed across respondents for each item. The formula to calculate percentage agreement for units for an item was:

$$\frac{\textit{\# of agreements for unit}}{\textit{\# of occurrences of the unit the PO coded}} \times 100.$$

Units consisted of the categories of monitoring, providing consequences, and antecedents, and the major subcategories of each of them -- positive consequences that were warranted, positive consequences that were not warranted, negative consequences that were warranted, negative consequences that were questionably warranted, negative consequences that were not warranted,

tacked-on traditional antecedents, and relying on antecedents. The units scored on each item differed. A scoring option developed specially for this scoring system was one referred to as "not identified" (or "ni"). When a respondent's reply did not fit any of the definitions or examples given, but the coder judged it to be a monitor, for example, the coder had the option of coding it as a monitor "not identified" on the code sheet. These responses coded as "not identified" were also considered as a particular unit.

Agreements were counted when the PO and RO both identified the same unit. One agreement was counted, thus, when the PO and RO both identified monitoring in a response to a particular item. Monitoring generally ranged in quality, often with the highest score of six points and the lowest value of one. If both the PO and RO gave a response six points for monitoring, it would be an agreement for the unit of monitoring. The PO could code a response as a 4-point monitor and the RO a 2-point monitor, which would also constitute an agreement because both had identified the same unit. In addition, if the PO coded monitoring "ni" for a response and the RO assigned three points for monitoring, this would be an agreement for monitoring. When the RO did *not* code monitoring and the PO did, this would be a disagreement.

The denominator in the formula -- the number of occurrences -- was calculated as the number of times that the PO identified a unit. When the PO coded a warranted positive consequence for a response to an item, for example, that counted as an occurrence of the unit. If the PO scored monitoring "ni," this was an occurrence of monitoring.

Interrater agreement for sub-units, or types of units, was calculated as follows:

$$\frac{\text{\# of agreements for sub-unit}}{\text{\# of occurrences of the sub-unit the PO coded, when PO and RO agreed on the unit and neither coded ni}} \times 100.$$

Sub-units consisted of methods of monitoring (i.e., work sampling, self-report) and special types of positive when warranted consequences (e.g., acknowledging the bearer of bad news, broadcasting to multiple parties, relaying a problem non-judgmentally).

Similar to the agreement for units, agreements for sub-units were counted when the PO and RO both identified a particular sub-unit. If the PO coded a response as acknowledging the bearer of bad news for three points and the RO did the same, it would be an agreement for the sub-unit of acknowledging the bearer of bad news. For the method of monitoring, if the PO chose a 5-point monitor and the RO a 6-point monitor for a response on an item, it would be an agreement for work sampling (the sub-unit), because both were defined as monitoring via work sampling. If, however, the RO had coded a 2-point monitor, this would have been a disagreement for work sampling because the 2-point monitor is defined as via self-report.

The denominator in the formula was calculated as the number of occurrences of the sub-unit the PO coded, when the PO and RO agreed on the unit and neither coded "ni." For instance, both the PO and RO had to code a monitor to assess whether there was an agreement for the method. Then and only then could it count in the denominator for the method sub-unit. The exception was any unit coded as "not identified" on the scoring sheet -- these were not counted as an occurrence for any sub-unit. If the PO scored monitoring "ni" and the RO had monitoring via work sampling, for example, no occurrence for any sub-unit of monitoring was counted.

Percentage scores were calculated by summing the agreements and occurrences (for units and sub-units) across all respondents and in-basket items. Hence, when coders agreed on 8 instances of monitoring via self-report out of a total of 10 identified by the PO, the self-report sub-unit score would be 80%.

Table 16 shows the interrater agreement percentage scores for each of the categories and subcategories of behavior, averaged across all items scored. Scores for the major categories were good: for monitoring, the average was 91.4% (ranging from 78.3 to 100.0% across the items in

which it was scored), 89.0% for providing consequences (ranging from 33.3 to 100.0%), and 82.5% for providing antecedents (ranging from 0.0 to 100.0%). These average percentage scores were within the limits of acceptability for a new measure, 80% (Miller, 1997). These results indicate that the in-basket resulted in reliable scoring, in the sense that two people independently judged the same behaviors to be the same category.[1]

Results for subcategory agreement ranged from 64.3 to 100%, with most subcategories having agreement percentages greater than 87. For monitoring, results were excellent -- when the scorers agreed monitoring was present in a response (and both could categorize it within the scoring system), they agreed nearly always on the method (99.0%). Agreement percentages for positive and negative evaluations of consequences are presented separately because, as mentioned above, each was calculated as an individual unit. Scores for warranted positive consequences and the subcategory of directness were fairly good (90.1 and 89.4%, respectively). Acceptable scores were achieved for subcategories of special types (81.6%) and antecedent-like (83.3%) positive consequences, even though many of the special types and the antecedent-like form were new to the scoring system. The agreement for positive consequences that were not warranted was fairly low – 64.3% -- but the frequency of their occurrence in the sample was also low. The subcategories for negative consequences were new additions to the score sheets as well, and scores for them were fairly good, with one exception. Agreement percentages for the three types of deservedness, severity, and form ranged from 87.0 to 100.0. The one aspect that had the lowest agreement was frequency (65.0%). This subcategory proved to be difficult in both definitional development and scoring because counting the occurrences of the behaviors was not simple and straightforward. (Refer to Table 8 for examples.) Subcategory scores for providing antecedents, a completely new

---

[1] Another way of assessing agreement is to compare the actual values each scorer assigned for a category (or subcategory). This method has an advantage in that agreement is calculated very strictly. A disadvantage, however, is that agreement percentages may be underestimated because values are assigned based on subcategories or other qualities.

category, were within an acceptable range (77.8 and 87.8% for the two types). The agreement for whether an antecedent was "sweeping" was 100%, but only one was found in the sample.

Table 16

Average Interrater Agreement Scores for Categories and Subcategories

Category/ Subcategory	Interrater Agreement[a] Frequency[b]	Percentage
Monitors	**256/280**	**91.4%**
Method: Work sampling or self-report	98/99	99.0%
Consequences	**389/437**	**89.0%**
Evaluation: Positive	227/256	88.7%
Deservedness: Warranted	218/242	90.1%
Special types: Relaying problem non-judgmentally, etc.	146/179	81.6%
Directness: Directly or indirectly delivered	34/38	89.4%
Deservedness: Not warranted	9/14	64.3%
Antecedent-like: Antecedent-like or not	15/18	83.3%
Evaluation: Negative	162/181	89.5%
Deservedness: Warranted	67/72	93.1%
Deservedness: Questionably warranted	94/108	87.0%
Deservedness: Not warranted	1/1	100.0%
Severity: Diplomatic, harsh, or injudicious	139/157	88.5%
Frequency: Once, twice, or three or more times	106/162	65.4%
Form: Simple negative, antecedent-like, or pseudo-monitor	172/195	88.2%
Antecedents	**85/103**	**82.5%**
Type: Tacking on traditional antecedents	42/54	77.8%
Type: Relying on antecedents	43/49	87.8%
Expanse: Sweeping or not	1/1	100.0%

[a]Agreement refers to occurrences of categories/subcategories; not to the numerical scores assigned to them [b]Frequency = number of agreements/number of occurrences.

Intraclass Correlations

The consistency of raters' scores was also assessed through calculating intraclass correlation (ICC) coefficients. ICCs are an index of the relationship among variables of a common class – those that share metric and variance (e.g., ratings provided by different raters). (*Inter*-class correlations, such as the Pearson *r*, on the other hand, are indices of the relationship between

variables that share neither metric nor variance [e.g., grade point average and IQ].) ICCs are measures of the degree of homogeneity of nominally-scaled classes relative to the total variability in interval-scale variables. When there is no within-class variation (the categories are perfectly homogeneous), the value of the intraclass correlation is +1.0. The correlation will be equal to zero when the categories are as homogeneous as would be expected by chance if there were no relationship between the variables (Blalock, 1960).

Numerous versions of ICCs are available for different purposes and research conditions (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Shrout and Fleiss (1979) described three models based on the judges or raters of each target. Model one is appropriate in situations where each target is rated by a unique rater, sampled from a population of raters, and is known as a one-way random effects model. The second model is used when all raters randomly sampled from a population of raters make judgments on each target, and is known as a two-way random effects model. The third model is similar to the second in that all targets are rated by each rater, but the raters are assumed to be the population of raters. This model is called a two-way mixed model because raters are a fixed factor (rather than random).

Two additional types of two-way models are available, which differ in terms of how error is defined (McGraw & Wong, 1996). One type of ICC uses a consistency definition in which rater variance is considered irrelevant. The other uses an absolute agreement definition in which rater variance is considered error and is included in the denominator of the ICC formula. The consistency definition provides an index of the similarity of rank or relative ordering of targets between raters, whereas the absolute agreement definition provides an index of the similarity of the magnitude of ratings between raters. The former is actually a measure of interrater reliability and the latter interrater agreement (McCloy & Putka, 2004).

In the present study, the in-basket scores for each participant on each category were summed over the items for each scorer (primary observer, or PO, and reliability observer, RO)

separately. ICCs for each category were computed using two-way mixed effects models. This model was chosen because the PO and RO comprised the population of raters – they were not a randomly chosen sample – and they provided scores on all categories for all participants. Both consistency and absolute agreement definitions were used to calculate the ICCs, since differences between the raters in the actual value of scores was of interest, not simply the rank ordering.

As shown in Table 17, consistency ICCs for categories ranged from .73 to .88 and absolute agreement ICCs ranged from .68 to .87. The scores, overall, were fairly high (with many exceeding reasonable reliability estimates of .70), indicating that raters scored participants' behaviors consistently. Absolute agreement ICCs for providing antecedents and negative consequences, however, fell slightly below the ".70 mark". Providing antecedents was a new category and some subcategories of negative consequences were never assessed before, so it was expected that agreement and consistency might be lower than for other categories, such as monitoring and providing positive consequences. (Results, also, were generally consistent with interscorer percentage agreements.)

For the major subcategories of behaviors, ICCs ranged more widely -- consistency ICCs were from .42 to 1.0 and absolute ICCs from .33 to 1.0. Many of these coefficients were above .70, indicating that raters scored the subcategories of behavior consistently. The three exceptions were positive consequences that were not warranted, negative consequences that were questionably warranted, and tacked-on traditional antecedents. Unwarranted positive consequences and tacked-on traditional antecedents also had relatively low percentage agreement scores, and were discussed earlier. The subcategory with the lowest ICCs, questionably warranted negative consequences (a brand new subcategory), did *not* have a low percentage agreement score. The relatively low ICCs might be explained by the fact that ICCs are based on comparisons of the actual scores. (Percentage agreement, on the other hand, is based on comparisons of the presence or absence of a code for category or subcategory, not on the numerical score.) Scores for questionably warranted negative

consequences were determined by the coding of other aspects, such as the severity/diplomacy of the response, the frequency of consequences given, and the form in which the consequences were given. (Refer to Table 11 on the scoring of negative consequences.) Thus, ICC coefficients might be lowered if raters frequently disagreed in the values assigned to questionably warranted negative consequences, even though they agreed that the subcategory was present.

Table 17

Intraclass Correlations for In-basket Category and Subcategory Scores

Category/Subcategory	ICC (C,1)[a]	ICC (A,1)[b]
Monitors	0.88	0.87
Work Sampling	0.93	0.93
Self-report	0.88	0.88
Positive Consequences	0.81	0.74
Warranted	0.95	0.94
Not Warranted	0.48	0.48
Negative Consequences	0.75	0.68
Warranted	0.87	0.87
Questionably Warranted	0.42	0.33
Not Warranted	1.00	1.00
Antecedents	0.73	0.68
Tacking on traditional antecedents	0.62	0.59
Relying on antecedents	0.79	0.78

[a] Intraclass correlation using a consistency definition. [b] Intraclass correlation using an absolute agreement definition.

Descriptive Statistics

Table 18 shows the summary of scores on the in-basket for the categories of monitoring, providing consequences, and providing antecedents, selected subcategories of each, and overall additive and multiplicative combinations. For easier interpretation, each category/subcategory is presented with two lines of data in the table. The top line represents scores calculated by dividing raw score points by the total points possible for the category/subcategory and then multiplying by 10. Raw scores are shown on the bottom line in parentheses. The calculated scores, in general, range from zero to ten, with higher values indicating behaviors that were exhibited more often than ones with lower scores.

As shown in Table 18, scores varied somewhat among the categories and subcategories, and when considering the calculated score means, they were generally fairly low. The behaviors that were exhibited the most in regard to what was possible were monitoring via self-report (*M* = 4.53), monitoring in total (*M* = 3.82), providing positive when warranted consequences (*M* = 3.53), and providing negative when warranted consequences (*M* = 3.46). Providing antecedents in total (*M* = 0.72), the various subtypes of antecedents measured (e.g., relying on them, *M* = .78; tacking on traditional ones, *M* = 1.00) and several of the special types of positive consequences (e.g., informing someone out of the contact loop, *M* = 0.05; letting sender know of action, *M* = 0.75; acknowledging the bearer of bad news, *M* = 1.00, simple, *M* = 1.07) were behaviors that occurred least often.

For monitoring, scores ranged from 1.45 to 6.77 (raw point equivalents of 9 and 42, respectively), and the average was 3.82 (23.66 points). The mean score for monitoring via work sampling was 1.76 (7.37 points), with a low of zero to a high of 5.48 (23 points). This behavior was not very common. In fact, in examining the frequencies of scores, it appears that the majority of respondents monitored via work sampling one time or less on the entire in-basket – 60% scored 6 points or less. Nine people (approximately 26% of the sample) received no points for work sampling, meaning that they monitored completely via self-report. Monitoring via self-report occurred often -- the mean was the highest of all behaviors, 4.53 (16.29 points), and scores ranged fairly widely from 0.83 (3 points) to 7.50 (27 points).

For providing consequences, scores ranged from -1.22 (-10 raw points) to 4.02 (33 points), and the average was 1.38 (11.34 points). This overall score was comprised of the sum of raw points for positive and negative consequences, divided by the maximum possible points for all consequences (76). As described in the Method chapter, some subcategories were scored negatively (i.e., positive consequences that were not warranted, negative consequences that were questionably and not warranted), and the points for these were subtracted from the raw scores for positive or

negative consequences. (No points were deducted from the maximum possible points in the denominators.) Also, some aspects of the severity, frequency, and form of *warranted* negative consequences resulted in points being deducted (e.g., if statements were harsh or injudicious, delivered three or more times, or in the form of a pseudo-monitor). Thus, it was possible for respondents to get a negative score for positive and negative consequences, as well as consequences overall.

Considering the scores for those consequences that were positive (which consisted of points for those that were warranted minus the points for those that were not), the mean was 3.19 (16.57 raw points), and ranged from 0.77 to 6.15 (4 to 32 points). The mean for those positive consequences that were warranted only was thus slightly higher, at 3.53 (18.34 points). Of the special types of positive consequences identified, most respondents gave them once or not at all throughout the in-basket, so means for them were fairly low. For example, 22 of the sample of 35, or 63%, gave a simple one- or two-word positive consequence (such as “Thank you”) once or not at all, 83% did the same for acknowledging a bearer of bad news, and 100% for broadcasting to multiple parties. For a few of the special types, one respondent “stood out” with a high score – for instance, one person scored 7.33 (11 points) for giving simple one- or two-word positive consequences, and another scored 6.67 (12 points) for relaying problems to others non-judgmentally, which he/she did four times in the in-basket. In terms of *unwarranted* positive consequences, the majority of respondents did not give them, thus their scores were zero (57%). Scores ranged from zero to 5.71 (-8 points), and the mean was 1.27 (-1.77 points). (Because this subcategory was negatively valuated, scores were calculated by dividing raw points by the minimum possible, a negative value. For example, the score of 5.71 was achieved by dividing the raw score of -8 by -14 [the minimum score possible], and then multiplying by 10.)

Scores for those consequences that were negative overall (which was equal to the sum of the points for those that were warranted, questionably warranted, and not warranted) ranged from

-9.00 to 4.33 (-27 to 13 points), and the mean was -1.74 (-5.23 points). (This mean was negative because many respondents gave questionably warranted negative consequences, which essentially "wiped out" any points they received for those that were warranted.) For negative when warranted consequences only, there was a wide range of scores from -1.25 to 7.92 (-3 to 19 points), and the mean was 3.46 (8.31 points). For questionably warranted negatives, the range was 0.69 to 3.89, indicating that every respondent had points subtracted for this type of behavior – from 5 to 28. The mean score was 1.86, or -13.37 raw points. (Given that respondents gave a fair number of these behaviors, more attention should be paid to them in future research.) Negative consequences that were not warranted were rarely seen in the in-basket. In fact, only one respondent provided one that resulted in a deduction of six points (equivalent to a score of 3.75) from the raw negative consequence score. Thus, 97% of the sample had a score of zero, and the mean was 0.11 (-0.17 points). (Because there was so little variability in this behavior, it will not be discussed any further in the results.)

For providing antecedents, total scores ranged from -0.13 (-1 point) to 1.65 (13 points), with a mean of 0.72 (5.71 points). The amount of points respondents received for tacking on traditional antecedents was fairly low -- the mean was 1.00 (equivalent to 4 raw points) and ranged from zero to 2.50 (10 points). Only 11% of the sample did not receive any points for this type of antecedent, 23% scored 4 points, and one person (2.9%) each received the two highest values (9 and 10 points). For those antecedents given in an unclear situation without monitoring, the average score was 1.13 (2.03 points), the minimum score was zero, and the maximum was 4.44 (8 points). Many of the respondents (54%) did not exhibit this behavior, and received scores of zero. For antecedents that were chosen over consequences, the mean was -0.79 (-0.31 points) and scores ranged from -10.00 to 5.00 (-4 to 2 points). (Respondents could have points subtracted if they indicated a preference for consequences over antecedents.) Many of the respondents (49%) either

did not indicate a preference or had opposing preferences on the two items used to measure this aspect, so their scores were zero.

Two overall scores were also calculated using the categories of monitoring and providing consequences. The additive score, which was the sum of the calculated scores for monitoring and providing consequences, ranged from 1.69 to 9.94, and the mean was 5.20. Scores were spread out through the range (the median was 4.81). For the multiplicative score (calculated by taking the z score conversions of the raw scores for monitoring and providing consequences, adding five, and then multiplying the results) the mean was 25.31 and scores ranged from 12.98 to 47.25. Scores were also spread throughout the range (the median was 23.34).

Table 18

Summary of In-basket Scores

Category/ *Subcategory 1/* Subcategory 2/ Subcategory 3	Mean	SD	Minimum	Maximum
Monitoring	**3.82** **(23.66)**	**1.24** **(7.70)**	**1.45** **(9)**	**6.77** **(42)**
Work sample	1.76 (7.37)	1.41 (5.90)	0.00 (0)	5.48 (23)
Self-report	4.53 (16.29)	1.28 (4.62)	0.83 (3)	7.50 (27)
Providing consequences	**1.38** **(11.34)**	**1.27** **(10.45)**	**-1.22** **(-10)**	**4.02** **(33)**
Positive	3.19 (16.57)	1.18 (6.14)	0.77 (4)	6.15 (32)
When warranted	3.53 (18.34)	1.15 (6.00)	1.35 (7)	6.15 (32)
Simple	1.07 (1.60)	1.50 (2.25)	0.00 (0)	7.33 (11)
Letting sender know of action	0.75 (1.94)	0.84 (2.20)	0.00 (0)	3.08 (8)
Relaying problem non-judgmentally	1.52 (2.73)	1.68 (3.03)	0.00 (0)	6.67 (12)
Acknowledging bearer of bad news	1.00 (2.40)	1.21 (2.88)	0.00 (0)	5.00 (12)

Table 18, cont.

Category/ *Subcategory 1/* Subcategory 2/ Subcategory 3	Mean	SD	Minimum	Maximum
Informing someone out of contact loop	0.05 (0.11)	0.28 (0.68)	0.00 (0)	1.67 (4)
Broadcasting to multiple parties	1.29 (1.30)	2.22 (2.20)	0.00 (0)	5.00 (5)
When not warranted[a]	1.27 (-1.77)	1.58 (2.21)	0.00 (0)	5.71 (-8)
Negative	-1.74 (-5.23)	3.14 (9.42)	-9.00 (-27)	4.33 (13)
When warranted	3.46 (8.31)	2.22 (5.32)	-1.25 (-3)	7.92 (19)
When questionably warranted[a]	2.03 (-13.37)	0.88 (5.80)	0.76 (-5)	4.24 (-28)
When not warranted[a]	0.14 (-0.17)	0.85 (1.01)	0.00 (0)	5.00 (-6)
Providing antecedents	**0.72** **(5.71)**	**0.42** **(3.34)**	**-0.13** **(-1)**	**1.65** **(13)**
Tacking on traditional antecedents	1.00 (4.00)	0.65 (2.60)	0.00 (0)	2.50 (10)
Relying on antecedents	0.78 (1.71)	1.28 (2.82)	-1.36 (-3)	4.09 (9)
In unclear situations	1.13 (2.03)	1.45 (2.62)	0.00 (0)	4.44 (8)
In choosing between antecedents and consequences	-0.79 (-0.31)	3.58 (1.43)	-10.00 (-4)	5.00 (2)
Overall additive (monitoring + consequences)	***5.20***	***2.04***	***1.69***	***9.94***
Overall multiplicative (monitoring X consequences)	***25.31***	***8.44***	***12.98***	***47.25***

Note. N = 35. Numbers shown represent raw scores divided by points possible, multiplied by 10. (Higher values indicate that behaviors occurred more often.) Numbers in parentheses are raw scores.

[a] Subcategory was scored negatively, and the *minimum* points possible (a negative value) was used to calculate the score.

Relationships Among the Categories and Subcategories

To assess how the major categories and subcategories of each behavior were interrelated, Pearson correlations were calculated between them. Results are presented in Table 19. Monitoring overall was significantly related to providing consequences overall ($r = .32$, $p < .05$), indicating that those who gathered information about performance also communicated evaluations of performance. Of the subcategories of consequences, two were related to overall monitoring. The correlation for negative when warranted consequences and monitoring was .32 ($p < .05$), showing that those people who gathered information about performance also let people know about errors they had discovered in actual samples of performance. The correlation with positive when not warranted, $r = -.41$ ($p < .01$), meant that those who gathered information about performance more often were also likely less likely to give noncontingent positive consequences.

Providing consequences overall was also related to providing antecedents overall ($r = -.32$, $p < .05$). This is a negative relationship, indicating, for example, that respondents who had low scores for consequences had high scores for antecedents. This relationship seems to be due primarily to how antecedents given in unclear situations were scored. Antecedents given in unclear situations were scored in four items in which it was possible to receive points for questionably warranted negative consequences (which also lowered the overall consequence score because they were scored negatively). (Antecedents were scored only when the respondent did not monitor.) In fact, the correlation between antecedents given in unclear situations and questionably warranted negative consequences was .56 ($p < .01$). Interestingly, neither providing antecedents overall, nor any of the types of antecedents was related to providing *positive* consequences (rs = .02 for overall, -.02 for tacking on traditional, and .04 for relying on antecedents, $p > .05$).

Table 19

Intercorrelations Between Major Categories and Subcategories of Behavior in the OSIBA

Category/ *Subcategory 1/* Subcategory 2 Subcategory 3	1	1a	2	2a	2a1	2a2	2b	2b1	2b2	3	3a	3b	3b1	3b2
1. Monitoring	*.30*													
1a. Work sample	.80**	*.07*												
2. Consequences	.32*	.23	*.16*											
2a. Positive	.21	.13	.45**	*.30*										
2a1. Warranted	.07	-.02	.32*	.93**	*.41*									
2a2. Not warranted	-.41**	-.41**	-.38**	-.24	.12	*-.31*								
2b. Negative	.21	.17	.81**	-.15	-.25	-.27	*.32*							
2b1. Warranted	.32*	.31*	.72**	-.04	-.17	-.36*	.83**	*.37*						
2b2. Questionably warranted	-.07	.00	-.66**	.18	.22	.07	-.86**	-.44**	*-.05*					
3. Antecedents	-.16	.06	-.32*	.02	.03	.02	-.36*	-.10	.51**	*-.31*				
3a. Tacking on traditional	-.05	-.05	-.08	-.02	.01	.08	-.08	-.05	.08	.57**	*-.002*			
3b. Relying on antecedents	-.14	.12	-.30*	.04	.02	-.05	-.35*	-.08	.52**	.66**	-.24	*-.11*		
3b1. In unclear situations	-.23	.08	-.40**	-.03	-.02	.03	-.42**	-.15	.56**	.48**	-.33*	.86**	*.14*	
3b2. In choosing between	.14	.09	.14	.14	.08	-.15	.07	.13	.00	.42**	.12	.39**	-.12	*-.39*

Note. Coefficients in italics shown on the diagonal are Cronbach's alphas.

** $p <= .01$. * $p <= .05$

The correlations between overall categories and subcategories “within them” give an indication of the proportion of the overall score that is comprised of the particular subcategory. For example, the correlation between monitoring via work sampling and overall monitoring was .80 ($p < .01$), showing that work sampling accounted for a high percentage of the variability in the monitoring score. All subcategories of consequences were significantly related to the overall consequence score (*r*s ranged from -.66 to .81). All subcategories of antecedents were also related to antecedents overall (*r*s ranged from .42 to .66).

Other interesting relationships of note from Table 19 include those among subcategories for providing consequences and antecedents. As shown, positive consequences were not significantly related to negative consequences ($r = -.15$, $p > .05$), meaning that respondents likely to communicate negative evaluations of performance were not necessarily likely to communicate positive evaluations, and vice versa. However, providing negative when warranted consequences was related to providing positive when *not* warranted consequences, $r = -.36$ ($p < .05$). This result indicates that respondents who more often let people know about errors they had discovered were less likely to provide noncontingent positive consequences. The major subcategories of antecedents, tacking on traditional and relying on antecedents were not significantly related ($r = -.24$, $p > .05$). Antecedents given in unclear situations, however were negatively related to those that were coded as tacked-on traditional ($r = -.33$, $p < .05$). This seems counterintuitive, in that respondents who give many reminders, directives, and instructions when they were not needed or requested gave fewer antecedents in unclear situations and without monitoring (or vice versa). It may be helpful to study these behaviors further, perhaps defining them better.

Item Analyses

A reliability analysis was performed using respondents’ scores for categories (and some of the major subcategories) on each item in order to assess how consistently the categories were measured across the in-basket items. Of interest was whether respondents monitored performance,

for instance, across the items in which the behavior was scored. Because the items themselves presented respondents with different situations and the opportunities to demonstrate behaviors differed, it was not predicted that scores would be the same across items. As with other in-basket exercises (and also assessment centers), some items (or exercises, in the case of assessment centers) provide more of a chance for respondents to demonstrate particular behaviors (e.g., providing positive consequences) than other items and some items offered opportunities to demonstrate several behaviors, such as monitoring and providing negative or positive consequences (J. Smither, personal communication, April 7, 2006). (This is different than traditional tests or questionnaires in which items provide an equal chance for respondents to demonstrate a construct/behavior [e.g., knowledge of a particular subject, extraversion] or homogeneous tests or questionnaires in which a single dimension is measured. In these measures, it would be expected that items were correlated with one another.) Inter-item correlations (e.g., bivariate correlations among participants' monitoring scores on item numbers 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 15, and 16) and Cronbach's coefficient alpha (α), an internal consistency estimate, were calculated for each category and major subcategory of behavior. (Results are shown in Table 20.)

For monitoring, scores from the 12 items in which the behavior was pre-identified on the score sheets were included in the analysis. (Monitoring "not identified" was coded for a few respondents on four other items, but data from those items were not used in the analysis because of the small amount of variability in scores.) Most of the inter-item correlations were not statistically significant – only four were significant, and one was marginally. The pattern of results was not very clear. Many correlations were negative, indicating that respondents, in general, had high scores for monitoring on one item and low scores for it on the other. The inter-item correlations ranged widely between -.35 and .38. The mean of these correlations was .03, indicating that there was essentially no relationship between monitoring scores on the items, or that people who monitored on one item

were not necessarily likely to monitor on other items. The alpha coefficient was .30, which suggests that internal consistency was fairly low.

The alpha coefficient for monitoring via work sampling was even worse, at .07. Among the six items in which the behavior was scored, the mean inter-item correlation was .02. When the data for work sampling are considered, however, it is possible that these results are due to low frequency of response and lack of variability. For every one of the six work sampling items, more than 70% of the sample had a score of zero and, as mentioned previously, 60% performed the behavior one time or less on the entire in-basket.

Data for consequences in total were more varied on the 16 items in which some form was scored, but the alpha coefficient was very low (.16). Only 4 inter-item correlations were statistically significant, and the mean of these correlations was .003. Separate reliability analyses were also conducted for positive and negative consequences. Results for these subcategories were somewhat better, but again fairly low. Similar to the results for monitoring, most inter-item correlations for both types of consequences were not statistically significant. For positive consequences, seven were significant, and for negative consequences only one was. In addition, many were negative. They ranged between -.37 and .51 for positive consequences and -.26 and .41 for negative consequences, with means of .04 and .06, respectively. Coefficient alphas were .30 and .32, respectively.

Additional reliability analyses were conducted for consequences of different types of deservedness within positive and negative evaluations. Positive and negative consequences that were warranted had the highest coefficient alphas of all behaviors, though they were still fairly low -- .44 and .37, respectively. The intercorrelations between scores on the 16 items that were scored for warranted positive consequences ranged from -.35 to .51, and nine were significant. The average correlation was .05, however. For warranted negative consequences, scored on four items, the average of the inter-item correlations was .13, and they ranged between -.11 and .24. The results for positive consequences that were *not* warranted and negative consequences that were *questionably*

warranted seemed odd, in that the alphas were negative (-.31 and -.05). Cronbach's alpha is not usually negative, but is possible when there are mixed signs among the inter-item correlations (Thompson, 2003). This was the case for the questionably warranted negatives, but not so for the unwarranted positives (all inter-item correlations were negative). To summarize, similar to monitoring, it appears that respondents did not provide consequences of any type consequences very consistently across the items in which the behaviors were scored.

For providing antecedents in total, 12 of the 13 items in which some type of antecedents were designated on the score sheet were included in the analysis (no one received points for tacked-on traditional ones on one item). Inter-item correlations ranged from -.42 to .33, with a mean of -.03. Coefficient alpha was again observed to be negative (-.31). Separate analyses were also completed for the two major types of antecedents, tacked -on traditional antecedents and relying on antecedents. For the tacked-on traditional variety, inter-item correlations ranged from -.25 and .33, and many items were negatively correlated with each other (though not significantly). The mean of the inter-item correlations was -.03 and coefficient alpha was -.002. The scores in each of the items were checked and it was found that for several, very few respondents received points for antecedents (and variability was quite small). The reliability analysis was run again without three items in which less than five respondents received a score for providing antecedents. In this analysis, the mean of the inter-item correlations was .09 and coefficient alpha rose to .22.

Results for relying on antecedents were similarly not very clear. For the five items in which this behavior was measured, the mean inter-item correlation was -.03 and Cronbach's alpha was -.11. It was thought this result might have occurred because the two items in which respondents chose providing antecedents over consequences might be different than those in which antecedents were given in an unclear situation without monitoring. Reanalyzing the groups of items separately yielded a correlation between the two "choice" items of -.17 and Cronbah's alpha of -.39, and an

average inter-item correlation of .06 and alpha coefficient of .14 for the three "unclear situation" items.

The results of these analyses suggest that respondents did not consistently monitor, provide consequences, or provide antecedents across different items in the OSIBA. The pattern of correlations between scores for the same behavior over different items was not clear (nor always positive) – correlations ranged from moderately positive to moderately negative. In addition, the mixed signs of these correlations likely had a detrimental effect on the estimation of the internal consistency statistic, Cronbach's alpha. Low frequencies of some behaviors may have also impacted alphas. Interestingly, a similar pattern has been observed in the results for other in-basket measures (e.g., Brannick et al., 1989; Brass & Oldham, 1976; Hakstian & Scratchley, 1997).

Table 20

Summary of Reliability Analyses for Categories or Subcategories Measured in Different In-basket Items

Behavior Category or Subcategory	# Items	Mean Inter-item *r*	Minimum Inter-item *r*	Maximum Inter-item *r*	Coefficient α
Monitoring	12	.03	-.35	.38	.30
Via work sampling	6	.02	-.19	.29	.07
Consequences	16	.003	-.41	.44	.16
Positive	16	.04	-.37	.51	.30
Warranted	16	.05	-.35	.51	.44
Not Warranted	3	-.10	-.13	-.05	-.31
Negative	9	.06	-.26	.41	.32
Warranted	4	.13	-.11	.24	.37
Questionably warranted	8	-.003	-.33	.41	-.05
Antecedents	12	-.03	-.42	.33	-.31
Tacked-on traditional antecedents	7	-.03	-.25	.33	-.002
Relying on antecedents	5	-.03	-.17	.33	-.11

Checks for Differences in In-Basket Version and Demographic Variables

To assess whether differences in in-basket scores might be observed for respondents with different demographic characteristics or for those who had taken different versions of the in-basket, analyses of variance (ANOVAs) and correlational analyses were conducted. ANOVAs were

performed with groups who took different versions of the in-basket, held different job titles, were in different locations, or were of different gender, and the major categories of monitoring, providing consequences, and providing antecedents were separate dependent measures, as well as the overall additive and multiplicative scores. Refer to Table 21 for a summary of the results.

Version of OSIBA. Because of study procedures, different versions of the in-basket could not be distributed evenly among participants. (Versions were randomly mailed to all 53 potential participants and completing the exercise was voluntary.) Thus, the numbers of participants that took each version varied. Of the five versions, 6 took the first, 4 took the second, 7 took the third, 11 took the fourth, and 6 took the fifth. One person was excluded from this analysis because the version he/she completed could not be determined. Results of the ANOVAs showed that for groups who took different versions of the exercise, there were no differences in scores for monitoring, providing consequences, and providing antecedents, and in the overall additive and multiplicative scores (*F*s ranged from 0.03 to 2.01, $p > .05$). Essentially, this indicated that participants who took any particular version scored the same as those who took the other four (as would be expected).

Job Title. To improve the analyses for job title, the lone Assistant Vice President (AVP) was added to the Vice President (VP) group. There were, however, still fairly significant differences in the sizes of the groups (i.e., Managing Directors = 7, Directors = 9, Vice Presidents/Assistant Vice President = 14, and Associates = 4). Again, results showed no differences for monitoring, providing consequences, providing antecedents, and overall additive and multiplicative scores among people in different job titles (*F*s ranged from 0.35 to 1.75, $p > .05$). Thus, Managing Directors, for example, scored no differently overall or for any of the major categories on the in-basket than Directors, Vice Presidents, or Associates.

Location. Participants were also not evenly distributed among the three geographic regions (i.e., U.S. = 14, Europe = 16, Asia = 5). The ANOVAs yielded no differences in monitoring, providing consequences, providing antecedents, and overall additive and multiplicative scores for

the three regional groups (*F*s ranged from 1.02 to 2.87, $p > .05$). This means that respondents in the U.S., for example, did not monitor, provide consequences, or provide antecedents any more or less than those in Europe or Asia, nor did they score any differently on the in-basket overall.

Gender. As was the case with the other demographic variables analyzed, the group sizes for gender differed greatly (i.e., males = 29, females = 6). Results from the ANOVAs showed no differences between males and females in monitoring, providing consequences, providing antecedents, and in the overall additive and multiplicative scores (*F*s ranged from 0.00 to 2.21, $p > .05$). Thus, women monitored and provided consequences and antecedents about the same as men, and received essentially the same overall scores on the in-basket.

Table 21

Results from ANOVAs for Job Title, Location, and Gender with Several Dependent Measures of In-basket Scores

Dependent Measure	Source	*Df*	*F*
	Version of OSIBA		
Monitoring			
	Between-Groups	4 (0.06)	0.03
	Within-Groups	29 (1.76)	
Providing Consequences			
	Between-Groups	4 (0.76)	0.42
	Within-Groups	29 (1.80)	
Providing Antecedents			
	Between-Groups	3 (0.29)	2.01
	Within-Groups	29 (0.14)	
Overall Additive			
	Between-Groups	4 (1.02)	0.22
	Within-Groups	29 (4.73)	
Overall Multiplicative			
	Between-Groups	4 (14.06)	0.17
	Within-Groups	29 (80.96)	
	Job Title		
Monitoring			
	Between-Groups	3 (2.53)	1.75
	Within-Groups	31 (1.45)	
Providing Consequences			
	Between-Groups	3 (0.60)	0.35
	Within-Groups	31 (1.73)	

Table 21, cont.

Dependent Measure	Source	*Df*	*F*
	Job Title, cont.		
Providing Antecedents			
	Between-Groups	3 (0.10)	0.56
	Within-Groups	31 (0.19)	
Overall Additive			
	Between-Groups	3 (2.41)	0.56
	Within-Groups	31 (4.35)	
Overall Multiplicative			
	Between-Groups	3 (44.89)	0.61
	Within-Groups	31 (73.73)	
	Location		
Monitoring			
	Between-Groups	2 (1.57)	1.02
	Within-Groups	32 (1.54)	
Providing Consequences			
	Between-Groups	2 (4.20)	2.87
	Within-Groups	32 (1.46)	
Providing Antecedents			
	Between-Groups	2 (25.31)	2.18
	Within-Groups	32 (8.90)	
Overall Additive			
	Between-Groups	2 (8.85)	2.28
	Within-Groups	32 (3.88)	
Overall Multiplicative			
	Between-Groups	2 (130.15)	1.93
	Within-Groups	32 (67.50)	
	Gender		
Monitoring			
	Between-Groups	1 (0.00)	0.00
	Within-Groups	33 (1.59)	
Providing Consequences			
	Between-Groups	1 (3.47)	2.21
	Within-Groups	33 (1.57)	
Providing Antecedents			
	Between-Groups	1 (0.25)	1.39
	Within-Groups	33 (0.18)	
Overall Additive			
	Between-Groups	1 (3.45)	0.82
	Within-Groups	33 (4.20)	
Overall Multiplicative			
	Between-Groups	1 (57.18)	0.38
	Within-Groups	33 (71.61)	

Note. ANOVA is the abbreviation for analysis of variance. Values enclosed in parentheses represent mean squares.

Bivariate Pearson correlations were computed between the in-basket scores (i.e., monitoring, providing consequences, providing antecedents, overall additive score, and overall multiplicative score) and the remainder of the demographic variables assessed, including age, education, tenure in the organization, tenure in the position, supervisory experience, and days of supervisory training. See Table 22 for a summary of the results.

Age. The age of respondents was not significantly related to scores for monitoring, providing consequences, or providing antecedents, nor to the overall additive and multiplicative scores (rs ranged from -.04 to .19, $p > .05$).

Education. The amount of respondents' education in years was also not significantly related to any of the behaviors or overall scores (rs ranged from -.22 to .17, $p > .05$).

Tenure in organization. The number of years that respondents reported working in the organization was significantly correlated with their scores for monitoring ($r = -.42$, $p < .05$). This result means that the longer a person was employed at the bank, the less he/she monitored on the in-basket. No other scores were significantly related to tenure in the organization (rs ranged from -.23 to .19, $p > .05$).

Tenure in current position. The years respondents were in their position was not significantly related to any of the scores for monitoring, providing consequences, or providing antecedents, or to the overall additive and multiplicative scores (rs ranged from -.28 to -.04, $p > .05$).

Supervisory experience. The number of respondents' years as a supervisor was also not significantly related to any of the behaviors or overall scores (rs ranged from -.10 to .25, $p > .05$).

Supervisory training. Similar to the many of the other demographic variables, the number of days of supervisory training reported by respondents was not significantly correlated with scores for the behaviors of monitoring, providing consequences, providing antecedents, or with overall additive or multiplicative scores (rs ranged from -.18 to .09, $p > .05$).

In sum, it does not appear that respondents who took different versions of the in-basket or those with different demographic characteristics responded any differently on the measure, with one exception. There was a significant negative correlation between monitoring scores on the in-basket and the number of years respondents reported working in the organization ($r = -.42$, $p < .05$), indicating that the longer a person was employed at the bank, the less he/she monitored. The lack of significant differences (though none were expected) might have been a result of the lack of consistency in scores among the behaviors measured in the in-basket items (which could have inflated error variance in the difference scores between groups) (J. Lefkowitz, personal communication, March 29, 2006).

Table 22

Correlations Between Selected Demographic Variables and In-basket Scores

Demographic Variable	Monitoring	Providing Consequences	Providing Antecedents	Overall Additive	Overall Multiplicative
Age	-.25[b]	.19[b]	-.05[b]	-.04[b]	-.04[b]
Education	-.22	-.08	.17	-.18	-.18
Tenure in organization	-.42*	-.06	.19	-.22	-.23
Tenure in position	-.28	-.04	-.02	-.20	-.19
Supervisory experience	-.08[b]	.25[b]	-.10[b]	.11[b]	.08[b]
Supervisory training[a]	-.18[b]	-.02[b]	.09[b]	-.08[b]	-.07[b]

Note. Except where indicated, the unit of measurement is years and $n = 35$.

[a] Unit of measurement is days. [b] $n = 34$.

* $p < .05$, two-tailed.

Psychometric Properties of the Criterion and Technical Skill Measure

The ratings provided by the 35 people who returned the in-basket were included in all the statistical analyses. (Ratings from five people who consented but did not complete the in-basket were not used.) All 52 people on the original list of potential participants (and one extra person who was added after the project began) were rated by at least one person in the group of 35. In fact, each

person received from 1 to 27 ratings, 10.0 effectiveness ratings on average and 11.1 technical skill ratings on average.

For both measures, the minimum rating was one (equivalent to the "grade" of *F*) and maximum rating was five (equivalent to *A*). The overall means for the group that completed the in-basket (35 people) were 3.86 ($n = 343$, $SD = 0.97$) for motivational effectiveness, and 4.25 for technical skill ($n = 379$, $SD = 0.83$). (Considering the ratings given for all 53 in the group, the means were 3.76 [$N = 529$, $SD = 1.04$] for motivational effectiveness, and 4.15 [$N = 585$, $SD = 0.91$] for technical skill.)

Alternate Calculations of Motivational Effectiveness

Two additional, exploratory measures of motivational effectiveness were calculated, which utilized the ranking information collected. These were described in detail in Chapter 7, in the section about the criterion. The range for the calculation that yielded gradations in the numeric translation of the letter grade ratings based on rank ("gradation" measure) was 0.2 to 5.0. The overall mean for the entire group on this measure was 3.42 ($N = 529$, $SD = 1.08$). The range for the second score, which was calculated by multiplying the numeric grade translation by the overall percentile for each person rated ("percentile" measure), was 0.01 to 5.00. The mean for the entire group was 2.13 ($N = 529$, $SD = 1.58$).

These alternate calculations and the "regular" measure of motivational effectiveness (the simple average of the numeric grade ratings) were very highly correlated. The Pearson correlation between the "regular" and "gradation" measures was .95 ($N = 53$, $p < .01$), and .86 ($N = 53$, $p < .01$) for the "regular" and "percentile" measures. The correlation between the two alternate calculations was .92 ($N = 53$, $p < .01$). Because the "regular" and alternate measures were so highly correlated and no major differences in results were found, the simple average of all ratings received was chosen to be used as the criterion.

Relationship between Criterion and Technical Skill Measures

To assess how related the criterion of motivational effectiveness was to the orthogonal measure of technical skill, a Pearson correlation was computed. Motivational effectiveness was moderately and significantly related to technical skill -- the correlation was .44 ($N = 53$, $p < .01$). Thus, the two measures were not statistically orthogonal -- there was some relationship between the two scores. In other words, if a banker were rated high in motivational skill, he/she would likely have a high score for technical skill.

The correlation between the two was not anticipated to be high because people who are good at motivating others might not be technically skilled, and vice versa. The relationship found, however, might, at least in part, be attributable to halo error “committed” by the raters or to common method variance from the same rating source using the same response format.

Rater Perspectives

Because one-item measures were used for motivational effectiveness and technical skill, formal assessment of reliability was not possible. As discussed in the chapter on the method, the rating form had a space to note if the respondent was or had been a subordinate of each person he/she rated. Thus, two rating sources were identified in the rating data: subordinates or supervisees and others (which could have included supervisors, peers, or others who might have worked on deal teams with a particular person but were not peers). This information was used to give a rough indication of agreement between the sources. (Not all participants formally supervised employees and rating data were received from subordinates for 20 of the 35 participants.)

Pearson correlations were computed between the two sources for both motivational effectiveness and technical skill ratings. Results are presented in Table 23. Ratings provided by subordinates and former subordinates were similar to those given by others. Subordinates’ ratings of motivational effectiveness and technical skill were significantly correlated with ratings provided by other people, *r*s were 0.55 and 0.68 ($p < .01$), respectively. These correlations were somewhat

higher than what has generally been reported in the literature on multisource performance ratings and 360° evaluation feedback. In one study, for example, low but significant relationships were found between average subordinate and supervisor ratings (Adsit, Crom, Jones, & London, 1994). The correlation between overall performance as rated by subordinates on an upward feedback instrument with an overall performance appraisal rating from a supervisor (which was assigned partially on the basis of a forced distribution) was .14 ($p < .05$). A meta-analysis of correlations between rating source showed mean correlations of .22 for subordinate and supervisor ratings (K [number of coefficients analyzed] = 22) and subordinate and peer ratings ($K = 14$) (Conway & Huffcutt, 1997).

Table 23

Correlations among Motivational Effectiveness and Technical Skill Scores Provided by Various Sources

Source	1	2	3
	Motivational Effectiveness		
1. Subordinates	-		
2. Others	.55[a]**	-	
3. Self	.21[b]	.48[c]**	-
	Technical Skill		
1. Subordinates	-		
2. Others	.68[a]**	-	
3. Self	-.15[b]	.31[c]	-

[a] $n = 30$. [b] $n = 20$. [c] $n = 35$.

** $p < .01$, two-tailed. * $p = .01$, two-tailed.

Respondents were asked to rate their own motivational effectiveness and technical skill (on the same A through F scales) upon completing the in-basket. These self-ratings were not included in the computation of the average rating used in the analyses. Pearson correlations, however, were computed between them and the ratings from other sources to see how similar bankers' views of themselves were to their co-workers'. As shown in Table 23, self-ratings of motivational effectiveness were significantly correlated with others' ratings ($r = .48$, $p < .01$), but not with

subordinates' (r = .21, p > 05). Self-ratings of technical skill were not significantly correlated with those from other sources (r = -.15 with subordinates and r = .31 with others, p > 05). With the exception of the one significant correlation with others' ratings of motivational effectiveness, the results were more in line with those found in other studies. Conway and Huffcutt's (1997) meta-analysis showed mean correlations of .22 (K = 50), .19 (K = 17), and .14 (K = 19) between self-ratings and supervisor, peer, and subordinate ratings, respectively. Even lower correlations were found in a more recent study -- .15, -.03, and -.08 for self- and supervisor, peer, and subordinate ratings (Atkins & Wood, 2002).

Checks for Differences in Context of Observation

Study participants were asked to provide ratings of motivational effectiveness and technical skill only for people in the group whom they had directly observed in action. They were asked to mark which of three contexts they had actually observed each person they were rating: supervising in his/her role as a manager or supervisor; creating or helping to create deal-specific teams within or outside of the group; or supervising informally, perhaps providing coaching. More than one context could be checked.

To assess whether ratings might be different depending upon the context in which the ratee was observed, each rating was assigned to one of five arbitrary groups based upon the context(s) noted: (1) supervising in his/her role as manager/supervisory only, (2) creating or helping to create deal-specific teams only, (3) supervising informally only, (4) any two contexts, or (5) all three contexts. All ratings submitted by the 35 study participants were used in this analysis (ratings were given for all 53 of the potential participants in the group). Sample sizes (the number of ratings) for each observational context group were somewhat uneven. For motivational effectiveness, the numbers of ratings in each group were as follows: (1) 44, (2) 131, (3) 59, (4) 128, and (5) 167; for technical skill (1) 44, (2) 143, (3) 58, (4) 130, and (5) 169. The average for all 529 motivational effectiveness ratings was 3.76 and mean scores ranged between 3.50 and 3.91 for the observational

context groups. Means for each group were: (1) 3.50, (2) 3.63, (3) 3.88, (4) 3.73, and (5) 3.91. For technical expertise ratings, the mean over all 544 ratings was 4.15, and means for the groups ranged from 4.01 to 4.30. The mean technical expertise scores for each group were: (1) 4.05, (2) 4.01, (3) 4.19, (4) 4.14, and (5) 4.30. ANOVAs were performed for the five groups with motivational effectiveness and technical skill as separate dependent variables. Results are shown in Table 24. The F statistics for both ANOVAs were not significant ($F = 2.26$, $p = .06$ for motivational effectiveness; $F = 2.15$, $p = .07$ for technical skill). Essentially, these results indicated that mean ratings for groups based on observational context did not significantly differ.

Table 24

Results from ANOVAs for Observational Context Groups with Dependent Measures of Motivational Effectiveness and Technical Skill

Dependent Measure	Source	*Df*	*F*
Motivational Effectiveness			
	Between-Groups	4 (2.45)	2.26
	Within-Groups	524 (1.08)	
Technical Skill			
	Between-Groups	4 (1.78)	2.15
	Within-Groups	539 (0.83)	

Note. ANOVA is the abbreviation for analysis of variance. Values enclosed in parentheses represent mean squares.

Checks for Differences in Demographic Variables

Analyses similar to those to assess differences in predictor scores for the various demographic variables were performed for the motivational effectiveness and technical knowledge ratings. ANOVAs were performed for job title, location, and gender groups with motivational effectiveness and technical skill scores as dependent measures and Pearson correlations were computed for the ratings with age, education, tenure in the organization, tenure in the position, supervisory experience, and supervisory training. (The same imbalances in group sizes were present for these ANOVAs.) Refer to Table 25 for a summary of results.

Job Title. Results from the ANOVAs showed no differences among job title groups for motivational effectiveness ($F(3, 31) = 2.59, p > .05$) and technical skill ratings ($F(3, 31) = 1.33, p > .05$).

Location. No differences were found in either measure between the three regional areas of U.S., Europe, and Asia ($F(2, 32) = 1.55, p > .05$ for motivational effectiveness and $F(2, 32) = 1.28, p > .05$ for technical skill).

Table 25

Results from ANOVAs for Job Title, Location, and Gender with Dependent Measures of Motivational Effectiveness and Technical Skill

Dependent Measure	Source	*Df*	*F*
	Job Title		
Motivational Effectiveness			
	Between-Groups	3 (0.46)	2.59
	Within-Groups	31 (0.18)	
Technical Skill			
	Between-Groups	3 (0.43)	1.33
	Within-Groups	31 (0.32)	
	Location		
Motivational Effectiveness			
	Between-Groups	2 (0.30)	1.55
	Within-Groups	32 (0.20)	
Technical Skill			
	Between-Groups	2 (0.42)	1.28
	Within-Groups	32 (0.33)	
	Gender		
Motivational Effectiveness			
	Between-Groups	1 (0.02)	0.11
	Within-Groups	33 (0.21)	
Technical Skill			
	Between-Groups	1 (1.50)	5.05*
	Within-Groups	33 (0.30)	

Note. ANOVA is the abbreviation for analysis of variance. Values enclosed in parentheses represent mean squares.

** $p < .01$. * $p < .05$

Gender. Results indicated no differences between motivational effectiveness scores for males and females, F (1, 33) = 0.11, $p > .05$. There was, however, a significant difference for technical skill scores, F (1, 33) = 5.05, $p < .05$. The men in the group received higher technical skill ratings than women, with the average score for men being 4.22 (n = 29, SD = 0.49) and 3.68 (n = 6, SD = 0.78) for women.

For the rest of the demographic variables, bivariate Pearson correlations were computed between them and the motivational effectiveness and technical skill measures. See Table 26 for a summary of these results.

Age. Age was negligibly (and not significantly) correlated with both motivational effectiveness and technical skill ratings ($r = .07$, $p > .05$ and $r = -.03$, $p > .05$, respectively).

Education. Motivational effectiveness ratings were not significantly correlated with respondents' amount of education, $r = .18$, $p > .05$. There was, however, a somewhat counterintuitive significant correlation for technical skill. Years of education was negatively related to technical skill ratings, $r = -.37$, $p < .05$. In other words, the more education a banker had, the lower his/her technical skills were rated.

Tenure in the organization. The years respondents have worked in the organization was not significantly related to either measure ($r = .06$, $p > .05$ for motivational effectiveness and $r = -.06$, $p > .05$ for technical skill).

Tenure in current position. Correlations between the years in position and the two measures were not significant ($r = .14$, $p > .05$ for motivational effectiveness and $r = .27$, $p > .05$ for technical skill).

Supervisory experience. The correlation between years of supervisory experience and motivational effectiveness was not significant ($r = .27$, $p > .05$), nor was the one with technical skill ($r = .12$, $p > .05$).

Supervisory training. A significant correlation was found between the amount of supervisory training and the motivational effectiveness rating, $r = .35$, $p < .05$, indicating, in general, that those who had more supervisory training were rated higher in terms of motivating others than those who had fewer days of training. The correlation between technical skill and days of supervisory training approached significance ($r = .33$, $p = .06$).

In sum, for motivational effectiveness, none of the participants' demographic characteristics significantly "explained" the variability in ratings, with the exception of days of supervisory training. In general, those bankers who had more supervisory training were rated higher in effectiveness at motivating others. For technical skill, gender and amount of education made a difference. Men were rated as more technically savvy than women and those with more education were rated lower than those with less education. (These results may have occurred due to unreliability of the criteria, each of which consisted of only one item.)

Table 26

Correlations Between Selected Demographic Variables and Ratings of Motivational Effectiveness and Technical Skill

Demographic Variable	Motivational Effectiveness	Technical Skill
Age	.07[b]	-.03[b]
Education	.18[c]	-.37[c]*
Tenure in organization	.06[c]	-.06[c]
Tenure in position	.14[c]	.27[c]
Supervisory experience	.27[b]	.12[b]
Supervisory training[a]	.35[b]*	.33[b]

Note. Except where indicated, the unit of measurement is years.
[a] Unit of measurement is days. [b] $n = 34$. [c] $n = 35$.
* $p < .05$, two-tailed.

Hypothesis Testing

Hypotheses 1a and 1b

To test hypotheses 1a and 1b, zero-order Pearson correlations were computed between the categories and subcategories of supervisory behavior assessed in the in-basket, motivational

effectiveness, and technical knowledge. Hypothesis 1a was partially supported and Hypothesis 1b was supported (Table 27). A series of multiple regressions was also performed in which the categories and subcategories were entered as predictors and ratings of motivational effectiveness and technical skill as criteria.

Table 27

Correlations Between In-basket Category and Subcategory Scores and Ratings of Motivational Effectiveness and Technical Skill

Category/ *Subcategory 1/* Subcategory 2 Subcategory 3	Motivational Effectiveness	Technical Skill
Monitoring (overall)	**.30***	**.11**
Work sample	.28*	.08
Self-report	.13	.08
Providing Consequences (questionably and not warranted subtracted)	**.23**	**.00**
Providing Consequences (only warranted)	**.29***	**.04**
Positive	.36*	.14
When warranted	.33*	.11
Simple	.06	-.19
Letting sender know of action	-.11	-.27
Relaying problem non-judgmentally	.26	-.05
Acknowledging bearer of bad news	.40**	.25
Informing someone out of contact loop	.07	.00
Broadcasting to multiple parties	.14	-.10
When not warranted[a]	-.10	-.10
Negative	.02	-.10
When warranted	.03	-.07
When questionably warranted[a]	-.06	.11
Providing Antecedents (overall)	**-.02**	**-.29**
Tacking on traditional antecedents	.06	-.44***
Relying on antecedents	-.08	.07
In unclear situations	-.04	.14
In choosing between antecedents and consequences	-.08	-.11

[a] Subcategory was scored negatively.

*** $p < .01$ (2-tailed), ** $p <= .01$ (1-tailed), * $p <= .05$ (1-tailed).

Hypothesis 1a: Relationships with Supervisory Effectiveness

Overall monitoring ($r = .30$, $p < .05$) and one type of monitoring in particular, via work sampling ($r = .28$, $p < .05$), were significantly correlated with motivational effectiveness. (Refer to Table 27.) This means that those rated most highly monitored, or gathered performance information. They gathered information in a particular way, as well. Instead of relying on reports of performance from secondary sources, highly rated managers requested to inspect work products or directly observe employees working.

To further illustrate these correlational results, the group of 35 participants was ranked in terms of their motivational effectiveness scores, and in-basket responses for the top and bottom thirds were examined. Contrasting responses to an in-basket item from participants who were rated in the top and bottom thirds of the group for motivational effectiveness are shown in Table 28. The item is from a manager describing that his group is not progressing on a project that had been successfully completed before. A banker who was highly rated in terms of motivational effectiveness suggested the manager monitor to discover what is going on in the team: "Please talk with each one of your team members and try to find out what their issues are." In contrast, a participant who was rated in the bottom third of the group suggested the group meet, not to gather information, but to revisit the timeline and get everyone's commitment to meeting their assigned goals.

Table 29 shows the difference in monitoring via work sampling for top and bottom rated bankers in a different in-basket item. In the item, the Director of the Art Department recommended that the magazine revert back to using the previous version of a reimbursement form because a contracted photographer had difficulty completing the new one. The highly rated banker wanted to gather information by taking a look at both the old and new forms (a product of the Accounting Department's work), which is monitoring via work sampling. He/she wrote, "Please let me have a copy of the old and the new reimbursement forms (see the complaint from Rosso). I want to discuss

Table 28

Monitoring in Responses to an OSIBA Item from Top and Bottom Rated Motivators

Item 2	
TO: You FROM: the desk of L. Rivera, Mgr., Mktg./Circ., FOODSTUFF DIGEST	
I just don't understand my group. I worked with them to create a timeline and milestones, with a final deadline in plenty of time. They just aren't making any progress! I've heard some complaints -- someone isn't doing their part and they're having disagreements over which clients to target. This resistance isn't like my group -- we've done promos many times before and they were successful! Can you give me some advice – what should I do about this situation? Thanks, Luis	
Responses	
Top Rated Banker	**Bottom Rated Banker**
There must be something wrong. Maybe no one wants to cover a particular client. *Please talk with each one of your team members and try to find out what their issues are*.	Perhaps a staff meeting -- go back and re-examine the timeline -- but be very clear to individual responsibilities and get each person's buy-in.

Note. Monitoring is shown in italics.

Table 29

Monitoring via Work Sampling in Responses to an OSIBA Item from Top and Bottom Rated Motivators

Item 3
TO: You FROM: J. Rosso, Dir., Art SUBJECT: Accounting reimbursement forms
Here we finally get Maria to do the shoot, but we can't even get her reimbursed. The reason: She didn't fill the form out right! I got a copy of the new form. Now I understand why she couldn't fill it out. It's a mess! I think that they ought to go back to the old form. I'm sure you'd agree. P.S. Don't worry. After several phone calls and begging, I finally got Accounting to send Maria a check.

Table 29, cont.

Responses	
Top Rated Banker	**Bottom Rated Banker**
To Accounting Dir.: [Fwd message] *Please let me have a copy of the old and the new reimbursement forms* (see the complaint from Rosso). I want to discuss with you once I have a chance to review them.	To Acctg. Dir.: Perhaps you could take some time to talk the Art Dept. through these new forms. I know they hate every form they're faced with, but let's see if we can help them come to terms with a change in process. To Rosso: I've had a word with Accounting and they're going to give your department a quick review of the new forms. That should resolve any issues going forward.

Note. Monitoring via work sampling is shown in italics.

with you once I have a chance to review them." The lower-rated banker simply asked Accounting to provide training or instruction on how to fill out the form, attributing the photographer's difficulty and the Art Director's complaint to resistance to change. He/she did not attempt to find out if the forms really were "a mess."

Providing consequences that were warranted (warranted only) was positively and significantly correlated with motivational effectiveness ($r = .29$, $p < .05$). (Refer to Table 27.) When questionably and not warranted types were subtracted from this consequence score (the overall score was used), however, the correlation was slightly lower and not significant ($r = .23$, $p > .05$).

One type of consequences in particular, positive consequences, was significantly related to effectiveness ($r = .36$, $p < .05$), as was a special sub-type of positive consequences. Respondents' scores for providing consequences in which they thanked the bearer of bad news were related to their ratings of motivational effectiveness ($r = .40$, $p < .01$). Thus, bankers who acknowledged people for drawing attention to unwanted or undesirable information in the in-basket were rated higher in motivating others than those who did not provide that type of positive consequence.

Contrasting responses for positive consequences are shown in Table 30. In the item shown, an employee raised his concerns about the quality of images being printed and the production workers' lack of concern for the problem. The higher rated banker provided a positive consequence,

acknowledging the bearer of bad news, by writing, “Thanks for your advice.” to the sender. The lower rated banker addressed the sender’s concerns, but did not acknowledge the sender for drawing attention to the problem.

Table 30

Providing Positive Consequences by Acknowledging the Bearer of Bad News in Responses to an OSIBA Item from Top and Bottom Rated Motivators

Item 15	
TO: You FROM: Mr. Puck, Art Department	
I’m concerned with the quality of the images that are coming off Press 1 – it’s really very poor. I’ve talked with Bobby Flay and his print shop workers several times about the importance of quality, but no one seems to care enough to do anything. Would you be willing to speak with the group also? Maybe then they’ll care.	
Responses	
Top Rated Banker	**Bottom Rated Banker**
Thanks for your advice. Please proceed and show me the deliverables. I need more info before I can talk to Bobby. As it's quite urgent, I would appreciate if you can drop by my office this afternoon. *Thanks*.	I will speak with the group. Is it possible that there is a mechanical problem?

Note. Acknowledging the bearer of bad news is shown in italics.

In sum, as hypothesized, monitoring and providing consequences (defined as warranted only) were positively and significantly related to motivational effectiveness. A subcategory of providing consequences – communicating positive evaluations – was also found to be correlated significantly with effectiveness.

Because zero-order correlations do not show the relationship between the behaviors and effectiveness while holding other behaviors constant, several multiple regressions were also performed. Total or overall scores for monitoring and providing consequences were predictors in the first regression analysis. So that the relationships between the criteria and various aspects, types, or subcategories of all behaviors measured in the in-basket could be assessed, different combinations of subcategories were used as predictors in two additional multiple regressions. In all

analyses, partial correlations (which are indices of the relationship between predictors and the criterion while controlling for other predictors) between the categories and subcategories of supervisory behavior, and motivational effectiveness were computed. The results, in general, did not confirm the correlational analyses.

As shown in Table 27, overall monitoring was significantly related to motivational effectiveness (zero-order $r = .30$, $p < .05$). However, in the first multiple regression (Model 1 in Table 31), the partial correlation, which reflects the relationship between monitoring and effectiveness while controlling for consequences, was lower and not significant ($r = .24$, $p = .08$). As mentioned previously, providing consequences (calculated by subtracting questionably and not warranted types) was not significantly related to motivational effectiveness (zero-order $r = .23$, $p > .05$, partial $r = .15$, $p > .05$). The multiple correlation was not significant ($R = .33$, $p > .05$) and neither of the beta weights for these predictors in the multiple regression equation was statistically significantly different from zero, indicating that together, they were not useful predictors of effectiveness.

An alternative regression model was also tested using overall monitoring and consequences defined as those that were warranted only as predictors (Model 1a in Table 31). Though the zero-order correlation between warranted-only consequences and motivational effectiveness was significant ($r = .29$, $p < .05$), the partial correlation was not ($r = .23$, $p > .05$). The multiple correlation was higher at .37, but not significant. Neither of the beta weights was significant.

Two additional multiple regressions were conducted using various types of monitoring, providing consequences, and antecedents as predictors to assess their relationships with effectiveness (Models 2 and 3 in Table 31). (The results of these should be viewed with caution, however, because using such large numbers of predictors with a relatively small sample probably resulted in very low power to detect results.) In Model 2 (with nine predictors), providing positive consequences when warranted, negative consequences that were not warranted, and monitoring via

Table 31

Multiple Regression and Correlation Results Using Categories and Subcategories as Predictors and Ratings of Motivational Effectiveness and Technical Skill as Criteria

Predictors	Motivational Effectiveness					Technical Skill				
	Zero-order *r*	Partial *r*	*ß*	*t*	*R*	Zero-order *r*	Partial *r*	*ß*	*t*	*R*
Model 1					.33					.12
Monitoring (overall)	.30*	.24	0.25	1.42		.11	.12	0.13	0.68	
Consequences (questionably and not warranted subtracted)	.23	.15	0.15	0.86		-.004	-.04	-0.04	-0.24	
Model 1a					.37					.11
Monitoring (overall)	.30*	.23	0.23	1.35		.11	.10	0.11	0.59	
Consequences (warranted only)	.29*	.23	0.23	1.32		.04	.01	0.01	0.06	
Model 2					.54					.52
Monitoring via work sampling	.28*	.25	0.25	1.30		.08	.09	0.09	0.46	
Monitoring via self-report	.13	.07	0.07	0.34		.08	.00	0.002	0.01	
Positive consequences when warranted	.33*	.32*	0.31	1.70		.11	.11	0.10	0.53	
Positive consequences when not warranted[a]	-.10	-.12	-0.12	-0.59		-.10	-.04	-0.04	-0.20	
Negative consequences when warranted	.03	-.13	-0.14	-0.67		-.07	-.03	-0.03	-0.13	
Negative consequences when questionably warranted[a]	-.06	-.17	-0.20	-0.84		.11	.18	0.23	0.93	
Negative consequences when not warranted[a]	.33*	.29	0.27	1.49		-.10	-.14	-0.13	-0.72	
Tacking on traditional antecedents	.06	.11	0.10	0.56		-.44**	-.48**	-0.51	-2.73**	
Relying on antecedents	-.08	.04	0.05	0.22		.07	-.16	-0.20	-0.82	

Table 31, cont.

Predictors	Motivational Effectiveness					Technical Skill				
	Zero-order *r*	Partial *r*	*ß*	*t*	*R*	Zero-order *r*	Partial *r*	*ß*	*t*	*R*
Model 3					.68					.68
Monitoring via work sampling	.28*	.25	0.26	1.08		.08	.04	0.04	0.15	
Monitoring via self-report	.13	.24	0.25	1.03		.08	.06	0.06	0.25	
Simple positive consequence	.06	.01	0.01	0.04		-.19	.05	0.05	0.22	
Letting sender know of action	-.11	.09	0.08	0.39		-.27	-.36	-0.34	-1.62	
Relaying problem non-judgmentally	.26	.25	0.25	1.09		-.05	-.07	-0.07	-0.31	
Acknowledging bearer of bad news	.40**	.29	0.40	1.28		.25	.11	0.15	0.47	
Informing someone out of the contact loop	.07	.31	0.27	1.38		.00	.03	0.03	0.13	
Broadcasting to multiple parties	.14	.12	0.11	0.50		-.10	-.01	-0.01	-0.06	
Other positive consequences	-.04	-.01	-0.01	-0.05		.18	.24	0.25	1.06	
Positive consequences when not warranted[a]	-.10	-.18	-0.20	-0.78		-.10	-.06	-0.07	-0.26	
Negative consequences when warranted	.03	-.30	-0.31	-1.32		-.07	-.05	-0.05	-0.22	
Negative consequences when questionably warranted[a]	-.06	-.32	-0.42	-1.41		.11	.08	0.10	0.32	
Negative consequences when not warranted[a]	.33*	.15	0.19	0.75		-.12	-.16	-0.21	-0.70	
Tacking on traditional antecedents	.06	.29	0.31	1.30		-.44**	-.43*	-0.49	-2.00	
Relying on antecedents in unclear situations	-.04	.32	0.46	1.43		.14	-.08	-0.11	-0.35	
Relying on antecedents in choosing between antecedents and consequences	-.08	-.13	-0.11	-0.55		-.11	-.04	-0.04	-0.18	

[a] Subcategory was scored negatively.

** $p <= .01$, * $p <= .05$.

work sampling were found to have significant zero-order correlations with effectiveness (*r*s equal to .33, .33, and .28 respectively, $p < .05$). (Providing unwarranted negative consequences was included in the analysis because it was a component of both negative and overall consequences, but the results for this subcategory in particular were disregarded because of limited variability. Only one person scored points for this behavior on one item.) The partial correlation for positive when warranted consequences was statistically significant ($r = .32$, $p < .05$), showing that highly rated bankers were likely to communicate positive evaluations of performance, provide positive feedback, or recognize effort/excellence when merited (while holding various types of monitoring, negative consequences, and antecedents constant). For monitoring via work sampling, the partial correlation was not significant at the .05 probability level ($r = .25$, $p = .10$). This result could mean that monitoring via work sampling is duplicative of other predictors in the analysis or reflecting method variance in common (J. Lefkowitz, personal communication, March 29, 2006). Again, none of the beta weights for the variables in this regression was significant. The largest beta weight, for positive when warranted consequences, was .31, $t = 1.70$, $p = .10$.

Another multiple regression (Model 3) was completed with a similar set of predictors, except that two behaviors were “broken down” further into their component parts. Positive consequences that were warranted was replaced by seven predictors (six special types and other) and relying on antecedents was replaced by two (relying in unclear situations and in choosing between antecedents and consequences), yielding a total of 16 predictors in the regression. One special type of positive when warranted consequence, acknowledging the bearer of bad news, had a fairly high and significant zero-order correlation with effectiveness ($r = .40$, $p < .01$), but the partial correlation for this behavior was not significant ($r = .29$, $p > .05$). In addition, not surprisingly, none of the beta weights in the regression equation was significant.

Hypothesis 1b: Relationships with Technical Knowledge

Table 27 shows that Hypothesis 1b was supported. Technical skill ratings were *not* related to monitoring overall, nor to either of the calculations for consequences in total (nor to any particular method of monitoring or type of consequence). In addition, a second set of multiple regressions were conducted using the same predictors as discussed previously, but with technical skill ratings as the criterion. As shown in Table 31, none of the partial correlations or beta weights for monitoring or providing consequences was significant.

An unanticipated and interesting finding in the results was that providing tacked-on traditional antecedents was fairly strongly related to technical skill. The zero-order correlation between the two was -.44 ($p < .01$). This means that those bankers high in technical skill provided few directives, instructions, or reminders when they were not needed or requested.

Table 32 shows contrasting responses to an in-basket item. In the item displayed, the Director of Accounting submitted a report on production costs and how much was spent on last-minute alterations over a five-month period. A banker who was in the bottom third for ratings of technical expertise noted the improvement in December and gave antecedents to the entire team: “Please note the costs due to last minute changes. As mentioned to you all in a previous email, it is everyone's responsibility to make sure the information is correct and that no room is left for last minute changes. Please, let's keep up the improvement in December and aim to reduce that to zero.” The banker in the top third of the ratings noted the improvement and asked for information about how the costs were reduced and how the improvement might sustained. This banker did not give a directive, reminder, or instruction in an effort to motivate the group to keep costs low.

The partial correlations when tacked-on traditional antecedents were included in multiple regressions were also significant. As shown in Table 31, the partial correlations were -.48 ($p < .01$) and -.43 ($p < .05$), for regression models 2 and 3 respectively. The beta weight was significant in model 2, $t = -2.73$ ($p < .01$), and marginally so in model 3, $t = -2.00$, ($p = .06$), indicating that

tacking on traditional antecedents was a significant predictor of ratings of technical skill (and, in fact, was the only one) while controlling for the effects of all the other behaviors.

Table 32

Tacked on Traditional Antecedents in Responses to an OSIBA Item from Top and Bottom Rated Technical Experts

Item 9		
TO: You		
FROM: M. Bulgarcu, Dir., Accounting		
I'm attaching these figures as you asked. The costs for last-minute alterations are due to Editorial. I sent Editorial a copy.		
Month	Total Production Costs	Costs Due to Last-Minute Alterations
August	$25,600.00	$5,800.00
September	$26,700.00	$4,900.00
October	$28,000.00	$8,907.00
November	$29,200.00	$6,900.00
December	$19,300.00	$1,100.00

Responses	
Top Rated Banker	**Bottom Rated Banker**
To Editorial Dir.: I assume you have seen these figures. I'm pleased to see the reduction in costs due to alterations but I'd like to hear your views as to how they can be minimized. To Production Dir.: I note the significant drop in production costs. How has this been achieved and is it sustainable?	To Team: [Fwd message] Please note the costs due to last minute changes. As mentioned to you all in a previous email, *it is everyone's responsibility to make sure the information is correct and that no room is left for last minute changes. Please, let's keep up the improvement in December and aim to reduce that to zero.*

Note. The tacked on traditional antecedents are shown in italics.

Hypotheses 2a and 2b

The additive and multiplicative combinations of monitoring and providing consequences were assessed in the traditional way of analyzing information from multiple predictors – through hierarchical multiple regression. Hierarchical regression can be described as a series of simultaneous analyses with optimally weighted combinations of predictors that all use the same criterion. The first step in the regression uses one or more predictors, and the next step adds one or more new predictors to those from the first. The change in R^2 between consecutive hierarchical

multiple regression analyses represents the proportion of variance in the criterion that is shared exclusively with the newly added variables (Licht, 1995).

Hypothesis 2a was tested in the first step of a hierarchical multiple regression – by entering monitoring and providing consequences (calculated by subtracting questionably and not warranted types) as predictors, and using motivational effectiveness ratings as criteria. This multiple regression was discussed in the section on testing Hypothesis 1a, and results were shown in Table 31 (see also Table 33 below). The multiple correlation coefficient was not significant, $R = .33$, $p > .05$. Thus, the additive combination of monitoring and providing consequences was not positively and significantly related to effectiveness, and Hypothesis 2a was not supported. In addition, when the warranted only calculation of consequences was used in the multiple regression equation (Model 1a), the resulting multiple correlation coefficient was slightly higher but still not significant ($R = .37$, $p > .05$).

Because specific types of monitoring and providing consequences were assessed in the OSIBA and found to have significant zero-order correlations with effectiveness (and marginally or significant partial correlations), exploratory multiple regressions were performed with these behaviors as predictors (not shown). The result for a model utilizing monitoring overall and providing *warranted positive* consequences was a statistically significant R of .43 ($p < .05$). These two behaviors, thus accounted for 18.5% of the variability in effectiveness. (The beta weights for both predictors were not significant, however. For monitoring, the beta weight was equal to 0.23 [$t = 1.41$, $p > .10$] and for positive consequences was equal to 0.31 [$t = 1.91$, $p = .06$].) Another multiple regression was performed with monitoring via work sampling and providing positive consequences, and results were similar – the R was .43 ($p < .05$). (The beta weight for monitoring via work sampling was not significant at 0.24 [$t = 1.50$], but was for providing positive consequences [ß = 0.33, $t = 2.06$, $p < .05$].)

Multiple regression analysis is used to statistically test for the possibility of an interaction between variables. A multiplicative term is added to the multiple regression equation as a predictor with each of the component parts of the interaction, and if an interaction is present, the increment in R^2 when the interaction term is added will be significant (Cortina, 1993; Guion, 1998). Hypothesis 2b concerned the interaction between monitoring and providing consequences and predicted that the interaction of the two would significantly add to the prediction of supervisory effectiveness, over and above both alone. Thus, this hypothesis was tested in the second step of the hierarchical multiple regression analysis.

As shown in Table 33, the multiple correlation coefficient for a model with overall monitoring and providing consequences, and a monitoring times consequence interaction term as predictors was .33, $p > .05$ (Model 2). The change in R^2 from model 1 was .001, which was also not significant. In other words, adding the interaction term to the multiple regression equation did not significantly improve the amount of variance in motivational effectiveness accounted for, over and above monitoring and providing consequences alone. (Results using the alternative warranted only calculation for consequences in the regression equation [Model 2a] were very similar – the R including the interaction term was .37, $p > .05$, and the change in R^2 was equal to .003, $p > .05$.) Thus, Hypothesis 2b was not supported.

Table 33

Hierarchical Multiple Regression Results Using Monitoring, Consequences, and an Interactive Term as Predictors

Predictors	ß	*t*	*R*	R^2 Chg.
Model 1			.33	
Monitoring (overall)	0.25	1.42		
Consequences (questionably and not warranted subtracted)	0.15	0.86		

Table 33, cont.

Predictors	ß	*t*	*R*	R^2 Chg.
Model 2			.33	.001
Monitoring (overall)	0.05	0.05		
Consequences (questionably and not warranted subtracted)	-0.05	-0.05		
Interactive term (M overall x C questionably and not warranted subtracted)	0.32	0.20		
Model 1a			.37	
Monitoring (overall)	0.23	1.35		
Consequences (warranted only)	0.23	1.32		
Model 2a			.37	.003
Monitoring (overall)	-0.05	-0.06		
Consequences (warranted only)	-0.07	-0.07		
Interactive term (M overall x C warranted only)	0.47	0.30		

Additional exploratory hierarchical regression analyses were carried out by adding interaction terms to regression equations using specific types of monitoring and providing consequences as predictors (results not shown). For the first, a new multiplicative score was calculated by converting the raw scores for positive consequences into z scores, adding a constant of five, and then multiplying them by the transformed monitoring scores (monitoring scores were also converted to z scores and five points added). The result utilizing predictors of overall monitoring, providing positive consequences, and the interaction term was .45, which approached significance ($p = .07$). The change in R^2 from the model with overall monitoring and positive consequences as predictors was .02, which was not significant. For the second exploratory analysis, another interaction term was calculated using monitoring via work sampling and providing positive consequences. The *R* adding this interaction term as a predictor to monitoring via work sampling and providing positive consequences was .48 ($p < .05$) and the change in R^2 was .04, $p > .05$.

Exploratory Analysis of Additive and Multiplicative Combinations

A second way that the combinations of monitoring and providing consequences were analyzed was by creating overall additive and multiplicative scores for each participant (as described in Chapter 7), and then calculating zero-order Pearson correlations between these scores

and average supervisory effectiveness ratings. (See Table 18 for a descriptive summary of the overall scores and Table 34 for the correlational results.) It was expected that the correlations would be positive and significant. Because multiple regression utilizes monitoring and providing consequences as two independent predictors and weights them optimally in the analysis, the correlations were not expected to be higher than the multiple correlations found in the multiple regression analyses.

Using the overall scores for monitoring and providing consequences (or the warranted minus questionably and not warranted calculation), both the additive and multiplicative combinations were found to be positively and significantly related to motivational effectiveness ($r = .32$, $p < .05$ for monitoring + consequences and $r = .33$, $p < .05$ for monitoring x consequences). The same was true when the overall score for monitoring was combined with the warranted only definition of consequences ($r = .37$, $p < .05$ for monitors + consequences and $r = .37$, $p < .05$ for monitors x consequences). (The magnitudes of these zero-order correlations were nearly the same as the multiple correlations found in regression Models 1 and 1a above, however the zero-order correlations were significant.) In other words, adding scores for monitoring and providing consequences in equal weights was positively and significantly related to average ratings of motivational effectiveness. Multiplying the scores, which does not “allow” high scores on one behavior to compensate for low scores on the other, was also related to motivational effectiveness. There was, however, essentially no difference between these correlations.

Table 34

Combinations of Participant Monitoring and Consequence Scores, and Their Relationships with Effectiveness

Score Combinations	*r* with effectiveness
Monitoring (overall) + Consequences (questionably and not warranted subtracted)	.32*
Monitoring (overall) x Consequences (questionably and not warranted subtracted)	.33*
Monitoring (overall) + Consequences (warranted only)	.37*
Monitoring (overall) x Consequences (warranted only)	.37*

* $p < .05$, one-tailed.

Hypothesis 3

To investigate whether providing antecedents aided in the prediction of effectiveness, the focus of Hypothesis 3, another hierarchical multiple regression analysis was performed. Refer to Table 35 for the results. The first step was identical to the previous hierarchical analysis – overall monitoring and consequences calculated as warranted minus questionably and not warranted types were the predictors and motivational effectiveness the criterion. The second step in the analysis was to add the total or overall score for antecedents to the regression equation. Contrary to Hypothesis 3, the addition of antecedents did not significantly improve the multiple correlation ($R = .34$, $p > .05$) nor explain any additional variability in effectiveness (R^2 change = .005, $p > .05$). The results were the same using the warranted only calculation for consequences (Model 2a), $R = .37$ ($p > .05$) and R^2 change = .001 ($p > .05$). (The addition of antecedents overall did not improve the multiple correlations over and above those for monitoring and providing *positive* consequences [$R = .43$, $p < .05$; R^2 change = .000, $p > .05$] or monitoring *via work sampling* and providing *positive* consequences [$R = .43$, $p < .05$; R^2 change = .002, $p > .05$] either.)

To assess the whether the two specific types of providing antecedents alone might be useful in prediction, additional multiple regressions were conducted (Models 3 and 4). Results were similar to those for antecedents in total. The addition of tacking on traditional antecedents did not

significantly improve the multiple correlation (R = .34, p > .05) nor explain any additional variability in effectiveness (R^2 change = .007, p > .05) over and above overall monitoring and providing consequences. Also, the multiple correlation did not change with the addition of relying on antecedents (R = .33, p > .05, R^2 change = .00, p > .05).

Table 35

Hierarchical Multiple Regression Results Using Monitoring, Consequences, and Antecedents as Predictors

Predictors	ß	t	R	R^2 Chg.
Model 1			.33	
Monitoring (overall)	0.25	1.42		
Consequences (questionably and not warranted subtracted)	0.15	0.86		
Model 2			.34	.005
Monitoring (overall)	0.25	1.42		
Consequences (questionably and not warranted subtracted)	0.17	0.93		
Antecedents (overall)	0.08	0.42		
Model 3			.34	.007
Monitoring (overall)	0.25	1.41		
Consequences (questionably and not warranted subtracted)	0.16	0.88		
Tacking on traditional antecedents	0.09	0.50		
Model 4			.33	.000
Monitoring (overall)	0.25	1.39		
Consequences (questionably and not warranted subtracted)	0.15	0.82		
Relying on antecedents	0.003	0.01		
Model 1a			.37	
Monitoring (overall)	0.23	1.35		
Consequences (warranted only)	0.23	1.32		
Model 2a			.37	.001
Monitoring (overall)	0.24	0.06		
Consequences (warranted only)	0.23	0.07		
Antecedents (overall)	0.03	0.30		

Chapter 9

DISCUSSION

Results and Their Implications

As hypothesized, two supervisory behaviors measured via an in-basket exercise were related to motivational effectiveness. Bankers' overall monitoring ($r = .30$, $p < .05$) and providing warranted consequences ($r = .29$, $p < .05$) were linked to ratings of how well they motivated others. The correlation between motivational effectiveness and another way of operationalizing providing consequences, by subtracting scores for not and questionably warranted types, was in the predicted direction, but not significant ($r = .23$, $p > .05$).

Bankers in the sample who were rated as effective at motivating others by their supervisors, subordinates, peers, or others monitored performance more than those who were rated lower. Those who were highly rated asked for details or sought information about performance or potential problems (or suggested others do the same). They also did such things as finding out first-hand how employees were performing to verify complaints, asking to view documents before they were released for production, and requesting to see evidence of problems for themselves before making decisions. For example, the top-rated banker in Table 28 suggested that a manager get information from his group to help them finish a troublesome project, "Please talk with each one of your team members and try to find out what their issues are."

More effective bankers, it was also found, communicated their knowledge of employee performance when it was deserved. They recognized effort or excellence, in particular. They acknowledged good results -- they noticed positive trends in reports, for instance, and commented on the progress being made. They also thanked people for bringing problems to their attention and for dealing with angry customers. In Table 30, for example, the higher rated banker acknowledged

the memo sender for bringing up his concerns about the quality of images being printed and the production workers' lack of concern for the problem by writing a simple "Thanks for your advice."

Technical expertise was the focus of Hypothesis1b. As hypothesized, neither monitoring nor providing consequences were related to average ratings of technical skill ($r = .11$ for monitoring overall, .04 for warranted only consequences, and .00 for consequences overall, $p > .05$). In other words, there was no difference in scores for monitoring and providing consequences between those bankers who were rated high in technical expertise and those who were rated low.

Results in support of Hypothesis 2a, that an additive, weighted linear combination of monitoring and providing consequences will be positively and significantly related to supervisory effectiveness, were mixed. A multiple regression with overall monitoring and providing consequences as predictors was not significant ($R = .33$, $p > .05$), nor were any of the beta weights for the predictors. The multiple correlation for a similar regression utilizing warranted only consequences was higher, but also was not significant ($R = .37$, $p > .05$). Results from an exploratory analysis in which scores for monitoring and providing consequences were added together for each respondent and then correlated with their effectiveness rating did, however, support Hypothesis 2a. The magnitude of these correlations were similar to the multiple correlations ($r = .32$, $p < .05$ for monitoring + consequences overall and $r = .37$, $p < .05$ for monitoring + warranted only consequences), but were significant. In other words, adding scores for monitoring and providing consequences in equal weights was positively and significantly related to average ratings of motivational effectiveness. The contradictory results might be explained as being related to statistical power, which was likely lower for the multiple regression analysis.

For Hypothesis 2b, which predicted that the product of monitoring and providing consequences would improve the prediction of supervisory effectiveness over and above monitoring and providing consequences alone, no support was found. Adding the interaction term as a third predictor to the multiple regression (with monitoring and providing consequences overall) did not

increase significantly the amount of variability in effectiveness accounted for by the group of predictors (R^2 change = .001, $p > .05$). In an exploratory analysis of the relationships between effectiveness and individual-level additive and multiplicative scores, both the additive and multiplicative combinations of monitoring and providing consequence (overall) were found to be related significantly to effectiveness ($r = .32$ and .33, respectively, $p < .05$), but the correlation for the multiplicative score was not significantly greater than that for the additive score. Results were similar using the warranted only calculation for providing consequences. Thus, it appeared that scores for the categories could compensate for one another (e.g., a high score for monitoring could make up for a low score on consequences) and there was no interactive effect (e.g., the relationship between providing consequences and effectiveness is dependent upon the level of monitoring). A visual inspection of the data, moreover, did not show patterns of scores that would be indicative of an interaction -- many participants who had very high scores on one behavior and very low scores on the other.

No support was found for Hypothesis 3, concerning providing antecedents, despite earlier studies that showed a link between the number of antecedents given before monitoring and effectiveness (Reynard & Komaki, 1995) and that providing antecedents functioned as a suppressor variable (Jensen & Komaki, 1993). Adding antecedents overall as a predictor in the multiple regression with monitoring and providing consequences (both warranted only and with questionably and not warranted subtracted) did not significantly increase the amount of variability in motivational effectiveness that was accounted for by the group of predictors (R^2 change = .001 for warranted only and .005 for overall consequences, $p > .05$). Results were similar when the two types, tacking on traditional antecedents (R^2 change = .007, $p > .05$) and relying on antecedents (R^2 change = .001 for warranted only and .000, $p > .05$), were used as predictors.

Providing antecedents, however, was related to the criterion of technical skill. Providing the tacked-on traditional type of antecedents in the in-basket was found to be negatively related to

technical skill ratings (zero-order r = -.44, p < .01). The correlation was even larger while controlling for all other behaviors measured in the in-basket (partial r in model with nine predictors = -.48, p < .01). In other words, those rated high in technical skill, in general, gave the *least* directives, reminders, or instructions in order to motivate others. For example, when presented with a report of production and alteration costs, a banker rated in the top third of the group in technical expertise (shown in Table 32) noted improvement and asked about how the costs were reduced. He/she did not give any directives or instructions. In contrast, a banker rated in the bottom third also noted the improvement, but gave antecedents to the entire team: “Please note the costs due to last minute changes. As mentioned to you all in a previous email, it is everyone's responsibility to make sure the information is correct and that no room is left for last minute changes. Please, let's keep up the improvement in December and aim to reduce that to zero.” A member of the author’s dissertation committee suggested that these results might be explained by people who are high in technical expertise assuming that others are as well, and thus not needing directives (J. Lefkowitz, personal communication, June 19, 2005). It could also be surmised that being regarded as a technical expert could be affected by one’s behaviors. Those who provide many directives, reminders, and instructions when they are not requested, for instance, might be viewed as lacking in technical knowledge. In addition, providing tacked-on antecedents was found to be an independent predictor (inverse) of technical skill ratings, as its beta weight in a multiple regression analysis was significant (β = -0.51, t = -2.73, p < .01). This indicated that knowing a banker’s score on providing tacked-on traditional antecedents was predictive of his/her level of technical skill as rated by others.

Despite the observed relationships between monitoring and providing consequences and effectiveness, results for analyses in which the behaviors were combined (and each held constant) were not as clear. As discussed previously, the multiple correlation coefficients in regressions utilizing monitoring and both calculations of consequences as predictors were not significant. None of the beta weights for these behaviors in the regression equations were significantly different from

zero. In addition, partial correlations for both monitoring and providing consequences were lower than the zero-order correlations and not significant (partial *r*s = .24 for monitoring overall and .15 for positive consequences overall, $p > .05$, partial *r*s = .23 for monitoring overall and .23 for warranted only positive consequences, $p > .05$). These results did not confirm the correlational analyses.

This study has implications for the Operant Model. It has served to extend the generalizability of the Model because it was an investigation of the behaviors of the model in a very different field site. Previous field studies were conducted in such industries as newspaper publishing, insurance, construction, law enforcement, theater, film, and sailing (Komaki, 1998). The model, until now, has not been tested with a sample of people whose tasks were to structure deals in an investment bank setting. Furthermore, the participants were highly educated -- the majority had both undergraduate (97.0%, n=33) and graduate degrees (80.6%, n=31) – and received relatively large salaries.

The results of this study also have implications for different types or qualities of the major behaviors in the Model. In previous research, distinctions were made between subcategories of monitoring and providing consequences, some of which were also assessed in the OSIBA. As found in an earlier study (Komaki, 1986), a specific method of monitoring, via work sampling, was related to effectiveness ($r = .28$, $p < .05$). In other words, effective bankers requested to observe workers in action and/or examine the products of their work. For example, the top-rated banker in Table 29 wanted to look at both the old and new reimbursement forms when an employee complained because a contracted photographer had difficulty completing the new one: "Please let me have a copy of the old and the new reimbursement forms (see the complaint from Rosso). I want to discuss with you once I have a chance to review them."

An unanticipated finding was that a particular evaluation of consequences, those that were positive, was also found to be significantly related to ratings of effectiveness ($r = .36$, $p < .05$). In

developing the scoring of positive consequences in the in-basket, distinctions were made among six different special types, including, for example, giving simple, short acknowledgements; broadcasting positive consequences to multiple parties; and acknowledging the bearer of bad news. Although the specification of these types was new and none was expected to be better than any other, the special type in which the respondent acknowledged the bearer of bad news was found to be related to motivational effectiveness ($r = .40$, $p < .05$). (This correlation was one of the strongest found in the study.) This meant that effective bankers thanked people for bringing unwanted or undesirable information to their attention. They wrote “Thanks for letting me know,” “Thanks for the heads-up,” or “Thanks for your advice” in response to being presented with issues, concerns, or problems (such as in the example depicted in Table30). It was surmised that this behavior was particularly effective because people who encourage others for drawing attention to problems display an openness to hearing issues, and as a result, people will come to them for help. A former member of the group (who left at least one year before the study began) felt that the encouragement might engender trust (D. Prottas, personal communication, March 31, 2005). In sum, the results for the subcategories of monitoring and providing consequences were consistent with the Operant Model.

In keeping with traditional research, the OSIBA was investigated for its psychometric properties in the present study. In terms of reliability, the OSIBA had adequate levels of scoring consistency. Interrater percentage agreement scores based on the occurrences of behaviors were fairly high. For the major categories of behavior, scores ranged from 82.5 to 91.4% and for subcategories, 64.3 to 100%. All figures except for three subcategories were above 80%, the accepted threshold for new measures (Miller, 1997). Results were similar for intraclass correlations, in which raters’ scores were compared. Consistency ICCs ranged from .73 to .88 for categories and .42 to 1.0 for subcategories and absolute agreement ICCs ranged from .68 to .87 for categories, .33 to 1.0 for subcategories. Many coefficients exceeded reasonable reliability estimates of .70 – two

categories and three subcategories were exceptions. These results suggest that the OSIBA was reliably scored, in the sense that two people generally judged responses to be the same type or category.

Internal consistency indices, indicating the relationships between behaviors measured across individual in-basket items, were very low and sometimes negative, however. Alpha coefficients for overall categories ranged from -.31 to .30 and for major subcategories, -.31 to .44. Correlations between scores for the same behavior over different items ranged from moderately positive to moderately negative, and these mixed signs might have had a detrimental effect on the estimation of the alpha statistic. Low frequencies of some behaviors may have also attenuated alphas. The results suggested that respondents did not consistently monitor, provide consequences, or provide antecedents across different items in the OSIBA.

It was not expected that respondents would consistently display behaviors across items because an in-basket exercise differs from a traditional test or questionnaire in which items measuring the same construct should be correlated (J. Smither, personal communication, April 7, 2006). Test or questionnaire items provide an equal chance for respondents to demonstrate a construct/behavior (e.g., knowledge of a particular subject, extraversion). In-basket items (and also exercises in assessment centers), however, often provide respondents with more of a chance to demonstrate particular behaviors on some items as compared to others. For example, respondents could potentially give a simple one-or-two word positive consequence on every item of the OSIBA, but three items give the respondent specific information about subordinates' good performance and two show numerical reports with improvements embedded in them. In-basket items also offer respondents opportunities to demonstrate several behaviors, and this was the case for the OSIBA. It was possible to demonstrate at least two behaviors on every item. (Appendix B shows an example of a score sheet for an item in which several behaviors could be demonstrated. The two most likely

behaviors were monitoring and providing positive consequences, but questionably warranted negative consequences and antecedents were also possible.)

A lack of internal consistency seems to be a problem plaguing other in-baskets in the published literature as well (Brannick, Michaels, and Baker, 1989; Brass & Oldham, 1976; Hakstian & Scratchley, 1997; Schippmann, Prien, & Katz, 1990). For example, low alpha coefficients were found for five dimensions in two alternate forms of an in-basket developed by Brannick, Michaels, and Baker (1989). Coefficients, which were computed separately for odd and even items within both forms, ranged from -.41 to .61 in Form A and .09 to .58 in Form B. (Reliability estimates for alternate forms were moderately low as well, ranging from .21 to .43 across the dimensions.) Also, corrected split-half reliability coefficients for six leadership scoring categories ranged from .13 to .58 for Brass and Oldham's (1976) in-basket. The researchers attributed the low reliabilities to "(a) some of the leadership activities … were used rather infrequently in response to the in-basket and (b) some in-basket items elicited certain leadership activities more than other in-basket problems" (p. 655).

These results highlight the importance of having a sufficient number of items for which to score each of the behaviors. If respondents monitor via work sampling, for example, in only a small number of items in which it is possible, it is necessary to provide many opportunities so that an adequate amount of the behavior can be "caught" in the exercise. Also, a possible explanation for the lack of results for providing antecedents is that perhaps more opportunities were needed to measure them. Only three items were used to assess relying on antecedents in unclear situations, for example (as compared to six items for monitoring via work sampling and all 16 for positive consequences). Even for live observations of the behaviors in the Operant Model, many observations are necessary to get representative information – a generalizability study showed that 20 half-hour sessions at random times of day were needed (Komaki, et al., 1986).

This study has several implications for managerial selection. It is the first to explore the relationship between the behaviors in the Operant Model measured via an in-basket exercise and supervisory effectiveness, thereby providing a theory-based measurement tool that might (with further research) be used in a practical way to select managers. The content of the in-basket addresses a need in the "predictor space" in managerial selection as well. It assesses non-cognitive interpersonal dimensions that have to do with supervising employees, which are often neglected in other selection techniques and instruments.

In comparison to other measures in the selection literature, the results for the revised OSIBA were about the same as most. Meta-analyses have yielded average estimated predictive validities for overall job performance (corrected for artifacts) of .34 for situational judgment tests, .35 for biodata measures, .37 for assessment centers, and .51 for structured interviews and tests of general mental ability (McDaniel et al., 2001; Schmidt & Hunter, 1998). An average criterion-related validity coefficient for in-baskets is not available because only one qualitative review has been completed (Schippmann et al., 1990). Of the 22 studies reviewed, only six utilized performance ratings as a criterion. One other study has been published since the review (i.e., Hakstian & Scratchley, 1997), and across these studies, validity coefficients ranged from -.46 to .72. Results for the OSIBA, in most cases, were similar to other in-baskets. While it is interesting to consider these predictive validities in comparison to the OSIBA, the values confound method with content (or actual skills/knowledge areas). Arthur et al.'s (2003) meta-analysis of assessment center dimensions provides a better comparison of the content to some extent (many different dimensions were combined to perform the analyses and all were measured via various assessment center activities). Of the six "mega-dimensions" assessed, the most interpersonal noncognitive dimensions, consideration/awareness of others and influencing others, had estimated validities for criteria of supervisory ratings of performance or promotion of .25 and .38 respectively. Results for the OSIBA

are comparable to these. (A benefit of the OSIBA was that it achieved the level of validity in less time and effort than an assessment center.)

This research makes a valuable contribution to the literature on in-baskets as well. In the last (and only) review of the in-basket literature, Schippman et al. (1990) were critical of the lack of good research in the construction or evaluation of in-baskets. They concluded that research and opinion about the in-basket could be best described as “incomplete” and “arrested in time,” and that the technique had become institutionalized not based on compelling research evidence, but more on belief and conviction. The results found here directly address some of Schippman et al.’s (1990) concerns -- the OSIBA is an exercise well-grounded in theory with acceptable levels of scoring consistency, and some evidence has been shown that the behaviors measured are related to supervisory effectiveness.

Limitations

One limitation of the research presented here is that it is, of course, a single study conducted with a single sample, with results that have not been replicated across different conditions. As in all single research studies, the question is whether the results can be generalized to different types of subjects doing different types of tasks.

Because the current research investigated the relationship between a predictor (or set of predictors) and a criterion of job behavior, it was a criterion-related validity study. The design was concurrent, in that present employees for whom job performance criteria were available took the in-basket exercise. Another type of criterion-related validation study, in which there is a time interval between taking the test and collecting criterion information, is known as a predictive design (Landy, 1989). Ideally, in this method, all applicants for a job are given the test, they are hired without using those test scores to make the hiring decision, and then after some time has passed, performance data are collected. This type of study more closely matches what occurs in a real selection situation because the intent of the study is to investigate how earlier observations (i.e., test scores) are

predictive of later behavior (i.e., criterion scores). There is also the potential advantage that a wider range of scores on the test are achieved. The range restriction possible with a concurrent design could result in artificially depressed correlations that might not be statistically significant.

Another limitation of the present study is related to the OSIBA. Though the in-basket has advantages over live observation in selection situations, a difficulty is the scaling or valuation of behaviors. In live observation, the metric is a very straightforward percent of time. If a person delivered five positive consequences in 30 intervals of time, it would count the same as when they delivered five negative consequences in 30 intervals. In the in-basket, however, point values were assigned depending on the quality of the behaviors respondents demonstrated. Point values of -4 through 5 were assigned for positive consequences, and scores ranged from -6 to +6 for negative consequences. A score of -6, for example, would be given to a questionably warranted negative consequence in which the respondent delivered the same negative evaluation twice and threatened to fire or demote someone in the in-basket. It is not clear if the behaviors are valued appropriately. In fact, results of the present study indicate that the values subtracted for questionably and not warranted consequences in particular could be “off” in some way. Providing consequences calculated as the sum of points for warranted minus points for questionably and not warranted was not significantly related to motivational effectiveness ($r = .23$, $p > .05$). Using the warranted only calculation of consequences, however, the correlation was slightly higher and significant ($r = .29$, $p < .05$). It could be, for instance, that subtracting four points for an unwarranted positive consequence or six points for a questionably warranted negative consequence might be too much.

The criterion measure of supervisory effectiveness, which was a single item asking respondents to judge a person’s “grade” at motivating others, was also a limitation. Most validation research utilizes criterion measures that contain several items, comprising more than one dimension of performance, and the items are often summed or averaged to form an overall composite. As a result, internal consistency reliability estimates can be calculated for the measure. With the single

item used in the present study, reliability could not be calculated. Thus, the criterion predicted was questionably, or at least not "ideally," reliable, and might have resulted in lower correlations than would have been found with more reliable criteria. (A test-retest reliability coefficient of .76 was found in a past research study that utilized the same item [Jensen & Komaki, 1993], however. In addition, in the present research, the ratings from different sources were correlated and those from subordinates were found to be similar to those given by others [$r = .55$, $p < .01$].) The criterion measure was also, like most, subjective or judgmental, which assumes that when managers are effective, they will be regarded and evaluated accordingly (Komaki, 1998; Robertson & Smith, 2001). Furthermore, the measure did not require the respondent to rate effectiveness in terms of specific, observable behaviors. Because they could lead to a fuller understanding of the domain of managerial success, behaviorally-based performance measures or systematic observations have been recommended to supplement judgmental estimates of effectiveness (Campbell, Dunnette, Lawler, & Weick, 1970; Cascio, 1987). Behaviorally anchored rating scales (BARS), which contain anchors that describe various degrees of supervisory effectiveness as specific behaviors, were considered for this study (as they had by Komaki [1986]), but the time and resources needed to develop them made them prohibitive for the bank.

In addition, the size of the sample was a limitation of the research. Though some significant results were found in the study, the sample size of 35 was relatively small. A larger sample may have produced more variability in both the behaviors on the in-basket and the ratings of the criterion and would have helped raise the power of the analyses to detect significant results. A power analysis showed that for a sample size of 35, utilizing a multiple regression with three predictors and an alpha level of .05 and assuming a medium effect size ($R^2 = .13$), power was equal to .40.

Future Research

Additional research should be conducted involving the in-basket measure and its expanded version. The present study is only the second to utilize it. More research on different samples in

different settings would be beneficial for establishing the measure's psychometric properties. It would be helpful to assess reliability in other ways besides inter-scorer agreement, perhaps by developing a roughly equivalent form of the exercise (alternate forms) or by administering it at two points in time (test-retest), and then assessing the relationship between respondents' scores on the two administrations.

More work should also be done to develop evidence of construct validity -- to show that the behaviors that are thought to be measured in the OSIBA are actually being measured. Another study (or studies) such as the multi-trait multi-method one involving the original OSIBA (Komaki et al., 1990) would be ideal. Because the indices of the relationships of the behaviors measured over different items were so low and there was no other construct validity evidence gathered in the present research, there is some question on what the exercise actually measured. (Internal consistency measures can be an indication of construct/convergent validity -- i.e., if different items are measuring the same construct or behavior, they should be related. It is for this reason that some have noted parallels between the low internal consistency on in-baskets and the low convergent validity of dimensions across different exercises in assessment centers [e.g., Brannick et al., 1989; J. Smither, personal communication, June 19, 2005].) Komaki et al. (1990), however, did find that scores for monitoring and providing consequences in the original version of the in-basket were related to the amount of time spent performing them as observed on the job. This relationship has not been explored for providing antecedents because the category is new to the in-basket. (A possible reason for why the results of the present study did not parallel those of Jensen and Komaki [1993] or Reynard and Komaki [1995] is that what is being measured as antecedents on the in-basket might be different from that in direct observation.) More research is warranted, especially because the OSIBA is based in theory and the constructs measured fulfill a need in the "predictor space" in managerial selection.

In addition, it might be interesting to investigate the relationship of the scores on the in-basket to other cognitive and noncognitive predictors, such as tests of general mental ability and personality inventories. More criterion-related validation research is needed, as well. A predictive validation design would be a good addition to the research literature. Also, a cross-validation procedure, in which a regression equation developed for one sample is applied to a second to see how well it fits, would be interesting to perform. Ideally, before being used operationally, it would be shown that scores on the behaviors measured via the OSIBA are valid and generalizable predictors of motivational effectiveness. Because the intent is to use the OSIBA operationally, it would also be helpful to collect information from respondents regarding their experiences in the administration of the exercise. For all research, a larger sample size should be sought to help boost power and confidence in detecting significant results.

More studies should also be conducted to closely examine the proposition of the Operant Model that monitoring and providing consequences interact to produce an effect. The possibility that both monitoring and consequences (together) make a difference in effectiveness has been given little consideration in previous research. With only the present field study appropriately testing this relationship (and having relatively low power to detect results), much more work needs to be done to “flesh out” and test this part of the Operant Model.

Further research should be conducted on the category of antecedents and its role in predicting effectiveness, using both observational and in-basket measures. More work could be done on honing the definition and measurement of antecedents in the in-basket, starting perhaps with an in-depth analysis of the “antecedent responses” from the top- and bottom-rated bankers in this study. As mentioned earlier, it would be also beneficial to develop more items designed to elicit the behavior. Seven items had originally been generated to use to tap the behavior, but after field testing, only five were added to the in-basket -- three for the relying on antecedents type and two for choosing between antecedents and consequences. More work should be focused on developing

items for relying on antecedents so that an adequate number of items are available for respondents to demonstrate it and, if possible, for the tacked-on traditional type, to strengthen the observed relationship with technical expertise. In addition, the specific conditions under which providing antecedents might be negative or act as suppressor variables should be investigated. As Jensen and Komaki (1993) speculated, the tasks, personnel, and directions for an organization may place emphasis on initiation of performance. Future research needs to identify and carefully look at these conditions and make comparisons to suppressor conditions or conditions under which providing antecedents might be negative. This is important work since suppressor variables and their use in selection are controversial (Cascio, 1987; Guion, 1990; Tzelgov & Henik, 1991).

The finding of a significant relationship between providing antecedents and ratings of technical expertise, though unexpected, was very interesting. This was the first study of the behaviors in the Operant Model in which providing antecedents were related to any supervisory or managerial criterion. It would be helpful if future research efforts on providing antecedents (and other behaviors) included measures of technical knowledge, so that this result might be replicated.

In sum, this research has shown that two noncognitive, interpersonal, supervisory behaviors grounded in theory and measured via an in-basket assessment were related to motivational effectiveness. With more research, the in-basket exercise could have practical applications for selecting people for managerial positions.

A p p e n d i x A

In-basket Exercise

Participant No. b0501____

Motivational Effectiveness Exercise

Copyright © 2002 by Judith Komaki. All rights reserved. No part of this publication may be reproduced or transmitted in any form by any means, electronic or mechanical, including photocopy, recording, or any information storage or retrieval system, without permission in writing from copyright owner.

Participant No. b0501____

July 11, 2002

To: Project participant

From: Judi Komaki and Michelle Minnich

We're pleased that you're willing to participate and to complete the Motivational Effectiveness Exercise. It's an "in-box" exercise in which you'll find messages from your staff, financial reports, and reminders to yourself, much as you would in your own "in-box" at your office.

It should take you from 60 to 90 minutes to complete the exercise. Ideally, you should try to work uninterrupted through the entire exercise. It's great to do on the plane or train. However, if this is not possible, work through parts of the exercise as you get time.

When you are finished, please place the entire exercise in the envelope provided and mail it back to us:

Judith L. Komaki
Baruch College, Box B8-215
One Bernard Baruch Way
New York, NY 10010

DUE: August 1, 2002.

We'll provide confidential, individualized feedback to every participant. You will learn about your supervisory style, how you behave when you interact with others on the "in-box" exercise, what you do particularly well, and how you can improve to most effectively motivate others. We'll relay your feedback via videoconferencing or if possible in person in early September. No information about you as an individual will be provided to the firm. It will be given to you only for development purposes. Participants in other organizations have found the feedback we give insightful and sometimes illuminating.

If you have any questions, please contact us at the above address, via fax: 1+646-312-3781, or by email: Judi Komaki (Judith_ Komaki@baruch.cuny.edu), Michelle Minnich (kevmic@fast.net) .

Thank you,
Judi and Michelle

Before you turn this page, please read the following License agreement regarding your use of this material. By proceeding beyond this page you indicate that you accept the terms of this license and shall abide by them. If you do not accept, please return these materials unused.

License

This in-box exercise and all materials associated herewith (the "IBE") are the proprietary property of Judith L. Komaki. She grants to you and you hereby accept a non-transferable, non-exclusive license upon the following terms and conditions. I acknowledge that:

1. The IBE constitutes the proprietary property of Judith L. Komaki, and that she retains exclusive ownership of all IBE materials.

2. I shall not disclose the contents of the IBE to any person not authorized by Judith L. Komaki. I shall not allow others to have access to the IBE, directly or indirectly, (except those with express written authorization from Judith L. Komaki), and shall not make or allow others to make copies or reproductions of the IBE in any form without the prior written consent of Judith Komaki.

3. I shall disclose to Judith L. Komaki any breach of this license which may come to my attention.

Participant No. b0501____

July 2002

Welcome,

You are the Publisher responsible for producing the monthly magazine, FOODSTUFF DIGEST. You have a fine reputation for managing your people to make FOODSTUFF the best it can be.

Right now, you want to deal with the items in your "in-box." There are a variety of messages and reports from your staff as well as people outside FOODSTUFF. You may respond in whatever way and to whomever you wish. FOODSTUFF is a relatively flat organization and run informally, not hierarchically. Hence, it would not be unusual for you to deal with problems or make suggestions to anyone in your reporting line.

To get the most representative view of yourself, you should respond as you usually do. Please remember that your feedback is confidential.

Some pointers:

A. Apportion your time so that you deal with all 22 items.
B. When you turn the page, you'll see a message or a report on the top. The first page is a reminder to yourself. It lists the people who report directly to you. On the bottom of each page are spaces for you to write to two persons. If you wish to write to a third person, you can do so at the bottom of the page. Please keep your responses within the allocated space and on the front of the page.
C. Don't just arrange to meet. Instead, tell us what you would say when you met with the person.
D. It's not necessary to write a formal response. Just say what you want to say.
E. You may forward or copy just as you would in an email. Please note cc and the name of the person to whom you want to send.
F. You should allow 60 to 90 minutes to complete the "in-box."

Any questions, please email me at: Judith_Komaki@baruch.cuny.edu or send a fax to: 212-312-3781.

Regards,

Judi Komaki

i01

Reminder note to yourself

Budget projections due from staff TODAY!!
Need to be on boss' desk by TOMORROW

Department	Received
Art: J. Rosso	Yes
Editorial: M. Lohninger	
Advertising: J. Child	Yes
Production: J. Tower	Note: Help! Having Problems
Mktg./ Circ.: A. Ortega	Yes
Accounting: M. Bulgurcu	Yes
Human Resources: T. Yoneyama	Yes

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

i22
TO: You
FROM: Taka Yoneyama, Dir., Human Resources

Can I help it if Marketing can't get their people there on time? By the time the trainees finally got there, the trainer had left. Now we're having to reschedule. I told them it's going against their budget.

How much clearer can it be? This was what was sent. See attached map and directions.

Directions:
From FOODSTUFF go south,
then west 5 miles to Michelin Bldg.

Allow enough time – traffic on
Concord can be heavy!

FOODSTUFF
Training Center
Concord
Michelin Bldg.
N

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ______________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ______________________________

Participant No. b0501____

i32
TO: You
FROM: Bruce Poole, Mgr., Advertising

Status update for awards banquet two short months from now:

1. Ethel is still looking for keynote speaker.
2. Gokhan is having trouble tracking down the person who engraves the plaques.
3. We finally put Nobu at the job and he did his usual thorough search and found us a dynamite place for 80% of the estimated cost.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i60

TO: You
FROM: Sally Clarke, Publisher, GOOD EATING

Hey, what's happening*?!
I'm sure I sent you a copy? Yeah, remember it the only one in which you could clearly see the onions in that dish didn't turn red.
Remember it?
In fact, I sent it twice!

So all you need to do is to find it.
I know you've got it somewhere.
And then I need you to give me a quick up or down.
Okay?

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

---

i42
TO: Publisher
From: Babette F., Information Systems Mgr., Foodstuff Digest

Dinesen's still one of the most popular trainers. But word's out that she's writing plays in her spare time. That explains why more and more of her class examples deal with plays she's seen. You should see how lively she gets! I'm sure she's going to lose interest in her job. I think that someone should let her know that we are concerned about this.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i44
To: You
From: R. Mondavi, Shipping

Have you seen Bollinger's score? Her staff rated her in the bottom 10%. She's so high and mighty. She thinks she's better than everyone.

Maybe it's a mistake. I know these are preliminary results, but Human Resources must have screwed up. I've heard there are lots of mistakes in the scores. I would love to see her reaction. I thought I might forward her the results to see how she reacts. I know I'd eventually have to tell her that the results may be wrong. But it would give me satisfaction to prick her bubble even if it's only for a few days. What'd you think? If those figures are correct, we have to tell her anyway!

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i82
TO: You
FROM the desk of J. Tower, Dir. of Production, FOODSTUFF DIGEST

The rep has offered
- -Training to familiarize the employees with the machinery
- -To let them practice in a "training lab" on company premises.

Any ideas on how we can set up the schedule? We have 2 options.

		Tue. 23rd **Loading**	Wed. 24th **Printing**	Thurs. 25th **Cutting**	Fri. 26th
Option 1 Cost = $5,000	a.m.	Basic Lecture	Basic Lecture	Basic Lecture	Practice (loading, printing, & cutting)
	p.m.	Advanced Lecture	Advanced Lecture	Advanced Lecture	
Option 2 Cost = $10,000	a.m.	Basic Lecture	Basic Lecture	Basic Lecture	Practice (loading, printing, & cutting)
	p.m.	Practice	Practice	Practice	

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

Participant No. b0501____

i61
To: You
From: Anton, Art Dept.

That special banquet is scheduled. But we still need the name of the town of that special cheese. I thought you said you remembered. It's not good enough to say it's from western Turkey. It's the cheese that is generally eaten with raki.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i12

EMAIL MESSAGE from S. Joe, Editorial Assistant

TO: You@FOODSTUFF.com
FROM: sjoe@FOODSTUFF.com
SUBJECT: RUSH copy for NEXT issue

I know you think there have been way too many mistakes. But there're not my fault.

Everything is finally inorder for the next issue.

I received your modificantions. They're being inserted right now.

I'll take over the camera-ready copy and see that it get sent out tomorrow morning. I'll make sure that you get a copy tomorrow afternoon.

Joe

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i40
TO: Publisher
From: S.C. Smetana, Mgr., Accounting

I know we're not supposed to give out comp days without approval. Gil Martinez Soto worked the past 14 days straight, for at least 12 hours a day. You weren't around, so I made the decision that he could take a day off without charging him a vacation day. He asked for the day, the only thing is that now we are having problems getting it approved in the Human Resources dept.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

Participant No. b0501____

i21
TO: You
FROM: Hirohisa Koyama, Graphic Designer

Here's the artwork for the new restuarant column, as you asked.

FOODSTUF DIGEST

Best Restaurants

© FOODSTUFF, Inc.

I'm sending it to the Production Dept. tomorrow.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ___________________________

Participant No. b0501____

i50

EMAIL MESSAGE from Antonio Ortega, Dir., Mktg./Circ.
TO: You@FOODSTUFF.com
FROM: aortega@FOODSTUFF.com

Every month, I have scheduled the conference room for a departmental meeting with Rick Bayless.

Now your secretary tells me that the Art deptartment is going to be using the room.

I can’t believe it!
What’s going on?

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i31
PHONE MESSAGES taken by Sherry, Secretary

Nozaki, your favorite person (ha ha) – he wouldn't leave a message.

Auditor really impressed with Anne in Accounting.

Left messages for Kevin in Editorial. Twice. Still NO reply.

Sent copy of Rick Bayless' letter of resignation to Human Resources.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i43
TO: Publisher
From: S. Lukins, Asst. Treasurer

That damn postal increase… Pelin - - I told her not to, but she took the word of the MPA, and estimated too conservatively. Damn lobbyists, they snowed her too. She should have known. We got a double digit postal increase!!!

I know, I know. The board of directors let it go this time, but they won't tolerate a repeat. Yeah, we got through it okay. But, Pelin should have known. For next year get rid of her. She's a disaster!

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

i10
TO: You
FROM: J. Rosso, Dir., Art
SUBJECT: Accounting reimbursement forms

Here we finally get Maria to do the shoot, but we can't even get her reimbursed. The reason: She didn't fill the form out right!

I got a copy of the new form. Now I understand why she couldn't fill it out. It's a mess!

I think that they ought to go back to the old form. I'm sure you'd agree.

P.S. Don't worry. After several phone calls and begging, I finally got Accounting to send Maria a check.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i11
TO: You
FROM: Rick Bayless, Associate Publisher

I need to talk with you ASAP. Got another irate call from one of our largest accounts, the Pheasant Peasant Atlanta-based chain.

The owner is furious with Advertising Dept. He claims:

1. No one answers phones. Phone rings and rings.
2. When someone finally answers, they're incompetent and/or rude.

How do you want me to verify his complaints about Advertising?

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i91
TO: You
FROM: Mr. Puck, Art Department

I'm concerned with the quality of the images that are coming off Press 1 – it's really very poor. I've talked with Bobby Flay and his print shop workers several times about the importance of quality. Bobby is convinced that it's a problem with the press, but I think he and his crew just don't care. That press has been working just fine for the past few years. Would you be willing to speak with the group also? Maybe then they'll care.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ______________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ______________________________

Participant No. b0501____

i81
TO: YOU
From: M. Hazan, Training Mgr., Foodstuff Digest

To build each manager's skills in motivating employees to do a good job, the consultant from TrainingRUs, Inc. recommends:

Transformational Leadership

- will help participants master their evolving roles as organizational leaders
- will learn how to communicate and mobilize teams around a shared vision

Lead with Feedback

- will help participants gain an organizational skill which has been lost or abandoned -- giving and receiving feedback.
- participants will learn how to effectively exchange information with the people they manage, increasing team performance.

We can’t afford both. Which do you feel is more critical?

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

---

i100
TO: You
FROM: Mario Lohninger, Director, Editorial
RE: Dan Boulud

I know, I know, we both warned Dan when we promoted him. He's quick, and creative. But, he's still late. In the afternoons, he is sometimes 15 to 30 minutes late. I've heard rumors about why he is late, but who knows… Last week the whole staff was here at 2 p.m. We had to wait a half-hour for Dan.

One of my staff has just pointed out to me, Dan's not always late. In the mornings he's the first one here.

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i20
TO: You
FROM: J. Child, Dir., Advertising
SUBJECT: Advertising accounts by year

Here's the tabulation of the number of accounts as gained and lost by Advertising Dept.:

	# Accounts at beginning of year	Changes in account status during year	# Accounts at end of year
1998	70	+7	77
1999	77	+9	86
2000	86	-3	53
2001	83	+12	95

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i04
TO: You
FROM: the desk of L. Rivera, Mgr., Mktg./Circ., FOODSTUFF DIGEST

I just don't understand my group. I worked with them to create a timeline and milestones, with a final deadline in plenty of time. They just aren't making any progress!

I've heard some complaints -- someone isn't doing their part and they're having disagreements over which clients to target. This resistance isn't like my group -- we've done promos <u>many</u> times before and they were successful!

Can you give me some advice – what should I do about this situation?

Thanks,
Luis

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

i30

TO: You
FROM: M. Bulgurcu, Dir., Accounting

I'm attaching these figures as you asked. The costs for last-minute alterations are due to Editorial. I sent Editorial a copy.

Month	Total Production Costs	Costs Due to Last-Minute Alterations
August	$25,600.00	$5,800.00
September	$26,700.00	$4,900.00
October	$28,000.00	$8,907.00
November	$29,200.00	$6,900.00
December	$19,300.00	$1,100.00

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

From the desk of the Publisher, FOODSTUFF DIGEST
TO: ____________________________

Participant No. b0501____

Participant Information

Please fill out:

How old are you? (Please state in years.) _____

Are you male ____ or female_____?

Your citizenship? ____________________________

For how many years have you been in this organization? _____

How many years have you been in your present position in this organization? _____

How many years of supervisory experience do you have? If you have previous experience, please include that also. _____

How many years of education have you had? ____ (where high school = 12 years; college =16)
Do you have an undergraduate degree? _____ If so, what was your major area(s) of study? ____
Do you have a graduate degree? _____ If so, what degree(s), and in what area(s)?
____

How many days of supervisory training have you had
in this organization? _____
in other organizations? _____

Please rate yourself on motivating others and technical expertise:

Place a check mark (✓) in the most suitable box. As in school, the highest rating is A and the lowest rating is F.

	Rating				
	A	B	C	D	F
Your skills in motivating others					
Your technical expertise					

CONFIDENTIAL

Appendix B

Excerpt from the Score Sheet for Item 3, "Old Form"

Score	Definition	Example	P1	P2	P3
Note: Score highest of all Ms, and of Ps.					
P1	s) simple, often 1 or 2 words, re: routine actions.	*Thanks for the update.*			
M6	r) (di**R**ect) Publisher w) monitor **W**ork sample: old and new forms t) **T**imely before making decision	*I'd like to see those forms so 1 can make an informed decision.* **OR** *Send me a copy of new & old form.* **OR** *I'll review forms with Mitch.*			
M5	r) w) "the" or "new" form t)	*Please send me a copy of the new form, ASAP. Thank you.* **OR** *I'll look at the form.* **OR** *Want to see form immediately. I will make a decision, and see how it compares to the others' input.*			
M2	l) monitor se**L**f report: at least one form t)	*Please come by my office so that we can discuss the aspects of the new form that make it unclear.* **OR** *Tell John in advertising – have him get with accounting to review. Schedule John & acctg with me when I return this PM.*			
M1a	l) new or old form u) Untimely after making decision	*To Acct.: I just received notification that the new reimbursement forms are a mess. Either have them revised or go back to the old form. They are a source of great confusion. Let's keep things as simple as possible please. Fax me a copy of the revised form. (u) I appreciate it! If you have any questions, please do not hesitate to call!*			
M1b	l) nothing more specific than problem/issue	*Please call me; I need to find out more about the problem.*			
Mni	ni) **N**ot **I**dentified above	*Please set up a meeting with acct. to review the new form with you.* (Not direct) *I want to use this as an example of how we can better serve our business leaders (like you) with our acct. processes and forms.*			
P4	P) **P**ositive consequence to person out of contact loop (incl. apology to Maria)				
P3a	b) To recognize **b**earer of bad news or o) For **o**ut of the ordinary action.	*Thanks for follow-up so that Maria got paid.* **Not P3a**:*Thanks*			

Score	Definition	Example	P1	P2	P3
P3b	njr) Nonjudgmental relay of potential problem to Accounting	To Accounting: *It appears that the new reimbursement form is causing some confusion for those filling them out. Please review the form and maybe we can discuss some changes.* To Accounting: *There seems to be some confusion over the new reimbursement form. Can we meet tomorrow about it for a few minutes?*			
P3c	cc) acknowledging receipt of message or letting sender know of action	To Dir. Of HR*: Please set up training class for new expense forms. Make sure everyone attends.* [second memo] To sender*: We think the new forms work. We're setting up training class to make sure you don't have any further problems.*			
Pni	**N**ot **I**dentified above				
P Alk	check if A-like or accompanied by A:				
	a) P+A, positive consequence (P) accompanied by A OR A-like P				
NU	u); N) neg. cons. about Acctg. (assuming problem with form) **<u>OR</u>** u); N) neg. cons. to sender/Art (assuming Art hasn't accepted new process, doesn't understand how to fill out form), not seen firsthand and thus **U**nverified **score reliance, sweeping, diplomacy, frequency, and form below**	To Dir. Of Accounting: *We need to improve the form to make it more user friendly. Give it a try and work with J. Rosso to get his input.* To Dir. Of Accounting: *Please issue a memo on instructions for filling out the new accounting reimbursement forms.*			
NUni	ni) not identified above				
NU rl	check if reliance:				
	reliance on NU: No M				
NU sw	check if sweeping:				
	sw) involving entire group or requiring training	To Dir. Of HR*: Please set up training class for new expense forms. Make sure everyone attends.* [second memo] *We think the new forms work. We're setting up training class to make sure you don't have any further problems.*			
NU dp	check the worst diplomacy (only 1):				
	d) diplomatic -- non-denigrating, non-punitive, non-sarcastic, dispassionate manner	To Dir. Of Accounting: *Please review the form and make it simpler. Please discuss with users prior to implementation. Thanks.*			
	h) harsh -- non-denigrating, non-punitive, possibly sarcastic or abrupt manner	To Dir. Of Accounting: L*ooks like someone needs to review this form! You know my pet project: Q.U.A.L.I.T.Y.*			
	I); fd) threaten to **F**ire/**D**emote OR b) **B**eat up, overly punitive				
	I); z) Zinger fire or demote on the spot				

Score	Definition	Example	P1	P2	P3
	I); ni) Not Identified above				
NU fq	check the frequency (only 1):				
	NU1) once about one topic or complete thought to one recipient	*OK! I'll ask to accounting department to change old form and thank you for your trouble.*			
	NU2) twice, about 2 topics, or to 2 recipients	*Raise problems with new form formally.* [1] [second memo] To Accounting: *No exceptions to new expense form!* [2]			
	NU3) 3 or more times, about 3 or more topics, to 3 or more recipients, or some combination of above				
NU fm	check all that apply for form:				
	n) simple negative consequence	To Accounting: *I would like to go over with you the new forms for getting contract personnel paid. There are seeral areas on the form that are unclear.*			
	a) antecedent-like	*Please arrange a training session to review the new accounting reimbursement forms. Let her know next time she won't get paid. Thanks for taking care of this.*			
	m) pseudo-monitor	*To Acct.: What's up with those forms? Why are they messy? (neg. C) Let's get some new, clearer forms made o.k.?*			
A	A) Antecedent; nonessential instructions				
A rl	check if reliance:				
	reliance on A: No M or P				
A fq	check the frequency (only 1):				
	A1) once about one topic or complete thought to one recipient				
	A2) twice, about 2 topics, or to 2 recipients				
	A3) 3 or more times, about 3 or more topics, to 3 or more recipients, or some combination of above				
NA	**N**ot **A**pplicable	*Please call PP today and see if you can resolve concerns. Advise we will see what changes can be made, and we'll meet with them next week to resolve concerns. This acct is critical. What changes can you accommodate to make our client's billing more clear*?			
NC	**N**ot **C**ompleted (did not finish exercise)				
BL	**Bl**ank, no response				
	Comments/additions				

Appendix C

Effectiveness Rating Form and Instructions

Effectiveness Rating: Instructions

We would like you to rate persons on two different things, one on effectively motivating others and the other on technical expertise. You will use grades in the U.S. school system, where A is the highest. Just record A, B, C, D, or F -- not A-, C+, etc. -- in the appropriate cell of the table.

You will also rank every person for whom you have rated the same and given the same grade. We call this ranking within rating. For example, if you have given 3 persons B's for effectively motivating others, we would like you to further differentiate them and rank them from 1 to 3 (with one being the highest rank). You will do the ranking within rating on the motivation grades only, not expertise.

Please rate and rank only those persons you have *DIRECTLY observed.* Do NOT score if you have heard about the person only second-hand through others. *Please assess only those persons you have seen in action.*

Here is an example of a hypothetical rating form:

Check if you have *DIRECTLY observed this person*				Effectively motivating others		Technical expertise	
Supervising you or others in his/her role as a manager or supervisor	Creating or helping to create deal-specific teams within or outside of the group	Supervising informally, perhaps providing coaching	Name	Rate (from A to F)	Rank w/in grade rating (1 = highest)	Rate (from A to F)	Check if you are/ were a super-visee
X	**X**	**X**	Person 1	**C**	**2**	**A**	**X**
			Person 2				
X			Person 3	**A**	**1**	**A**	**X**
	X		Person 4	**B**	**1**	**B**	
		X	Person 5	**B**	**3**	**C**	
	X	**X**	Person 6	**B**	**2**	**A**	
X	**X**		Person 7	**C**	**1**	**C**	
	X		Person 8	**D**	**1**	**A**	
			Person 9				
X	**X**		Person 10	**F**	**1**	**B**	

Effectiveness Rating Form

Check if you have *DIRECTLY observed this person*			Name	Effectively motivating others		Technical expertise	Check if you are/ were a super- visee
Supervising you or others in his/her role as a manager or supervisor	Creating or helping to create deal-specific teams within or outside of the group	Supervising informally, perhaps providing coaching		Rate (from A to F)	Rank w/in grade rating (1 = highest)	Rate (from A to F)	

Appendix D

Email from Managing Director

Sent: Wednesday, May 15, 2002 4:08 PM
Subject: Effectively Motivating Others

As part of *the group's* commitment to providing training and other development opportunities to its professionals, we have entered into an agreement with Dr. Judith Komaki at Baruch College in NYC, whereby she and her associates will provide personalized and confidential feedback to individual members based on their completion of an "in-box" exercise that she has developed for this purpose.

Judi Komaki has written extensively on management effectiveness in her 1998 book, Leadership from an Operant Perspective (London: Routledge) and conducted numerous studies on the principles that have been incorporated into this exercise. This project with *the group* will be incorporated into her research work.

To answer questions, Judi Komaki and her Ph.D. student, Michelle Minnich, will be available during a scheduled conference call:

CALL DATE: MAY-21-2002 (Tuesday)
CALL TIME: 10:00 AM EASTERN TIME (NY)
DURATION: 1 hr
USA Toll Free Number: 888-664-9860
USA Toll Number: +1-212-519-0802
PASSCODE: 17259

You are also welcome to contact Judi at Judith_Komaki@baruch.cuny.edu or Michelle at kevmic@fast.net.

We have selected a number of *the group's* professionals in certain locations in the Advisory business to take advantage of this opportunity. Persons with current management responsibilities as well as transactors are included. Your participation in this project is strictly voluntary. You will be neither favored nor disfavored for your decision to participate because we will not know your decision. But I would strongly encourage you to participate, as the individual coaching and feedback should assist you -- as a manager, a transactor, or both -- in developing your own managerial skills.

Regards,
Managing Director
Note: Please take a moment to read the following memo from Judi Komaki.

Memo To: Invited Project Participants
From: Judi Komaki

I am pleased to be able to do this work in *the group*. The work stems from a model of supervisory effectiveness I developed which identifies what effective managers actually do when supervising or motivating others. The model's behaviors are backed up by seven research studies. The managers we have observed include Australian police sergeants, Finnish construction supervisors, and U.S. sailboat skippers. In this research with you, we want to use a paper-and-pencil exercise that allows us to capture qualitative information that we cannot obtain through

observations alone. We also want to see whether the (remarkably consistent) results we have obtained thus far will be replicated with a global sample of financial managers and transactors.

You are invited to take an "in-box" exercise in which you will receive emails, memos, and reports and respond to them just like you would in your own office and to provide ratings of effectiveness in motivating others. The exercise takes from 60 to 90 minutes. Michelle Minnich, a Ph.D. candidate, and I will provide information about results of the model, the behaviors found to distinguish between effective and lackluster managers, and the behaviors of participants as a group on the "in-box."

Furthermore, we will provide confidential, individualized feedback to every participant. You will learn about your supervisory style, how you behave when you interact with others on the "in-box" exercise, what you do particularly well, and how you can improve to most effectively motivate others. Michelle and I will relay your feedback via videoconferencing or if possible in person. No information about you as an individual will be provided to the Firm. It will be given only to you for developmental purposes. Participants in other organizations have found the feedback we give insightful and sometimes illuminating. We do hope you will be willing to participate.

If you decide to participate, the first step is to complete a Consent Form (attached as "Informed Consent.doc") that outlines the research, explains that your scores and responses will be confidential and your participation strictly voluntary. In fact, we will not divulge whether you choose to participate or not. Please sign on the signature line, date it, and send it to me as soon as possible.

We would also like you to complete an Effectiveness Rating (attached as "Effectiveness Rating DUE May 29.doc"). In it, you will be asked to rate and rank how effective people in *the group* are at motivating others and on technical expertise. This will help us identify what effective supervisors in *the group* actually do. DUE DATE: May 29, 2002.

All information should be sent directly to me. No individual results/information will be disclosed to *the company*. Because of concerns about maintaining confidentiality, we are providing three ways for you to return the Consent Form and the Effectiveness Rating to us:

1. Email: Please transfer the attached Word documents to your hard drive, complete them, and forward them as attachments to me via email: Judith_Komaki@baruch.cuny.edu.
2. Print and fax to me at the college: +1-646-312-3781
3. Print and send via regular mail to:

   Judith Komaki

   Baruch College, Box B8-215

   1 Bernard Baruch Way

   New York, NY 10010.

Sending via regular mail is the most secure option, emailing the least secure. U.S. courts have ruled that companies may track and read the emails of employees sent on company email systems. This possibility, while a possibility, is unlikely. You are welcome to choose any of these options. We would be pleased to hear from you, in whatever way you wish to return the information to us.

We look forward to speaking with you during the conference call. You are also welcome to contact Judi at Judith_Komaki@baruch.cuny.edu or Michelle at kevmic@fast.net.

Appendix E

Group and Individual Feedback

Motivational Effectiveness Exercise (MEE)

The MEE focuses on what managers actually do when supervising or motivating others. It shows managers in action as they:

a. Convey expectations of performance (*Antecedents*):
"You can just write 'no comment' in there if you want to."
"Let's try and coordinate these two if we can."

b. Gather information about performance (*MONITORS*):
Listening to report given by subordinate.
Watching women work on CRTs.
"Did you set up the rehab audit?"

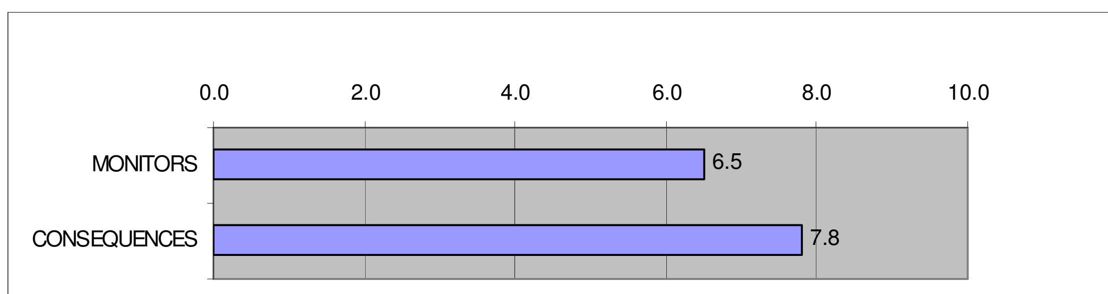
c. Provide compliments or corrections (*CONSEQUENCES*):
"I noticed you showed the 1008 in your report."
"You saved us from making a very big mistake."

Why Monitors and Consequences

The focus of the MEE is on Monitors and Consequences. It is not enough to just tell people what to do. It is important to do the next two steps, to find out what they do and to provide feedback. This emphasis is based on the theory of psychologist B. F. Skinner. And it is backed up by research studies, seven so far. The time managers spent gathering information and providing feedback made a difference in every situation, across lines of nationality: with American medical insurance managers, newspaper managers, and sailboat skippers in two settings, Finnish construction supervisors and government workers, and Australian police sergeants.

Group Results

The results are shown below for MONITORS and CONSEQUENCES. The scores can range from 0 to 10. A zero for Monitors means that no Monitors were given. Ten means that on occasions when it was appropriate, a Monitor was given and it was of the highest quality. The results for your group are shown below:



Your Results

Your results are shown below for Monitors and Consequences. In addition, we show you how well you gather information and provide feedback. For MONITORS, it is important to:

a. Observe workers in action and examine the products of their work (*Work Sample*):
(After receiving email about forwarding material) *Reply email to request a camera-ready copy before getting sent out. Given the typos in this email, I reasonably cannot trust that the mistakes are not his/her 'fault'.*

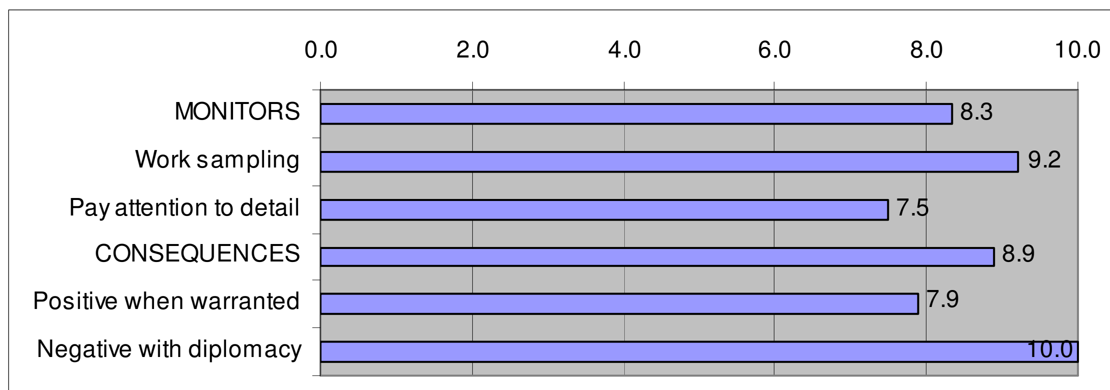
b. Make accurate observations and catch mistakes (*Pay Attention to Detail*):
(When shown a map and directions that contradict one another) *Please have a look at the instructions you sent. I believe you made a mistake providing wrong directions and this is why they arrived late.*

For CONSEQUENCES, it is critical to:

a. Recognize effort/excellence but only when deserved (Positive when Warranted):
(About person singled out by the auditor) *Please congratulate Anne on our behalf. Auditors were impressed with her. (P.S. Make sure they don't steal her away with a job offer!)*
(To person concerned with quality) *Thanks for the heads-up. You are right that the quality of those images are critical in achieving our vision of being the top gourmet industry mouthpiece in the country.*

b. Let persons know when they have erred or need to improve but do so in a way that is not sarcastic, abrupt, or demeaning (*Negative with Diplomacy*)
(To Advertising about tabulation mistake) *Year 2000: either the numbers are wrong or I'm missing something…*

Participant 15



For further information, you can refer to Prof. Komaki's book, Leadership from an operant perspective. London: Routledge, 1998. Or you can contact her: Judith_Komaki@baruch. cuny.edu or +1-646-312-3817.

BIBLIOGRAPHY

Adsit, D.J., Crom, S., Jones, D., & London, M. (1994). Management performance from different perspectives. *Journal of Managerial Psychology, 9(2),* 22-29.

Anderson, L.R. & Thacker, J. (1985). Self-monitoring and sex as related to assessment center ratings and job performance. *Basic and Applied Social Psychology, 64*, 345-361.

Arthur, W., 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*, 125-154.

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.

Avolio, B.J. O'Connell, M.S., Maritz, D., & Kennedy, C.J. (1999, April). Assessment of transformational leadership potential. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Barrick, M.R. & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M.R., Mount, M.K. & Judge, T.A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and Assessment, 9*, 9-30.

Bass, B.M. (1990). *Bass and Stogdill's handbook of leadership* (3rd ed.). New York, NY: Free Press.

Bass, B.M. (1985). Leadership: Good, better, best. *Organizational Dynamics, 13(3)*, 26-40.

Bayazit, M., Komaki, J.L., Redding, T., Haimovich, T., & Kalt, A. (2000). *Observing subordinates' reactions to supervisors' monitoring and pseudo-monitoring*. Poster presented at the conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Bernardin, H.J. & Alvares, K.M. (1976). The managerial grid as a predictor of conflict resolution method and managerial effectiveness. *Administrative Science Quarterly, 21*, 84-92.

Binning, J.F., Adorno, A.J., & LeBreton, J.M. (1999, April). "Sociotechnical" moderators of assessment center criterion-related validity. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Blake, R. & Mouton, J. (1964). *The managerial grid*. Houston: Gulf.

Blalock, H.M., Jr. (1960). *Social statistics* (2nd ed.). New York, NY: McGraw-Hill, Inc.

Bobko, P., Roth, P.L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

Bobrow, W. & Leonards, J.S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality, 12*, 217-236.

Borman, W.C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67,* 3-9.

Borman, W.C. & Brush, D.H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6(1),* 13-21.

Brannick, M.T., Michaels, C.E., & Baker, D.P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*, 957-963.

Brass, D.J. & Oldham, G.R. (1976). Validating an in-basket test using an alternative set of leadership scoring dimensions. *Journal of Applied Psychology, 61*, 652-657.

Bray, D.W. & Campbell, R.J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology, 52,* 36-41.

Bray, D.W., Campbell, R.J., & Grant, D.L. (1974). *Formative years in business: A long-term AT&T study of managerial lives.* New York, NY: Wiley.

Brewer, N. (1995). The effects of monitoring individual and group performance on the distribution of effort across tasks. *Journal of Applied Social Psychology, 25,* 760-777.

Brewer, N. & Ridgway, T. (1998). Effects of supervisory monitoring on productivity and quality of performance. *Journal of Experimental Psychology: Applied, 4,* 211-227.

Brewer, N., Wilson, C., & Beck, K. (1994). Supervisory behavior and team performance amongst police patrol sergeants. *Journal of Occupational and Organizational Psychology, 67,* 69-78.

Burns, J.M. (1984). *The power to lead*. New York: Simon & Schuster.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. (1970). *Managerial behavior, performance, and effectiveness.* New York, NY: McGraw-Hill.

Cascio, W.F. (1987). *Applied psychology in personnel management.* Englewood Cliffs, NJ: Prentice-Hall.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69,* 167-181.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgoup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Conger, A.J. & Jackson, D.N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement, 32*, 579-599.

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.

Cortina, J.M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management, 19*, 915-922.

Cortina, J.M., Goldstein, N.B., Payne, S.C., Davison, H.K., & Gilliland, S.W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*, 325-351.

Daniels, A.C. (1994). *Bringing out the best in people: How to apply the astonishing power of positive reinforcement.* New York, NY: McGraw-Hill, Inc.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly?. *Personnel Psychology, 55*, 827-849.

Evans, M.G. (1996). R.J. House's 'a path-goal theory of leader effectiveness'. *Leadership Quarterly, 7,* 305-309.

Fiedler, F.E. & House, R.J. (1988). Leadership: A report of progress. In C. Cooper (Ed.), *International review of industrial and organizational psychology* (pp. 73-92). Greenwhich, CT: JAI Press.

Fleishman, E.A. (1953). The description of supervisory behavior. *Personnel Psychology, 37*, 1-6.

Fleishman, E.A. (1973). Twenty years of consideration and structure. In E.A. Fleishman and J.G. Hunt (Eds.), *Current developments in the study of leadership* (pp. 1-37). Carbondale: Southern Illinois Universtiy Press.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology,72,* 493-511.

Ganzach, Y. (1998). Nonlinearity, multicollinearity and the probability of type II error in detecting interaction. *Journal of Management, 24,* 615-622.

Gill, R.W.T. (1979). The in-tray (in-basket) exercise as a measure of management potential. *Journal of Occupational Psychology, 52*, 185-197.

Goffin, R.D., Rothstein, M.G., & Johnston, N.G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81,* 746-756.

Goldstein, H.W., Yusko, K.P., Braverman, E.P, Smith, D.B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51,* 357-374.

Goldstein, H.W., Yusko, K.P., & Nicolopoulos, V. (2001). Exploring black-white subgroup differences of managerial competencies. *Personnel Psychology, 54,* 783-807.

Goldstein, H.W., Zedeck, S., & Goldstein, I.L. (2002). *g*: Is this your final answer?. *Human Performance, 15,* 123-142.

Guion, R.M. (1990). Personnel assessment, selection, and placement. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (vol. 2, pp. 327-397). Palo Alto, CA: Consulting Psychologists Press, Inc.

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hakstian, A.R. & Scratchley, L.S. (1997). In-basket assessment by fully objective methods: Development and evaluation of a self-report system. *Educational and Psychological Measurement, 57*, 607-630.

Hemphill, J.K. (1959). Job descriptions for executives. *Harvard Business Review, 37,* 55-67.

Hinrichs, J.R. (1969). Comparison of "real life" assessments of management potential with situational exercises, paper-and-pencil ability tests, and personality inventories. *Journal of Applied Psychology, 53,* 425-432.

Hinrichs, J.R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63,* 596-601.

Honig, W.K. (Ed.) (1966). *Operant behavior: Areas of research and application.* New York, NY: Appleton-Centruy-Crofts.

Honig, W.K. & Stadden, J.E.R. (1977). *Handbook of operant behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Horst, P. (1941). The role of the predictor variables which are independent of the criterion. *Social Science Research Council Bulletin, 48*, 431-436.

Horst, P. (1966). *Psychological measurement and prediction*. Belmont, CA: Wadsworth.

House, R.J. (1971). A path-goal theory of leader effectiveness. *Administrative Science Quarterly, 16,* 321-339.

House, R.J. (1977). A 1976 theory of charismatic leadership. In J. G. Hunt & L. L. Larson (Eds.), *Leadership: The cutting edge* (pp. 189-207). Carbondale: Southern Illinois University Press.

House, R.J. (1996). Path-goal theory of leadership: Lessons, legacy, and a reformulated theory. *Leadership Quarterly, 7,* 323-352.

House, R.J. & Aditya, R.N. (1997). The social scientific study of leadership: Quo vadis?. *Journal of Management, 23,* 409-473.

House, R.J. & Mitchell, T.R. (1974). Path-goal theory of leadership. *Journal of Contemporary Business, 3,* 81-97.

House, R.J. Spangler, W.D., & Woycke, J. (1991). Personality and charisma in the US presidency: A psychological theory of leadership effectiveness. *Administrative Science Quarterly, 36,* 364-396.

Huck, J.R. & Bray, D.W. (1976). Management assessment center evaluations and subsequent job performance of black and white females. *Personnel Psychology, 29,*13-30.

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

Inkson, K. & Moss, A.T. (1993). Transformational leadership – Is it universally applicable? *Leadership and Organization Development Journal, 14 (4)*, i-ii.

Jensen, M. & Komaki, J. L. (1993, May). Beware of too many directives: Spotting a suppressor variable in the operant model of effective supervision. In J. Cannon-Bowers (Chair), *Optimizing team performance through team leader behavior.* Symposium conducted at the conference of the Society of Industrial and Organizational Psychology, San Francisco, CA.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment centre. *Journal of Occupational Psychology, 64*, 1-21.

Jones, R.G. & Whitmore, M.D. (1995). Evaluating developmental assessment centers as interventions. *Personnel Psychology, 48*, 377-388.

Judge, T.A., Piccolo, R.F., & Ilies, R. (2004). The forgotten ones?: The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology, 89*, 36-51.

Kerr, S. & Schriesheim, C. (1974). Consideration, initiating structure, and organizational criteria – An update of Korman's 1966 review. *Personnel Psychology, 27,* 555-568.

Kim, H. & Yukl, G. (1995). Relationships of managerial effectiveness and advancement to self-reported and subordinate-reported leadership behaviors from the multiple-linkage model. *Leadership Quarterly,6,* 361-377.

Komaki, J.L. (1986). Toward effective supervision: An operant analysis and comparison of managers at work. *Journal of Applied Psychology, 71,* 270-278.

Komaki, J.L. (1998). *Leadership from an operant perspective.* London: Routledge, Chapman, and Hall, Inc.

Komaki, J.L., Coombs, T., Redding, Jr., T.P., & Schepman, S. (2000). A rich and rigorous examination of applied behavior analysis research in the world of work. In C.L. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, Vol. 15* (pp. 265-367). Sussex, England: John Wiley & Sons Ltd.

Komaki, J.L., Desselles, M.L., & Bowman, E. (1989). Definitely not a breeze: Extending an operant model of effective supervision to teams. *Journal of Applied Psychology, 74,* 522-529.

Komaki, J.L., Desselles, M.L., & Schepman, S. (1988, October). Testing critical linkages in the operant model of effective supervision: Assessing subordinate performance and attitudes. Paper presented at the conference of the Society of Organizational Behavior, Worthington, OH.

Komaki, J.L., Hyttinen, M., & Immonen, S. (1991, August). Cross-cultural research in work and leader behavior. In K.R. Thompson (Chair), *Managing across cultures from a behavioral perspective: Theory and research.* Symposium presented at the meeting of the Academy of Management, Atlanta.

Komaki, J.L., Newlin, M.H., & Desselles, M.L. (1990, April). Walking on the wild side: Criterion-related validation of an in-basket exercise of supervisory behaviors. In M.H. Newlin (Chair), *Simulated performance assessment: Fact or fantasy?.* Symposium presented at the conference of the Society for Industrial and Organizational Psychology, Miami, FL.

Komaki, J.L., Reynard Minnich, M., Lee, C., & Wallace, J. (1997, May). Apples and oranges at sea: Skippers' leadership styles make a difference. In B.P. Niehoff (Chair), *Research on performance monitoring: Refinements and boundaries.* Symposium presented at the conference of the Society for Industrial and Organizational Psychology, St. Louis.

Komaki, J.L., Zlotnick, S., & Jensen, M. (1986). Development of an operant-based taxonomy and observational index of supervisory behavior. *Journal of Applied Psychology, 71*, 260-269.

Kraut, A.I. & Scott, G.J. (1972). Validity of an operational management assessment program. *Journal of Applied Psychology, 56,* 124-129.

Landy, F.J. (1989). *Psychology of work behavior.* Pacific Grove, CA: Brooks/Cole Publishing Co.

Larson, J.R. & Callahan, C. (1990). Performance monitoring: How it affects work productivity. *Journal of Applied Psychology, 75*, 530-538.

Larson, L.L., Hunt, J.G., & Osborn, R.N. (1976). The great hi-hi leader behavior myth: A lesson from Occam's razor. *Academy of Management Journal, 19*, 628-641.

Lefkowitz, J., Gebbia, M.I., Balsam, T., & Dunn, L. (1999). Dimensions of biodata items and their relationships to item validity. *Journal of Occupational and Organizational Psychology, 72,* 331-350.

Licht, M.H. (1995). Multiple regression and correlation. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 19-64). Washington, DC: American Psychological Association.

Lubin, A. (1957). Some formulae for use with suppressor variables. *Educational and Psychological Measurement, 17,* 286-296.

McCloy, R. & Putka, D. (2004, April). Estimating interrater reliability: Conquering the messiness of real-world data. Tutorial session presented at the conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology. 86,* 730-740.

McDaniel, M.A. & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.

McEvoy, G.M. & Beatty, R.W. (1989). Assessment centers and subordinate appraisals of managers: A seven-year examination of predictive validity. *Personnel Psychology, 42,* 37-52.

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Methot, L.L., Williams, W.L., Cummings, A., & Bradshaw, B. (1996). Measuring the effects of a manager-supervisor training program through the generalized performance of managers, supervisors, front-line staff and clients in a human service setting. *Journal of Organizational Behavior Management, 16(2),* 3-34.

Miller, L.K. (1997). *Principles of everyday behavior analysis* (3rd ed.). Belmont, CA: Wadsworth.

Mintzberg, H. (1975, July-August). The manager's job: Folklore and fact. *Harvard Business Review, 53*, 49-61. Reproduced in J.R. Hackman, E. Lawler, & L. Porter, (Eds.), *Perspectives on behavior in organizations,* 1983, (pp. 5-15). New York, NY: McGraw-Hill.

Mitchel, J.O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573-579.

Moses, J.L. & Boehm, V.R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology, 60,* 527-529.

Motowidlo, S.J., Dunnette, M.D., & Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology. 75,* 640-647.

Mount, M.K., Barrick, M.R., & Stewart, G.L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145-165.

Muchinsky, P.M. (1986). Personnel selection methods. In C.L. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 37-70). Sussex, England: John Wiley & Sons Ltd.

Ones, D.S. & Viswesvaran, C. (2001). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. *International Journal of Selection and Assessment, 9*, 31-39.

Ones, D.S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Outtz, J.L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15*, 161-171.

Phillips, J.J. (1985). *Improving supervisors' effectiveness.* San Francisco, CA: Jossey-Bass.

Podsakoff, P.M., MacKenzie, S.B., Ahearne, M., & Bommer, W.H. (1995). Searching for a needle in a haystack: Trying to identify the illusive moderators of leadership behaviors. *Journal of Management, 21,* 423-470.

Podsakoff, P.M. & Schriesheim, C.A. (1985). Leader reward and punishment behavior: A methodological and substantive review. Unpublished manuscript. Cited in B.M. Bass (Ed.), *Bass and Stogdill's handbook of leadership*. New York: Free Press.

Pool, S.W. (1997). The relationship of job satisfaction with substitutes of leadership, leadership behavior, and work motivation. *The Journal of Psychology, 131*, 271-283.

Pulakos, E.D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*, 241-258.

Reynard, M. & Komaki, J.L. (1995, May). *The music of management: Rhythms of effective leaders*. Poster presented at the conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Rifkin, G. (1994, February 27). It's not IBM and it's not big, but he's got big ideas. *The New York Times,* p. 8.

Ritchie, R.J. (1994). Using the assessment center method to predict senior management potential. *Consulting Psychology Journal, 46*, 16-23.

Ritchie, R.J. & Moses, J.L. (1983). Assessment center correlations of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68,* 227-231.

Robertson, I.T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74,* 441-472.

Russell, C.J. & Domm, D.R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology, 68,* 25-47.

Salgado, J.F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*, 323-346.

Sayles, L. (1979). *Leadership: What effective managers really do… and how they do it.* New York, NY: McGraw-Hill.

Schippmann, J.S., Prien, E.P., & Katz, J.A. (1990). Reliability and validity of in-basket performance measures. *Personnel Psychology, 43,* 837-859.

Schleicher, D.J., Day, D.V., Mayes, B.T., & Riggio, R.E. (2002). A new frame of reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746.

Schmidt, F.L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33*, 272-292.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407-422.

Schmitt, N., Noe, R.A., Meritt, R., & Fitzgerald, M.P. (1984). Validity of assessment center ratings for the prediction of performance ratings and school climate of school administrators. *Journal of Applied Psychology, 69,* 207-213.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719-730.

Schmitt, N., Schneider, J.R., & Cohen, S.A. (1990). Factors affecting validity of a regionally administered assessment center. *Personnel Psychology, 43,* 1-11.

Schriesheim, C.A. (1982). The great high consideration-high initiating structure leadership myth: Evidence on its generalizability. *The Journal of Social Psychology,116*, 221-228.

Schriesheim, C.A. & Neider, L.L. (1996). Path-goal leadership theory: The long and winding road. *Leadership Quarterly, 7*, 317-321.

Scott, W.E. & Podsakoff, P.M. (1985). *Behavioral principles in the practice of management*. New York, NY: John Wiley & Sons.

Shore, T.H., Shore, L.M. & Thornton, G.C. III. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42-54.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Skinner, B.F. (1974). *About behaviorism.* New York, NY: Vintage.

Smith, R.L, Ager, J.W., Jr., & Williams, D.L. (1992). Suppressor variables in mulitple regression/correlation. *Educational and Psychological Measurement, 52,* 17-29.

Tett, R.P., Guterman, H.A., Bleier, A., & Murphy, P.J. (2000). Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance, 13*, 205-251.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-741.

The City of New York Department of Citywide Administrative Services (2002, February 6). *Notice of Examination: Promotion to Lieutenant (Fire Department), Exam No. 0540, Amended Notice.* Retrieved March 30, 2005 from http://www.nyc.gov/html/dcas/pdf/plieutenant_fdny.pdf.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score Reliability: Contemporary Thinking on Reliability Issues* (pp. 3-23). Thousand Oaks, CA: Sage Publications, Inc.

Tichy, N.M. & Ulrich, D.O. (1984). The leadership challenge – a call for the transformational leader. *Sloan Management Review, 26(1),* 59-68.

Tornow, W.W. & Pinto, P.R. (1976). The development of a managerial job taxonomy: A system for describing, classifying and evaluating executive positions. *Journal of Applied Psychology, 61,* 410-418.

Turnage, J.J. & Muchinsky, P.M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69,* 595-602.

Tzelgov, J. & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109,* 524-536.

Tziner, A. & Dolan, S. (1982). Validity of an assessment center for identifying female officers in the military. *Journal of Applied Psychology, 67,* 728-736.

Tziner, A., Ronen, S., & Hacohen, D. (1993). A four-year validation of an assessment center in a financial corporation. *Journal of Organizational Behavior, 14,* 225-237.

Whaley, D.L. & Malott, R.W. (1971). *Elementary principals of behavior*. New York: Appleton-Century-Crofts.

Wofford, J.C. & Liska, L.Z. (1993). Path-goal theories of leadership: A meta-analysis. *Journal of Management, 19,* 857-876.

Wollowick, H.B. & McNamara, W.J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53,* 348-352.

Worbois, G.M. (1975). Validation of externally developed assessment procedures for identification of supervisory potential. *Personnel Psychology, 28*, 77-91.

Yukl, G. (1994). *Leadership in organizations* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Yukl, G. (1998). *Leadership in organizations* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Yukl, G. & Van Fleet, D.D. (1990). Theory and research on leadership in organizations. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (vol. 3, pp. 147-197). Palo Alto, CA: Consulting Psychologists Press, Inc.

Yukl, G., Wall, S., & Lepsinger, R. (1990). Preliminary report on validation of the managerial practices survey. In K.E. Clark & M.B. Clark (Eds.), *Measures of leadership* (pp. 223-238). West Orange, NJ: Leadership Library of America.