EFFECTS OF RATER CHARACTERISTICS
ON VALIDITY COEFFICIENTS

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

By

KATE VAN BREMEN

M.S., Wright State University 2006

2016
Wright State University
I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Kate Van Bremen ENTITLED Effects of Rater Characteristics on Validity Coefficients BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

David LaHuis, Ph.D.
Dissertation Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D., Chair
Department of Psychology

Robert E. W. Fyffe, Ph.D.
Vice President for Research and Dean of the Graduate School

David LaHuis, Ph.D.

Nathan Bowling, Ph.D.

Valerie Shalin, Ph.D.

Gary Burns, Ph.D.
ABSTRACT

Van Bremen, Kate, Ph.D., Department of Psychology, Wright State University, 2016. The Effects of Rater Characteristics on Validity Coefficients

Researchers and practitioners look to validity coefficients as indicators of the relationship between predictors and a measure of performance, typically as rated by a supervisor. Previous research suggests that subjective ratings of performance are problematic as they contain variance unrelated to subordinate performance. Though previous research suggested variance in supervisor ratings was random and unsystematic, recent literature suggests distortions may be deliberate and goal-directed. A potential implication of systematic variance in job performance ratings is a possible effect on validity coefficients. In the present study, I used multilevel modeling to determine if supervisor level characteristics influence the average ratings as well as the validity of job performance predictors for entry-level manufacturing incumbents. Results indicate that supervisor intelligence and level of adaptability had a significant relationship with the average performance rating they assigned. Though average ratings were not influenced, as the number of subordinates managed increased the validity of predictors Positive Attitude, Process Monitoring, and Responsibility decreased; and as the length of time managing increased, the validity of the predictor Attention to Detail increased significantly.
# TABLE OF CONTENTS

## I. EFFECTS OF RATER CHARACTERISTICS ON VALIDITY

<table>
<thead>
<tr>
<th>Predictors of Job Performance</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Ability</td>
<td>5</td>
</tr>
<tr>
<td>Non-Cognitive Predictors</td>
<td>8</td>
</tr>
<tr>
<td>Job Performance</td>
<td>11</td>
</tr>
<tr>
<td>Supervisor Ratings of Job Performance</td>
<td>13</td>
</tr>
<tr>
<td>Validity</td>
<td>15</td>
</tr>
<tr>
<td>Objective versus Subjective Criteria</td>
<td>15</td>
</tr>
<tr>
<td>Systematic Variance in Validity Coefficients</td>
<td>17</td>
</tr>
<tr>
<td>Common Rater Errors</td>
<td>18</td>
</tr>
<tr>
<td>Goal-Directed Rater Effects</td>
<td>20</td>
</tr>
</tbody>
</table>

## II. METHODS

<table>
<thead>
<tr>
<th>Participants</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures</td>
<td>27</td>
</tr>
<tr>
<td>Procedure</td>
<td>29</td>
</tr>
<tr>
<td>Analyses</td>
<td>30</td>
</tr>
<tr>
<td>Results</td>
<td>32</td>
</tr>
<tr>
<td>Discussion</td>
<td>37</td>
</tr>
<tr>
<td>References</td>
<td>46</td>
</tr>
</tbody>
</table>
INDEX OF TABLES

Table 1 Means, Standard Deviations, and Intercorrelations of Subordinate Study Variables

Table 2 Means, Standard Deviations, and Intercorrelations of Supervisor Study Variables

Table 3 Estimated Variance Components (Random Effects)

Table 4 Supervisor Competencies Predicting the Intercept

Table 5 Effect of Supervisor-Level Competencies on Level One Slope
Effects of Rater Characteristics on Validity Coefficients

One of the major goals in I/O psychology, academically and in practice, is to predict job performance. Researchers and practitioners alike look to validity coefficients as an indication of how effective variables are at predicting a variety of outcomes. Specifically, in selection, researchers compare correlations between a range of predictors and the criteria job performance, typically as rated by a supervisor.

Over the past 90 years researchers have analyzed a number of potential predictors of job performance from cognitive ability to level of emotional competence (Schmidt and Hunter, 1998; Jayan, 2006) in an attempt to determine the core factors that differentiate those who succeed on the job from those who do not. Similarly, much attention has been devoted to the best way to measure job performance. In most validation studies, job performance is measured using supervisors’ ratings of subordinate performance. These ratings are thought to be subject to numerous rating errors and, as such, researchers have investigated how to eliminate these errors. For example, studies have targeted leniency (Murphy and Cleveland, 1995), halo (Thorndike, 1920), and similarity (Schmitt, Pulakos, Nason and Whitney, 1996) biases. Recently, researchers have suggested that there may be systematic supervisor effects on ratings that reflect goal-oriented behavior and not errors (Murphy and Cleveland, 1995; Murphy & DeShon, 2000). Subsequent research has supported the idea that some variance in supervisors’ performance ratings is systematic and goal oriented (Tziner, Murphy, and Cleveland, 2005).

Researchers have also given considerable attention to how to best estimate the relationship between a predictor and job performance. The focus is usually on how to correct for imperfection in job performance ratings. For example, validity coefficients are often corrected
for range restriction in the predictor and/or the criterion. In many cases, coefficients are corrected for unreliability in the criterion. There is some debate as to the best way to correct for unreliability. The debate centers around whether or not errors in performance ratings are viewed as random or systematic measurement errors (Murphy and DeShon, 2000a; Schmidt, Viswesvaran, and Ones, 2000).

Schmidt and colleagues (2000) contend that the classical measurement theory is an appropriate method to correct for unreliability in the criterion because, they contend, it accounts for the error in ratings. However, the classical model assumes that rater error has certain psychometric properties. Research has shown that rater error does not meet those properties. Specifically, the error is not uncorrelated and unsystematic. For example, arguably the most pervasive rater idiosyncrasy (error) is leniency (Murphy and Cleveland, 1995), which suggests that raters are consistently, or systematically more lenient to employees across the board.

One potential unexplored implication of systematic variance in performance ratings is the possibility that this variance would lead to systematic variance in validity coefficients across supervisors. That is, it is possible that there are supervisor differences in the relationships between predictors and performance ratings and that these differences may be explained using supervisor-level variables. In the present study, I suggest that there are contextual factors which account for systematic (non-performance related) variance in job performance ratings, and I will use multilevel modeling to determine if supervisor level variables significantly impact validity coefficients. Specifically, I contend that the rater’s cognitive ability, interpersonal skills, level of cognitive demands, and their “promotability” will significantly influence the criterion-related validity of the subordinate predictors of cognitive ability, Conscientiousness and interpersonal skills.
In the following sections I will discuss the most common predictors of job performance, including cognitive and non-cognitive predictors and their relative validity. I will also discuss job performance more broadly, reviewing researchers’ attempts to define and model job performance throughout the literature. Then I will more specifically examine supervisor ratings of job performance and review some strategies historically employed to reduce some of the biases and error associated with these subjective ratings. This “error” in the criterion leads to the topic of validity. In this section I will consider the importance of criterion-related validity in general and the effect of different factors on this relationship. Next, I will discuss systematic variance in validity coefficients and cover some common rater errors as well as goal-directed rater effects. Finally, I will conclude with hypotheses about possible effects of rater characteristics on validity coefficients.

Predictors of Job Performance

One major goal in I/O psychology is to predict future job performance in an attempt to determine which candidates will be successful on the job and which will not (Hunter, Schmidt, & Judiesch, 1990). Throughout the years a myriad of predictors have been proposed and analyzed. To date, the most valid predictor of job performance is cognitive ability (Schmidt and Hunter, 1998). However, other non-cognitive predictors have been shown to account for incremental variance in job performance over and above cognitive ability.

When evaluating the effectiveness of a selection tool or system, we often cite its validity coefficient. Essentially, the validity coefficient tells us the correlation between scores on the selection tool and corresponding measures of job performance. Criterion-related validity can be established by two different methods: Predictive and/or concurrent validation. In predictive validation, the selection tool is administered to candidates as part of the application process.
Once those successful candidates have been selected, become employees, and have been on the job a sufficient amount of time to demonstrate performance (generally at least six months), performance ratings or some relevant outcome measure is collected. A validity coefficient is calculated by evaluating the relationship between individuals’ scores on the selection tool and their corresponding performance ratings. In concurrent validation, the selection tool is administered to current incumbents who have already been selected and have been performing in the job. The selection tool may be administered to current employees and the performance ratings may be collected concurrently as the incumbents have already demonstrated performance on the job. As in the predictive design, the relationship between the selection tool and the performance outcome is evaluated and a correlation coefficient is calculated.

There are pros and cons to both methods. In a predictive design applicants who are competing for a job are completing the assessment, so they are likely motivated to perform to the best of their ability. However, because only the best candidates are selected the range has been restricted and there is less variance in predictor scores, undermining the correlation coefficient. Another drawback to the predictive design is the length of time needed to determine the relationship. Applicants complete the assessment and must be selected and perform on the job for a sufficient amount of time to demonstrate performance before the criterion measure can be collected. This may take six months to several years. Concurrent validity can be established quickly because the predictor and criterion measures can be collected concurrently as incumbents have been performing in the job and performance can be evaluated. However, scores on the predictor are likely contaminated. Because they already hold the position, they are likely not as motivated to perform well on the measure. Further, strong job performers may want to get back to their job tasks quickly, so they may rush through the assessment. Those incumbents who have
poor job performance may recognize that and try to perform well on the measure. There is also range restriction in a concurrent study. Variance is limited because incumbents are very similar: They have all passed through the same selection system, remained in the position, and have not turned over (voluntarily or involuntarily). The result is a relatively homogeneous group of individuals and there is likely less variance in both the predictor and criterion measure.

**Cognitive ability**

Schmidt and Hunter (1998) compared 19 different selection procedures as predictors of job performance and the validity coefficients of each, based on meta-analyses conducted over the previous 85 years. They determined that, when selecting employees without prior experience in that job, cognitive ability is the best predictor of future job performance with a validity coefficient of .51.

Furthermore, Schmidt and Hunter (1998) site four main reasons why using cognitive ability as part of a selection process is so valuable. The first reason they site is the notion that cognitive ability assessments are especially useful because they predict future performance across all jobs and for all levels. Further, of all the possible selection procedures that meet this standard, cognitive ability assessments are the most valid and have the lowest cost associated. Second, they cite the vast amount of studies conducted on the cognitive ability – job performance relationship throughout the literature and conclude that the evidence of this relationship is stronger than the evidence for any other predictor. Third, they assert that cognitive ability as a predictor of performance is so critical because it also predicts job-related learning; the research suggests that cognitive ability is the best predictor of the acquisition job knowledge and on-the-job training performance. Finally, they reference the theoretical foundation of cognitive ability in general. They cite the more than 90 decades of research clarifying the construct and conclude
that, as a result, we have significantly more information about cognitive ability than a more recently introduced construct, such as an assessment center.

Indeed, cognitive ability testing has a rich and varied history. The beginning of cognitive ability testing used for personnel decisions is traditionally traced back to World War I. Oral trade tests were among the first performance-based assessments used in selection (Guion, 1998, Chapman, 1921). During this time many candidates were unable to read, making it necessary to use oral tests to assess trade knowledge. It was necessary to use these oral tests to assess an applicant’s knowledge of a specific trade, as their self-reported level of proficiency in the trade was often unreliable.

Even in trade tests’ infancy, their test development was rigorous and job content experts created psychometrically strong assessments. Technical content experts created and regularly reviewed questions and (correct) answers for the specific trade tests (Guion, 1998; Chapman, 1921). In addition, practitioners and test administrators recognized the necessity of standardization. The Administration and Scope of the Oral Trade Test includes highly specific instructions for giving oral trade tests: “In order that the Oral Trade Tests be used effectively, it is necessary that examiners follow to the letter these ‘Instructions for Giving Oral Trade Tests.’ Although the tests have been carefully prepared, they will not give uniform results unless examiners use them uniformly. Consequently, no change must be made in the tests themselves, or in the manner of administering them, until official notice is given to that effect.” (p. 160 Administration and Scope of Oral Trade Tests, Instructions for giving oral trade tests, Chapman, 1921). Guion (1998) cites that, though “paper and pencil” tests are commonly used now, tests must continue to be well-standardized, items reliably scored, and be able to be administered to groups of people.
Though most researchers do not dispute, even the historical validity of cognitive ability as a predictor of job performance, relatively recent literature has reviewed the incremental validity of specific abilities over and above a general measure of intelligence, or “g” (Carretta and Ree, 2000). Carretta and Ree cited early researcher Charles Spearman who suggested all cognitive assessments measured general cognitive ability (g) and another specific ability. He purported that all cognitive assessments measured general ability and the specific ability measured depended on the test.

The research that followed Spearman suggested that there were equal, independent abilities (Thurstone 1938). Thurstone argued that intelligence was not one overarching factor but that it was comprised of seven unique Primary Abilities: word fluency, verbal comprehension, spatial visualization, number facility, associative memory, reasoning, and perceptual speed. In his later research, Thurstone was unsuccessful in finding his seven independent abilities and found support for and evidence of g. Ultimately, he revised his model of the factors of intelligence to include his seven mental abilities and g (Thurstone and Thurstone, 1941).

More recent research is a compromise of Spearman’s and Thurstone’s later findings and suggests that a higher-order ability source accounts for the majority of the variance in assessments, relative to specific, lower-order factors (e.g., verbal, math) (Carretta and Ree, 2000). In his review of cognitive ability testing and measurement, Guion (1998) suggests that despite concerted efforts to develop highly specific assessments, more general cognitive ability continues to account for most of the variance. Further, in their review of multiple ability tests, Carretta and Ree (2000) found that g accounted for 35 – 56% of the variance and that the most variance accounted for by any specific factor was only 8%.
As a result of its inherent value in the selection system, many organizations use measures of general mental ability to some extent in their selection process. One caveat, however, is that cognitive ability assessments have been shown to have adverse impact against protected groups. A selection tool is said to have adverse impact if the use of it to screen applicants results in different hiring rates for different (protected) groups. In order to mitigate this impact against minorities, organizations often use cognitive ability scores as simply one data point and consider this in combination with other measures with less adverse impact to gain additional information to predict performance.

**Non-Cognitive Predictors**

Similar to cognitive ability, there is a long history of personality assessment in selection. Early, researchers concluded there was little evidence to support the validity and utility of using personality measures in selection. Specifically, Guion and Gottier (1965) examined research published in the *Journal of Applied Psychology* and *Personnel Psychology* between 1952 and 1963. In their article summarizing the research on the predictive validity of personality assessments, they concluded, “It cannot be said that any of the conventional personality measures have demonstrated really general usefulness as selection tools in employment practice (p. 140).”

In response to this discouraging statement, scientists researched a myriad of different personality traits as predictors and a variety of performance criteria across different occupations in an effort to demonstrate useful, predictive validity of personality in personnel selection. Because of the variance in predictors, criteria, and occupations examined, nominal, sometimes negative correlations were found and most researchers acquiesced agreeing personality measures
were not as predictive of performance as they had originally maintained (e.g., Schmitt, Gooding, Noe, and Kirsch, 1984).

Over time, meta-analytic methods and techniques improved. In addition, the “five factor model” (FFM) became the widely accepted classification of personality into five characteristics: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Subsequently, by the 1990’s meta-analytic studies began to show much stronger relationships between personality and job performance. For example, Barrick and Mount (1991) used the Five Factor Model taxonomy of personality and examined the predictive relationship with job performance across multiple occupations and criteria. Specifically, they looked at job proficiency, training proficiency, and personnel data as outcome measures across five different job families: professional, police, managers, sales, and skilled/semi-skilled. Barrick and Mount found an estimated true score correlation between the FFM dimensions of personality and performance across job families and criterion types ranged from .04 for Openness to Experience to .22 for Conscientiousness.

After Barrick and Mount’s meta-analysis, Conscientiousness was often studied as a predictor of job performance. Individuals high on this construct are said to be dependable and achievement striving (Costa and McCrae, 1988). Accordingly, including a measure of this personality construct makes intuitive sense as conscientious people are likely to be better performers on the job than their less conscientious counterparts. Indeed, quantitative data supports this intuitive supposition. In their review of the meta-analytic predictors of job performance, Schmidt and Hunter (1998) found the validity coefficient for Conscientiousness to be .31. In a meta-analysis of the FFM and job performance, Barrick, Mount, and Judge (2001) found that Conscientiousness was correlated with performance in virtually all jobs. Also,
Salgado (1997) concluded that Conscientiousness and emotional stability are the best FFM predictors of job performance.

In addition, Conscientiousness and general mental ability are not highly correlated (Cortina, et al 2000) making Conscientiousness a useful predictor to include in a selection battery. Further, it has been shown to account for additional, incremental variance in job performance over and above cognitive ability (Schmidt and Hunter, 1998). In addition to its significant predictive power, Conscientiousness is a popular personality trait to include in selection batteries as it has not been shown to have adverse impact against protected classes; Hough, Oswald and Ployhart (2001) found no significant differences between groups on the global Big-Five measure of Conscientiousness.

In addition to the personality trait Conscientiousness, Schmidt and Hunter (1998) found that integrity tests significantly predicted job performance ($r = .41$). They suggest that, when paired with cognitive ability, integrity accounts for a large increase in incremental validity, likely due to the fact that it does not correlate at all with cognitive ability. In their study, Schmidt and Hunter also reviewed the validity of employment interviews. They concluded that the employment interview is a significant predictor of job performance and that structured is more valid than an unstructured interview ($r = .58$, $r = .38$, respectively). In a related study, Huffcutt, Roth, and McDaniel (1996) reviewed 49 studies and found an average correlation of .40 between interview ratings and cognitive ability. They concluded that employment interview ratings do reflect cognitive ability. However, Schmidt and Hunter found the structured interview accounts for incremental variance in job performance over and above cognitive ability, and that the two combined have a validity of .63, suggesting there is substantial utility to each and significant benefits to combining the two predictors.
In the present study there are three general categories of subordinate predictors which have been shown to significantly, positively relate to job performance throughout the literature (Hunter and Schmidt, 1998; Barrick, Mount, and Judge, 2001): Cognitive ability, conscientiousness, and interpersonal skills. Within each category there are multiple specific competencies. The category “Cognitive Ability” is made up of the competencies Qualitative Problem Solving, Quantitative Problem Solving, Process Monitoring and Perceptual Speed. The “Conscientiousness” category includes Conscientiousness, Attention to Detail, and Responsibility. Finally, the category “Interpersonal Skills” includes Positive Attitude, and Teamwork.

**Job Performance**

Job performance is complex and multi-dimensional (Hogan and Roberts 1996), and, consequently, throughout the literature there has not been a consistent definition of this criterion. Some researchers believe that it is a more general factor like “g” (Arvey and Murphy, 1998). Others break down the criterion of job performance into the two distinctions of task performance and contextual performance (Viswesvaran and Ones, 2000). Research supports the notion that these criteria are in fact distinct. Individual difference characteristics have been shown to correlate differently with these two dimensions: Conscientiousness has been shown to be related to contextual performance and cognitive ability is related to task performance (Motowidlo & Van Scotter, 1994).

In addition, task performance and contextual performance have been shown to contribute independently to overall job performance. Motowidlo and Van Scotter (1994) found that contextual performance explained 12 to 34 percent of the variance in overall performance beyond the variance explained by task performance. Moreover, these measures appear to be
relatively independent of each other; Motowidlo et al found only a .20 correlation between the measures of contextual and task performance.

In addition to task and contextual performance, Viswesvaran and Ones (2000) suggest counterproductive behaviors as another general dimension that can stand alone and apply to all jobs. Robinson and Bennett (1995) suggest these organizationally deviant behaviors have negative, destructive effects on the organization and can include major and minor deviance aimed at either the organization itself or individuals. Viswesvaran and Ones also include withdraw behaviors such as absenteeism and turnover in the category of counterproductive behaviors. Research suggests managers consider these deviant behaviors when assigning performance ratings (Orr, Sackett, and Mercer, 1989).

Viswesvaran and Ones (2000) also review models of job performance that apply to specific occupations. They considered entry-level jobs in the service industry, managers, and military jobs. For example, Hunt (1996) contends there are nine dimensions of job performance for entry-level jobs in the service industry that are unrelated to job-specific knowledge. The proposed dimensions are: adherence to confrontational rules, industriousness, thoroughness, schedule flexibility, attendance, off-task behavior, unruliness, theft, and drug misuse.

Several different researchers have developed models of job performance for managers. PROFILER®, an instrument developed by Personnel Decisions Inc, assesses 24 dimensions of managerial performance. Brumback and Vincent (1970) identified 26 job performance factors for managers. Eighteen dimensions of managerial performance were presented by Borman and Brush (1993). Further, they segmented the 18 dimensions into four broad job performance dimensions for managers: leadership and supervision, interpersonal relations, technical behaviors and mechanics of management, and useful behaviors and skills.
Viswesvaran and Ones (2000) also reviewed the job performance models proposed for military jobs. For 19 entry level army jobs Campbell, McHenry, and Wise (1990) found five performance dimensions: core technical proficiency, general soldiering proficiency, effort and leadership, personal discipline, and physical fitness and military bearing. Borman, Motowidlo, Rose and Hansen (1985) studied factors important for unit success in first-tour soldiers and determined that, in addition to the standard task performance, there were three additional dimensions of job performance: allegiance, teamwork, and determination.

Campbell (1990) proposes eight dimensions that explain the underlying structure of job performance. That is, these factors apply to performance in all jobs. The eight dimensions are: job-specific task proficiency, non-job-specific task proficiency, written and oral communication, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision, and management or administration.

In the present study, the job performance measures collected included supervisor ratings of task, contextual, and overall job performance. I used overall job performance as the criterion for this study as it accounted for both task and contextual performance of incumbents. This was appropriate as Ones and Viswesvaran (2000) proposed the theoretical value of each measure and Motowidlo and Van Scotter (1994) demonstrated that contextual performance accounts for incremental variance in task performance.

**Supervisor ratings of job performance**

Obtaining reliable ratings of job performance has been a significant problem faced by researchers and practitioners alike. The “criterion problem” (Smith, 1976) – finding an appropriate, valid measure of job performance – has vexed researchers for years. The most commonly used measures are based on the judgments, or evaluations of supervisors. Because of
their inherently subjective nature, there is general agreement that a large amount of error is associated with these ratings, leading to lower reliabilities (Sturman, Cheramie, and Cashen, 2005).

Viswesvaran, Ones & Schmidt (1996) discussed the distinction between interrater and intrarater reliabilities. Interrater reliability assesses the extent to which different raters agree on a ratee’s performance. Intra-rater reliability assesses the extent to which items of a performance appraisal hang together. In supervisor ratings they found the interrater reliability to be .52 and intrarater reliability to be .85.

Historically, two major strategies have been used to increase the reliability of ratings: Improving the instruments or scales used, and rater training. Starting in the 1950’s most research focused on the development and improvement of the rating format (Arvey and Murphy, 1998; Howe, D., 1952). There was a concerted focus on the best method of eliciting a valid measure of performance. Many looked at the effect of the scale used (Howe, D., 1952), while others investigated the benefits of ranking subordinate performance. In the 1980’s, there was a shift away from researching rating instruments toward an understanding of cognitive processes in assigning ratings (Arvey & Murphy, 1998).

Much of the performance appraisal research involved investigating the effect of training the rater as a means of increasing rating accuracy. The general goal was to reduce rater error in the forms of leniency, central tendency, and halo. Most research found that rater training can improve rater accuracy (Woehr & Huffcut, 1994). Frame of Reference (FOR) training teaches raters to employ a common conception of performance levels when rating performance. Woehr & Huffcut found that FOR training is best at increasing rating and observational accuracy, and decreasing leniency.
The most recent literature proposes that discrepancies between ratings and actual performance are not a result of the cognitive deficiencies of the rater (Kane, 1992; Murphy & Cleveland, 1995). Rather, researchers suggest that inaccuracies in ratings reflect deliberate distortions made by the raters in order to accomplish their own, personal goals (Murphy, Cleveland, Skattebo, & Kinney, 2004). Ultimately, distortions in the criterion, intentional or not, have the potential to influence the reported relationships between predictors and job performance.

**Validity**

The primary goal of validation in selection is to determine the extent to which an assessment or process successfully predicts important outcomes, such as performance ratings, from supervisors and objective measures of performance. The Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, Inc., 1987 and 2003) describes the importance of criterion-related validity: “Personnel selection procedures are used to predict future performance or other behavior. Evidence for criterion-related validity typically consists of a demonstration of a useful relationship between the selection procedure (predictor or predictors) and one or more measures of job-relevant behavior (criterion or criteria)” (Principles, 1987, p. 6). Essentially, with a criterion-related approach, the goal is to empirically demonstrate that the predictor(s) being used are significantly and meaningfully related to the criteria being measured. However, selecting the most appropriate criteria to assess this relationship is challenging.

**Objective versus Subjective Criteria**

There is disagreement as to the most appropriate measure of the criterion job performance. Supervisor ratings of subordinate performance are by far the most common measure, but some
researchers purport that more objective measures are best. In research, objective and subjective measures have shown different relationships with predictors (Borman, Johnson, Rich, Podsakoff, & MacKenzie, 1995). In their meta-analysis, Borman and colleagues found the correlation between objective and subjective criteria to be only .39 and concluded the two measures are not interchangeable (Borman, Johnson, Rich, Podsakoff, & MacKenzie, 1995). Most researchers agree with the suggestion that the two are indeed distinct; however, there is much debate in the literature as to which measure is appropriate.

On the positive side for objective ratings, they are free from human bias. Additionally, because objective measures can usually be converted into financial terms it can be relatively easy for business decision makers to interpret and these ratings may provide more direct links to their financial benefit to the organization. However, there are some drawbacks to using objective measures as job performance criteria. Unfortunately, they are sometimes difficult to obtain. In addition, they are often difficult to accurately, consistently interpret. For example, considering attendance may not give a great indication of an employee’s work ethic. They are often contaminated by external factors such as geographical location, season, or an incumbent’s specific job title. There is also usually low variance and/or poor reliability associated with objective ratings (Sturman, Cheramie, and Cashen, 2005). Furthermore, because objective measures usually have less variance (understating the validity), it makes predicting a relationship more difficult.

For the aforementioned reasons, most researchers use subjective, supervisor ratings as job performance criteria (Landy and Farr, 1980). A notable pro to using subjective criteria is that there is more variance in these ratings. In addition, these ratings are better suited to take into account extraneous variables that may impact a ratee's performance that is beyond the ratee's
control. For example, poor sales for a particular incumbent may be more indicative of a struggling economy than the employee’s actual performance. Also, objective factors of contextual performance or organizational citizenship behaviors are difficult to attain. Accordingly, subjective, supervisor ratings of those facets are more plausible. Two notable drawbacks to using subjective measures of job performance are contamination and deficiency in the criterion. Ratings could be contaminated with a myriad of factors. Characteristics of the rater, for example their goals and tenure, have been hypothesized to effect performance ratings (Murphy, et al., 2004; Murphy and Cleveland, 1995). Supervisor ratings may also exhibit criterion deficiency. Deficiency in subjective measures may be caused by factors such as opportunity to observe. For example, a supervisor may work remote from the job site and not have direct, personal interaction with the employee and thus have a limited perspective of her performance.

**Systematic Variance in Validity Coefficients**

As previously mentioned, one of the major issues with using subjective, supervisor ratings as measures of performance is that supervisor characteristics may account for some of the variance in the ratings of their subordinates’ performance. Early research suggested that this variance was unintentional and due to the cognitive limitations of the raters (Bernardin & Villanova, 1986). More recently, it has been proposed that the distortion of ratings may be the result of a deliberate decision by the raters to manipulate these ratings. This is an important distinction because, according to classical test theory, errors in measurement must be random and non-systematic. Error must also have certain psychometric properties such as similar means and standard deviations. The major point of contention is recent research suggests that differences in ratings are not random at all, but are likely systematic (Murphy and Cleveland,
There are a few rater characteristics that have been suggested to have a systematic effect on performance ratings and ultimately on validity coefficients.

**Common Rater Errors**

Throughout the literature rater errors have been studied in an attempt to highlight these common mistakes and train the raters to avoid committing them. A few of the rater errors that have been researched are leniency, similarity bias, and halo.

**Leniency.** As discussed above, the classical model assumes that rater error has certain psychometric properties. Research has shown that rater error does not meet those properties. For example, the error is not uncorrelated and unsystematic. One of the most pervasive rater idiosyncrasies (error) is leniency (Murphy & Cleveland, 1995), which suggests that raters are consistently, or systematically more lenient to employees across the board. Leniency then, by definition, is not random – scores are systematically higher (Murphy & DeShon 2000). Because this type of bias occurs independently of the ratee it leads to a similar effect on all ratees.

Research has found a systematic effect of supervisors’ levels of leniency on their ratings and different reasons for this finding have been suggested. For example, Bartels and Doverspike, (1997) found that leniency was related to personality characteristics of the rater. Specifically, they found that leniency was related to sensitivity, warmheartedness, and tough poise, such that individuals high in these traits were more likely to give higher mean ratings. In an article discussing the role of rater motivation in the appraisal process, Harris (1994) suggests that leniency is a function of the level of accountability the raters have to their subordinates. Specifically, he purports that when raters were held accountable for their ratings by their subordinates they were more likely to provide consistently higher ratings. Bernardin and Orban (1990) found that lower trust in the appraisal process led to leniency in ratings. They also found
that supervisors who perceived that they were given fair ratings from their supervisors gave their subordinates more accurate ratings.

**Similarity bias.** Different than leniency, similarity bias is a result of the interaction between a rater and a ratee (Schmitt, Pulakos, Nason & Whitney, 1996). Research on rater effects suggests that raters may have a similar-to-me bias (Murphy & Cleveland, 1995). They purport that raters give consistently higher ratings to those who they perceive to be similar to them. For example, in a meta-analysis, Kraiger and Ford (1985) found that ratees received higher ratings from same-raced raters. It is worth noting that one limitation of the study is that the methodology was limited, in that black and white raters did not rate the same people. Also, Schmitt, Pulakos, Nason and Whitney (1996) found a significant relationship between the supervisors’ perceived similarity to the ratee and the performance rating they gave to the ratee.

**Halo.** The halo effect was proposed initially by Thorndike (1920). In his article, he reviews a study where a flight commander was instructed to rate his cadets in four separate areas: Physical qualities, intelligence, leadership, and personal qualities. The instructions in the official rating plan required “very emphatically” that each area be rated independently. However, the intercorrelations were abnormally high. Specifically, the correlations of intelligence with physique, leadership, and character were .51, .58, and .64, respectively. Thorndike notes that the correlation of intelligence with physique should be about one-third the correlation of intelligence with character and/or leadership. Halo in performance ratings occurs when raters do not adequately differentiate between separate aspects of ratees’ performance. Lance, LaPointe and Stewart (1994) purport that halo can be defined in three ways (1) a general evaluation that affects all dimensional ratings, (2) a salient dimension that affects ratings on other dimensions, and (3) insufficient discrimination among dimensions (Solomonson and Lance, 1997). This ultimately
results in higher correlations between different dimensions of performance than true relationships. Feldman (1981) and Ilgen and Feldman (1981) suggest distinct behaviors are usually classified into categories and, because they are so similar, raters merge the otherwise distinct behaviors together and similar ratings, with abnormally high intercorrelations, result.

For example, a supervisor may give all high or all low ratings on different performance dimensions, in spite of the fact that the ratee’s performance does vary. This effect is important in the context of performance ratings because failing to discriminate between different areas of performance reduces accuracy, and ultimately validity (Borman, 1975). Throughout the literature, researchers have studied different methods of reducing halo (e.g., simply instructing raters to consider each item independently, statistically controlling for halo) (Follman, Wiley, Geiger, & Lavely, 1974; Harvey, 1982). For example, Palmer and Feldman (2005) found that halo is significantly weaker when raters are held accountable for their ratings. When raters were told they would have to justify the ratings given for each dimension they were better able to discriminate. It should be noted, however, they found that these results were more pronounced when the performance was good rather than poor. This supports previous research that suggests that poorer performance has a stronger effect than good performance on processing and ratings (Maurer, Palmer & Ashe, 1993) and, thus, halo, in these cases is more difficult to overcome.

**Goal-Directed Rater Effects**

In I/O literature, the notion that rater effects have an influence on reliability has been suggested (Murphy et al, 2004; Murphy & Cleveland, 1995). One hypothesis is that rater level variables, such as number of subordinates rated, could account for variance in individual ratings such as leniency (LaHuis & Avis, 2007). It has also been suggested that rater goals could account for variance in job performance ratings (Murphy et al., 2004). Murphy & Cleveland (1995)
suggest four broad categories of goals raters may pursue when appraising subordinate performance: (a) task-performance, (b) interpersonal goals, (c) strategic goals, and (d) internalized goals.

In task-performance goals, raters use performance appraisal to maintain or increase their subordinates’ level of performance. Murphy and Cleveland (1995) note that raters who employ task-performance goals may use performance appraisal to direct the development of employees in an effort to prepare them for future assignments or positions. Interpersonal goals involve using performance appraisal to keep relations smooth between rater and ratees. Assuming a rater cares how she is perceived by her subordinates, she may employ interpersonal goals and increase ratings in order to maintain a positive work environment. A rater employs strategic goals when evaluating performance when she uses performance appraisal to increase her team’s or her own standing in the organization. A rater may give consistently high ratings to those on her team to ensure that her workgroup is perceived as a high performing team in the organization. Murphy and Cleveland describe internalized goals as a reflection of the raters’ beliefs and ideals. For example, a rater with a high level of integrity may be more interested in rating accurately and honestly than giving uniformly high ratings.

Although little research has been done in this area it is logical to surmise there are other rater characteristics that may account for some of the variance in job performance ratings. For example, a supervisor’s cognitive ability, or their interpersonal skills could help explain some rating variance. There are four categories of supervisor characteristics that I modeled to determine if they account for significant variance in predictor-criterion relationships for their subordinates. The four broad categories are Cognitive Ability, Interpersonal Skills, Promotability, and Cognitive Demands. In two of the four supervisor categories (as with the
subordinates) there are multiple, specific competencies which comprise the broad categories. The category “Cognitive Ability” includes measures of Interpreting Information and Business Planning. “Interpersonal Skills” includes only a single measure of supervisors’ level of Adaptability. The category of “Promotability” reflects whether or not the supervisor has been promoted out of their supervisor or First Line Manager position since the ratings were originally assigned. Finally, the category of “Cognitive Demands” contains the length of time they supervised the subordinate and the number of subordinates they rated.

**Cognitive ability.** Traditionally, it was believed that, in performance appraisal, it was the goal of supervisors to give accurate, valid assessments of subordinates’ job performance. Additionally, early research suggests that discrepancies between true performance and the performance rating assigned by the supervisor were a function of the supervisors’ level of cognitive ability (Bartles and Villanova, 1988). Specifically, the proposition was that more intelligent supervisors, relative to their less intelligent counterparts, would more accurately rate the performance of their subordinates. In view of that, it would follow that more intelligent supervisors would be more likely to attend to the relative levels of significant predictors of their subordinates and to consider those when assigning performance ratings. In the present study, I have measures of supervisor levels on two competencies that measure cognitive ability: Interpreting Information and Business Planning. Specifically, Interpreting Information is a measure of the supervisor’s ability to systematically investigate a problem to better understand a situation. As part of this measure, supervisors identify and interpret underlying trends, interrelationships and cause-effect relationships among seemingly unrelated pieces of data. They also draw logical inferences and make rational recommendations based on the application of inductive and deductive reasoning skills. Business Planning measures supervisors’ skills in
planning projects such that specific tasks, timelines, milestones, and objectives are clearly established in advance. It also measures their ability to institute effective methods for keeping track of project details and organize or adjust information and materials to ensure accurate and timely completion of tasks. Accordingly, I hypothesize the overall mean rating will decrease as the supervisors’ intelligence increases and the predictor-criterion relationships (validity) for subordinates will increase as the level of the supervisors’ intelligence increases. Specifically:

*Hypothesis 1a: Supervisors’ levels of Interpreting Info will be negatively related to subordinate performance ratings.*

*Hypothesis 1b: Supervisors’ level of Interpreting Info will be positively related to validity coefficients.*

*Hypothesis 2a: Supervisors’ levels of Business Planning will be negative related to subordinate performance ratings.*

*Hypothesis 2b: Supervisors’ level of Business Planning will be positively related to validity coefficients.*

**Cognitive demands.** Included in the category “Cognitive Demands” are the measures of the length of time supervisors have managed the subordinates and the number of subordinates rated. These variables are seen as proxies for cognitive demands in this study because, generally, the longer a supervisor has been with the organization, the more responsibilities (e.g., committee memberships, work tasks) he or she is given. These multiple tasks and responsibilities demand more cognitive resources. In addition, the number of subordinates a supervisor has is seen as a proxy for level of cognitive demands as managing more subordinates leads to an increase in manager responsibilities (e.g., checking timesheets, approving vacations, resolving employee conflicts). It is reasonable to think that a supervisor may be likely to spend less time on
accurately assessing the job performance of subordinates, if she has a high level of cognitive demands placed on her. For example, if a supervisor is on several committees, has a large number of subordinates, and is responsible for a large, complex production line, it reduces the amount of time she is able to observe her subordinates’ performance directly. Findley, Giles, & Mossholder (2000) found that the more observations managers did, the more likely they were to give high contextual performance ratings.

In addition, based on an extensive review of the relevant research, Krosnick (1991) proposes that increased cognitive demands decrease both the rater’s ability and motivation to accurately assess and report. He suggests that when raters face significant cognitive demands “they perform incomplete or biased searches of memory and integration of retrieved information… they eliminate the steps of retrieval and judgment all together and generate responses by interpreting questions and selecting response alternatives that seem reasonable without ever making reference to the relevant knowledge stored in memory.” (p. 229). Based on Findley et al.’s., (2000) and Krosnick’s (1991) findings from the relevant research I propose that as the cognitive demands on the supervisors increase, their ability and motivation to accurately reference necessary predictors and indicators when rating performance will decrease; the average rating will increase and there will be a subsequent decrease in the validity of the predictors as the cognitive demands of the supervisor increase.

Specifically:

_Hypothesis 3a: Supervisors’ number of subordinates will be positively related to subordinate performance ratings._

_Hypothesis 3b: Supervisors’ number of subordinates will be negatively related to validity coefficients._
Hypothesis 4a: Supervisors length of time supervising will be positively related to subordinate performance ratings.

Hypothesis 4b: Supervisors length of time supervising will be negatively related to validity coefficients.

**Promotability.** The category “Promotability” denotes a variable that distinguishes those supervisors who have been promoted out of their First Line Manager position since they rated their subordinates (within the last 16 months) from those who have remained. With respect to the four broad categories of rater goals Murphy and Cleveland (1995) proposed, those supervisors who have been promoted are likely to have used performance appraisal to achieve *strategic* goals. Supervisors who strive to achieve this goal use appraisal to enhance their position, or their team or workgroup’s position, within the organization. For example, in order to gain valued rewards or promotions for their team, or individual contributors of the team, the supervisor may give consistently higher ratings to members of his team. It is possible, then, that those supervisor’s ratings of their team were higher than their counterparts with more accurate (lower) ratings of their respective workgroups, and that lead, in part, to the promotion of those supervisors.

Considering the variable “Promotability” in light of Murphy and Cleveland’s (1995) proposed strategic goals category, I suggest that as the supervisors’ propensity to be promoted increases, subordinate ratings will also increase and the validity of the predictors will decrease. Specifically:

*Hypothesis 5a: Supervisors’ Promotability will be positively related to subordinate performance ratings.*
Hypothesis 5b: Supervisors’ Promotability will be negatively related to validity coefficients.

Interpersonal skills. As previously mentioned, there are several interpersonal skills found to have an effect on ratings. For example, Bartels & Doverspike, (1997) found that the propensity to be lenient when assigning ratings is related to rater characteristics. They found that raters were more likely to be lenient when they had increased levels of sensitivity and warmheartedness. The measure in this study of supervisors’ Adaptability assesses the extent to which they prefer to be amenable and not “rock the boat,” and also the extent to which they feel the strong need to keep relations smooth within the team. Researchers have suggested that supervisors with such goals are more likely to be lenient and rate less accurately (Murphy and Cleveland, 1995; Murphy, Cleveland, Skattebo, and Kinney, 2004). Accordingly, I hypothesize that the overall mean will increase and the validity for subordinates’ predictors with job performance will decrease as the supervisors’ level of Adaptability increases.

Specifically:

Hypothesis 6a: Supervisors’ level of Adaptability will be positively related to subordinate performance ratings.

Hypothesis 6b: Supervisors’ number of Adaptability will be negatively related to validity coefficients.

Method

An archival data set was used to test the hypotheses for this study.

Participants

Participants were entry level subordinates (N = 494) and their supervisors (N = 82). Each supervisor managed an average of six subordinates. They are employed by a large international
manufacturing organization, though all participants were from plants located in North America. For the subordinates 52% were white, 39% were African American, and 5% were Hispanic; 76% were males. No race or gender information was available for the supervisors. 91% of supervisors had been supervising their employees for 0 – 12 months; 9% had been supervising them for more than one year.

**Measures**

**SAM.** Subordinates completed the Select Assessment for Manufacturing (SAM). They completed this assessment either, as part of a larger selection process, or as part of a concurrent validation study completed in 2006 or a predictive validation study completed in 2007. This test incorporates multiple assessment methodologies. Specifically, the assessment includes a 1) personal beliefs section, where team members respond to what extent they agree or disagree with a statement, 2) comparison, where candidates look for a match to a given number in a grid located in another part of the screen, measuring speed and accuracy, 3) situational judgment where candidates read a scenario and rate three possible responses, 4) process monitoring where candidates complete two tasks, simultaneously. The first task is to monitor gauges and the second is to compare two sets of numbers and decide if they are the same or different, 5) problem solving – quantitative, where candidates answer math questions related to a production line process, 6) problem solving – qualitative; in this section candidates answer non-math, multiple choice questions about the production line.

SAM measures 11 competencies. The 11 competencies measured and their scale reliabilities are: Attention to Detail (.70), Positive Attitude (.70), Process Monitoring (.70), Qualitative Problem Solving (.61), Quantitative Problem Solving (.64), Responsibility (.64), Risk Reliability (.66), Safety Orientation (.63), Teamwork (.64), Work Ethic (.69), and Perceptual
Speed (.75). The test-retest reliability for the composite overall score is .82. The means, standard deviations, and intercorrelations for subordinate variables can be found in Table 1.

Supervisors completed either the Select Assessment for Leadership Performance (SALP) or the Select Assessment for Leader Development (SALD). All supervisors completed only one assessment as either part of a larger process for selection, or as part of a developmental project. Both assessments contain 1) situational judgment items, where the candidate is presented with realistic scenarios dealing with key issues faced by leaders and is asked to rate effectiveness of several possible options for dealing with the situation; 2) logical reasoning items which presents candidates with a series of three objects that begin a pattern and a fourth object that completes each pattern series. The candidate’s task is to select the object that completes the series from the options presented; 3) personal belief questions which are self-report items covering multiple competency areas where candidates are provided with a series of statements and asked the degree to which they agree or disagree; and 4) applied problem solving scenarios where the candidate is presented with a description of a problem followed by a set of logical conditions and are asked to use the information presented to answer a series of questions.

**SALP.** This assessment consists of multiple scales measuring 10 constructs: (1) Adapting to Change, (2) Business Planning, (3) Competitive Drive, (4) Conscientiousness, (5) Decision Making Style, (6) Empowering Others, (7) Initiative, (8) Interpreting Information, (9) Leadership Style, and (10) Working with Others. The internal consistency of the scales in SALP are all acceptable (> .70).


Criterion Measure: Subordinate Performance Evaluation Form. A performance rating form was administered to the subordinates’ immediate supervisors. The criteria rating forms consisted of 33 items assessing 9 different performance areas: (1) Adaptability, (2) Organizational Citizenship Behavior (OCB) – loyalty, (3) OCB - persistence with enthusiasm, (4) OCB – volunteering, (5) Problem Solving, (6) Quality Awareness, (7) Responsibility, (8) Teamwork, and (9) Perceptual Speed. Underlying this nine-factor scale were two primary factors, task behavior and contextual behavior. For example, an item that reads “maintain a good work pace” measures task performance/work tempo whereas an item that reads "volunteers for extra work” measures contextual performance/OCB-volunteering. Item responses are indicated on a 7-point likert-type scale. An overall measure of job performance was used for this study. This rating consists of 50% task performance, 30% contextual performance and 20% from the overall performance rating. Criterion scores ranged from 2.55 – 7.00. This and an earlier version of the rating form and format have been used in previous research (Hattrup, O’Connell, & Wingate, 1998; Hattrup, O’Connell, & Labrador, 2005).

Procedure

The data for this study was collected during multiple stages. The criterion data for the production-level incumbents (i.e., their performance ratings) were collected during either a 2006
concurrent validation study or in a 2007 predictive validation study. Production-level predictor data (i.e., their SAM assessment scores) were either collected during a 2006 concurrent validation or at various times between 2006 and 2008 as part of the organization’s routine selection process. A portion of the supervisor level predictors (i.e., SALD and SALP) were collected at various times for selection purposes. Another portion of these supervisors were assessed as part of performance evaluation I conducted in 2008.

Analyses

Typically, job performance ratings are assigned by a supervisor to the team of subordinates they manage. Because these ratings are nested within supervisors, there is potentially some variance within those ratings which can be attributed to supervisor level characteristics. It is appropriate, then to use multi-level analysis in the performance rating context to determine the effect these supervisor-level characteristics have on 1) the ratings they assign and, ultimately, 2) the validity or relationship between subordinate predictor and criterion (i.e., rating); rather than assuming that any variance attributable to the supervisor is error and unsystematic (as the classical model would suppose), multilevel analysis allows us to acknowledge that homoscedasticity predict and model that variance.

Essentially, each of my six hypotheses can be grouped into two categories: Effect of supervisor-level predictor on subordinate average performance rating, and the effect of supervisor-level predictor on validity of subordinate predictors. In part “a” of each hypothesis, I predicted the effect of the supervisor level variable on the average subordinate performance rating. For example, Hypothesis 1a states “As supervisors’ level of Interpreting Information increases the average subordinate performance rating decreases.” In sections b of each hypothesis, I predicted the effect of the supervisor level variable on the relationship of the
subordinate predictor-performance rating relationship, or the validity of each subordinate predictor. For example, Hypothesis 1b states “As the supervisors’ level of Interpreting Information increases, the subordinate validity coefficient for Perceptual Speed will increase.”

In order to test my hypotheses, I used multilevel modeling. This technique is appropriate as I have subordinates nested within supervisors. To illustrate how I tested my hypotheses, I’ll use Hypotheses 1a and 1b as examples.

To begin, I have data from $j$ supervisors (in this case, $j = 82$), who manage a different number of subordinates, $n_j$. For the subordinates, I have the predictor variable Performance Rating measured using a 7-point Likert scale. My explanatory variable in this case is subordinates’ Perceptual Speed score. My supervisor-level predictor variable is Interpreting Information.

Below is the Level 1 regression equation for each supervisor used to predict the outcome variable $Y$ by the explanatory variable $X$:

$$\text{Performance Rating}_{ij} = B_{0j} + B_{1j} \text{Perceptual Speed}_{ij} + e_{ij} \quad (1.1)$$

In the equation above, $B_{0j}$ represents the usual intercept and $B_{1j}$ represents the usual regression slope. Because this is a multilevel problem, I am assuming that there are different intercepts and slopes for the subordinate level variables. I want to explain the variance in the regression coefficients $B_{0j}$ and $B_{1j}$ by adding supervisor-level predictor variables, below:

$$B_{0j} = \gamma_{00} + \gamma_{01} \text{Interpreting Information} + u_{0j} \quad (1.2)$$

$$B_{1j} = \gamma_{10} + \gamma_{11} \text{Interpreting Information} + u_{1j} \quad (1.3)$$
Equation 1.2 predicts supervisors’ average subordinate performance rating by supervisors’ Interpreting Information score. Hypothesis 1a will be supported if $\gamma_{01}$ is negative. That is, if $\gamma_{01}$ is negative, supervisors’ average subordinate performance rating is lower for those supervisors with higher levels of Interpreting Information.

Equation 1.3 states that the relationship (validity) between the performance rating and subordinates Perceptual Speed is a function of the supervisors’ level of Interpreting Information. Support for my hypotheses predicting effects on validity would be found by examining $\gamma_{11}$. The direction (i.e., whether it is positive or negative) will determine whether increases in the supervisor level variable are associated with increased or decreased validity in the subordinate predictor-criterion relationship. That is, if $\gamma_{11}$ is positive, (in the case of Hypothesis 1b) as supervisors’ level of Interpreting Information increases, the relationship between subordinate Perceptual Speed and performance ratings increase, supporting this hypothesis.

All other hypotheses will be analyzed using these methods.

Results

To test the “a” part of my hypotheses, I predicted effect of supervisor level characteristics on average overall performance rating for subordinates. For example, I expected those supervisors with higher levels of cognitive ability (i.e., higher Interpreting Information and Business Planning scores) to assign lower performance ratings. Table 4 presents the supervisor competencies predicting the intercepts, or the effect on the average performance ratings. In part “b,” I hypothesized the validity, or the relationship between subordinates’ performance ratings and predictor variables would be moderated by supervisor level variables. For example, I expected that validity would increase for those supervisors who have higher levels of cognitive
ability. Table 5 presents the gammas and standard errors associated with the slopes, or the effect of supervisor competencies on the relationship between subordinate predictors and average performance ratings.

Variance Components

Before I began testing my hypotheses, I first wanted to determine if there were in fact systematic differences in validity coefficients. In order to determine this, I tested the variance components for the slopes for all level one predictors.

The first step was to estimate the null model with all level one predictors included, and all slopes fixed. Next, I allowed one level one predictor slope to vary. I compared the $\chi^2$ from the random intercept model to the $\chi^2$ in a model where I allowed the single level 1 predictor’s slope to vary. I ran these analyses, one at a time, for each of the eleven level one predictors and compared each to the null model. I found significant differences between the null model and the level one predictor Safety ($\chi^2 = 6.36, p = .04$). The differences for the remaining level 1 predictors were not significant.

Table 3 presents the $\chi^2$ difference and the degrees of freedom between the slope in the null model and each of the slopes associated with the level one predictors. Finding significance for the level one predictor Safety indicated that the validity coefficients for these predictors did in fact vary across supervisors. The next step was to determine if any of the supervisor characteristics examined in this study accounted for that variance.
Cognitive Ability

Interpreting Information

Hypothesis 1a was supported; supervisors’ level of Interpreting Information had a significant relationship with the average overall performance rating they assigned to subordinates ($\gamma = -.12, p = .01$). Specifically, as supervisors’ levels of Interpreting Information increased, the average overall performance rating decreased.

Hypothesis 1b was not supported; supervisors’ level of Interpreting Information was not related to the regression coefficients for Attention to Detail ($\gamma = .01, p = .69$), Perceptual Speed ($\gamma = -.03, p = .12$), Positive Attitude ($\gamma = -.01, p = .33$), Process Monitoring ($\gamma = -.01, p = .55$), Qualitative Problem Solving ($\gamma = -.01, p = .68$), Quantitative Problem Solving ($\gamma = .00, p = .93$), Responsibility ($\gamma = -.00, p = .74$), and Teamwork ($\gamma = .01, p = .43$). Supervisors’ level of Interpreting Information was significantly related to the slopes for Conscientiousness and Safety, but in the opposite direction ($\gamma = -.03, p = .01$, and $\gamma = -.04, p = .05$, respectively).

Business Planning

Hypothesis 2a was not supported; supervisors’ level of Business Planning had no significant effect on the average overall performance rating they assigned to subordinates ($\gamma = .04, p = .79$).

Hypothesis 2b was not supported; supervisors’ level of Business Planning was not related to the regression coefficients for Attention to Detail ($\gamma = .04, p = .52$), Conscientiousness ($\gamma = .05, p = .19$), Perceptual Speed ($\gamma = .05, p = .41$), Positive Attitude ($\gamma = .01, p = .83$), Process Monitoring ($\gamma = .04, p = .29$), Qualitative Problem Solving ($\gamma = -.03, p = .64$), Quantitative
Problem Solving ($\gamma = .01, p = .86$), Responsibility ($\gamma = .07, p = .10$), Safety ($\gamma = -.01, p = .92$), and Teamwork ($\gamma = .05, p = .35$).

**Cognitive Demands**

The number of subordinates supervised and the length of time supervising those subordinates were the two variables used as proxies for supervisors’ levels of cognitive demands.

**Number of Subordinates**

_Hypothesis 3a_ was not supported; the number of subordinates a supervisor managed had no significant effect on the average overall performance rating they assigned to subordinates ($\gamma = -.00, p = .97$).

_Hypothesis 3b_ was partially supported; as supervisors’ number of subordinates increased, the relationship between subordinates’ average performance rating and the predictors Positive Attitude, Process Monitoring, and Responsibility decreased significantly ($\gamma = -.01, p = .02$; $\gamma = -.00, p = .04$; and $\gamma = -.01, p = .01$, respectively). However, supervisors’ number of subordinates was not related to the regression coefficients for Attention to Detail ($\gamma = -.01, p = .06$), Conscientiousness ($\gamma = -.00, p = .87$), Perceptual Speed ($\gamma = -.00, p = .14$), Qualitative Problem Solving ($\gamma = -.00, p = .85$), Quantitative Problem Solving ($\gamma = .00, p = .50$), Safety ($\gamma = -.00, p = .31$), and Teamwork ($\gamma = -.00, p = .33$).

**Length of Supervision**

_Hypothesis 4a_ was not supported; the length of time a supervisor managed his or her subordinates had no significant relationship with the average overall performance rating they assigned to subordinates ($\gamma = .06, p = .37$).

_Hypothesis 4b_ was not supported; supervisors’ length of time supervising subordinates was not related to the regression coefficients for Conscientiousness ($\gamma = .04, p = .09$), Perceptual
Speed ($\gamma = .03, p = .29$), Positive Attitude ($\gamma = .03, p = .23$), Process Monitoring ($\gamma = .03, p = .12$), Qualitative Problem Solving ($\gamma = -.02, p = .58$), Quantitative Problem Solving ($\gamma = -.01, p = .78$), Responsibility ($\gamma = .03, p = .13$), Safety ($\gamma = .05, p = .22$), and Teamwork ($\gamma = .01, p = .85$). In addition, as supervisors’ length of time supervising subordinates increased, the validity of subordinates’ average overall performance rating and the predictor Attention to Detail ($\gamma = .08, p = .03$) increased significantly.

**Promotability**

*Hypothesis 5a* was not supported; supervisors’ Promotability did not have a significant relationship with the average overall performance rating they assigned to subordinates ($\gamma = -.32, p = .12$).

*Hypothesis 5b* was not supported; supervisors’ Promotability was not related to the regression coefficients for Attention to Detail ($\gamma = -.082, p = .186$), Conscientiousness ($\gamma = -.03, p = .43$), Perceptual Speed ($\gamma = -.06, p = .26$), Positive Attitude ($\gamma = -.05, p = .16$), Process Monitoring ($\gamma = -.04, p = .26$), Qualitative Problem Solving ($\gamma = .02, p = .72$), Quantitative Problem Solving ($\gamma = .04, p = .19$), Responsibility ($\gamma = -.07, p = .07$), Safety ($\gamma = -.07, p = .29$), and Teamwork ($\gamma = -.06, p = .21$).

**Interpersonal Skills**

The variable Adaptability was used in this study as a measure of supervisors’ interpersonal skills.

**Adaptability**

*Hypothesis 6a* was supported; supervisors’ levels of Adaptability had a significant relationship with the average overall performance rating they assigned to subordinates ($\gamma = .08, p = .02$). Specifically, as supervisors’ levels of Adaptability increased, the average overall performance rating increased.
Hypothesis 6b was not supported; supervisors’ adaptability was not related to the regression coefficients for Attention to Detail ($\gamma = .00, p = .92$), Conscientiousness ($\gamma = .01, p = .58$), Perceptual Speed ($\gamma = -.01, p = .70$), Positive Attitude ($\gamma = -.01, p = .57$), Process Monitoring ($\gamma = .00, p = .64$), Qualitative Problem Solving ($\gamma = -.01, p = .44$), Quantitative Problem Solving ($\gamma = -.00, p = .58$), Responsibility ($\gamma = -.01, p = .45$), Safety ($\gamma = .01, p = .50$), and Teamwork ($\gamma = -.02, p = .06$).

Discussion

Of the six rater-level characteristics investigated, there were 2 that had a statistically significant relationship with the performance ratings that raters assign. As supervisors’ level in the competency Interpreting Information (a proxy for cognitive ability) increased, the average performance ratings they assigned to subordinates decreased. Thus, Hypothesis 1a was supported. Early research posited that any discrepancy between ratings and actual performance was due to the cognitive limitations of the rater and that more intelligent supervisors would rate more accurately (Bartles and Villanova, 1988). My results suggest that more intelligent supervisors assign lower ratings than their less intelligent peers. Research in rater errors posits that the most pervasive rater error is inflating ratings (Murphy and Cleveland, 1995). It is logical, then, to surmise that lower ratings are more accurate than inflated ratings.

If this is so, assuming the subordinate (level 1) predictors are valid, the validity coefficients should increase as supervisors’ level of Interpreting Information (cognitive ability) and the accuracy of the ratings increase. However, my results did not support this: Although an increase in raters’ level of Interpreting Information was associated with a significant decrease in the performance ratings they assigned, this rater characteristic was not associated with a corresponding increase in the validity of subordinate (level 1) predictors. Supervisors’ levels of
Interpreting Information had no effect on the validity of eight of the subordinate predictors. However, as supervisors’ Interpreting Information levels increased, the validity of two subordinate predictors (Safety and Conscientiousness) decreased.

Although, this is counter to what I hypothesized, there are a couple explanations for this unexpected finding. Those high in the construct Conscientiousness, as measured in this study, are very thorough, and careful. They are unlikely to be comfortable moving quickly and making rushed decisions with little direction. We know from several meta-analyses (e.g., Barrick and Mount, 1991) that Conscientiousness is an important predictor of job performance; generally, being deliberate and thorough is seen as important for success on a job. However, additional research suggests that just how important may depend on the position. Robie and Ryan (1999) suggest that a nonlinear Conscientiousness-job performance relationship may likely be seen in a fast-paced work environment.

In the context of this study, the work environment is very structured and fast-paced. A minimum level of Conscientiousness is required, even expected, to perform the job. However, as individuals’ levels of Conscientiousness increases, they are less likely to move quickly and be comfortable in a fast-paced setting. They may be so careful and thorough that they perform at a slower pace than their cohorts who are lower in Conscientiousness. That deliberate, cautious pace may decrease their output. Supervisors who are more intelligent may consider these more tangible performance outcomes (e.g., product produced per hour) when assigning performance ratings, as they are highly valued from an organizational perspective.

The explanation for the unexpected relationship with the predictor Safety is related. In fact, the two constructs share quite a bit of variance. Certainly in a manufacturing environment you would be unlikely to find a supervisor who doesn’t purport to value Safety. In practice,
though, perhaps only a minimum level of Safety is required and much beyond the requisite may hinder performance. Similar to those with high levels of Conscientiousness, individuals with high levels in Safety have a strong need to follow the rules. Those who are safety-oriented tend to be risk-averse; they are not comfortable trying new activities and do not seek out novel situations. Having subordinates with interest in learning new tasks or being able to shift to cover other positions would likely be important to supervisors. Again, seeing the value these individuals add to an organization in the long term, more intelligent supervisors may be likely to value employees who are willing and able to learn a different position, or train on a brand new machine. They are likely to see these employees as indispensible and, accordingly, assign them relatively higher job performance ratings than their more hyper-safe counterparts.

The second rater characteristic that was related to assigned performance ratings was the competency Adaptability. My findings suggest that as supervisors’ level of Adaptability increases, the average performance ratings they assign to subordinates also increases. These results provide support for Hypothesis 6a. These mirror the findings of Bartles and Doverspike (1997), who found that raters with increased levels of Sensitivity and Warmheartedness are more likely to be lenient (and thus, inflate) when assigning performance ratings. The measure of Adaptability assesses supervisors’ propensity to be amenable and their desire to keep relations between themselves and the team smooth.

Accordingly, my findings also lend support to research conducted by Murphy, Cleveland, Skattebo, & Kinney (2004) which suggests that those raters with such goals are more likely to be lenient. Taken a step further, they suggest that these ratings are likely to be inflated and less accurate or reflective of actual performance. If this were the case, again assuming the subordinate (level 1) predictors are valid, the subordinate validity coefficients should decrease as
supervisor levels of Adaptability increase. My findings, however, do not support this. Although an increase in supervisors’ level of Adaptability was associated with an increase in the average performance ratings assigned, there was no corresponding decrease in the validity of any level 1 (subordinate) predictors.

In Hypothesis 3 (a and b), I theorized that supervisors with higher numbers of subordinates were more likely to have increased cognitive demands, would rate less accurately (i.e., inflate their ratings) and thus, the validity of subordinate predictors would decrease. I reasoned that these demands would increase because managing more subordinates leads to an increase in other managerial responsibilities (e.g., checking timesheets, approving vacations, resolving employee conflicts). Supervisors under such increased demands may have less time, and less ability and motivation (Krosnick, 1991) to accurately assess the performance of subordinates. Providing additional support, Findley, Giles, & Mossholder (2000) found that the more observations managers did, the more likely they were to give high contextual performance ratings.

Although it was not significantly related to overall performance ratings, supervisors’ number of subordinates (a proxy for their level of cognitive demands) did moderate the subordinate predictor-performance rating relationship for three predictors. As supervisors’ number of subordinates increased, the relationship between subordinates’ Positive Attitude, Process Monitoring, and Responsibility score and their performance rating decreased significantly. These findings provided partial support for Hypothesis 3b. The validity coefficients for the other seven subordinate predictors were unaffected by the number of subordinates raters supervised.
Though previously mentioned, it bears mentioning again that the predictor data at levels 1 and 2 were collected at several different stages and for different purposes. In the case of the subordinate predictor data, some individuals completed this assessment as incumbents, participating in the concurrent validation study. Others completed the assessment while candidates as part of the selection process. Some supervisors completed the assessment as part of an annual performance evaluation to assess gaps and strengths and others completed the same assessment as candidates in part of their selection process. Obviously, differences likely exist between the motivation of a candidate applying for a job, a current incumbent being evaluated for administrative purposes, and one simply participating in a study. This inconsistency in predictor data may have potentially made it more challenging to find significant effects of rater characteristics. However, though not every hypothesis was supported, the fact that significance was found despite this lack of control provides evidence of external validity.

Limitations

A potential limitation in this study was that I was able to examine the possible moderating effect of supervisor-level characteristics on only one type of subordinate assessment. This assessment was created and customized for the manufacturing industry. Subordinates had only completed one assessment while on the job and, due to organizational constraints, I was only permitted to use existing predictor data for my study. To address this, I looked at the effect of supervisor characteristics on validity coefficients on several scales within the assessment that captured a wide variety of competencies including cognitive and non-cognitive predictors as well as interpersonal skills.
Another possible limitation is that the criterion measure, or subordinate performance ratings, was collected at 2 different times. Some ratings were collected in 2006 as part of a concurrent validation study, the others were collected in 2007 as part of a predictive validation study. However, when performance ratings were collected, in each instance, supervisors were given the same instructions and used the same rating form. They were also told that these ratings would be used for research and not administrative purposes. This is important as the belief that ratings are being used for research typically motivates raters to be more honest and accurate when assigning ratings (Zedeck and Cascio, 1982; McIntyre, Smith, and Hassett, 1984).

One limitation of multilevel modeling is that it assumes all subordinates are comparable across all supervisors. In my study this may not necessarily be the case as some areas of the plant are more desirable than others, and subordinates compete to work in specialized areas. As a result, those subordinates working under supervisors in these sought-after divisions may be more qualified and the performance ratings may reflect actual differences in performance rather than differences in rater characteristics. During the analysis I controlled for the subordinate characteristics for which I had data, in an attempt to address this concern.

Implications

Previous research, though limited, has found that characteristics of supervisors influence the ratings they assign to subordinates (Schmitt, Pulakos, Nason, and Whitney, 1996; Tziner, Murphy, and Cleveland, 2005). It would follow, and additional researchers suggest, that those non-performance related, systematic effects on performance ratings would ultimately influence validity coefficients (Murphy and Cleveland, 1995). However, there is no research that uses actual applied data to examine this potential effect. The dearth of research into this area is likely a result of methodological constraints (e.g., the fact that the rater characteristics lie at a higher
level than the ratings themselves). Multilevel problems and data typically violate the assumptions of independence of observations, usually required for most statistical tests. For example, two subordinates under the same supervisor – Supervisor 1 – are likely more similar than one subordinate under Supervisor 1 and another subordinate under Supervisor 2. This study extends the current body of research as it represents one of the first to use Multilevel Modeling to analyze the effects of supervisor level characteristics on the validity coefficients of level 1 predictors.

Another important implication of this study is that I found systematic variability in the validity coefficient for the level 1 predictor Safety. The majority of supervisor characteristics examined in this study did not significantly account for the variance in Safety’s validity. However, though it was not in the hypothesized direction, I was able to predict this variance with a measure of supervisor cognitive ability (as measured by Interpreting Information in the current study). I found that as supervisors’ cognitive ability increased, the validity of the predictor Safety decreased.

This study adds to existing research in the effects of supervisor characteristics on performance ratings. Specifically, my results support previous research that suggests that more intelligent supervisors may be less likely to commit typical rater errors such as inflating their ratings (Bernardin & Villanova, 1986). Findings from my study suggest that as supervisors’ level of intelligence (the competency Interpreting Information) increases, their average overall performance ratings decreased.

My results also indicate that as supervisors’ level of Adaptability increases, their average performance rating increases as well. This supports findings by Bartles and Doverspike (1997) which suggest that raters who are high in Warmheartedness and Sensitivity are more likely to
inflate their ratings. Murphy and Cleveland (1995) and Murphy, Cleveland, Skateboo, and Kinney (2004) would suggest that this is because their goals when assigning ratings are not necessarily to rate accurately, but to keep relations smooth. Accordingly, my study does lend support to the position that raters with different goals and motivations may in fact rate differently, which ultimately may affect various organizational outcomes.

**Future Research**

This study examined the effects of rater characteristics on average performance ratings assigned and validity coefficients of level 1 predictors in a manufacturing environment. It is possible that the effects of rater characteristics and goals may be more pronounced in other industries with different organizational structure and culture. In this particular instance there was a strong union in place in all but one of the locations where data was collected. Much of this performance rating data was collected shortly after union members went on strike and the plant needed to hire a large amount of new employees to replace the striking workers. In an effort to obtain honest, accurate ratings of subordinate performance, raters were assured the ratings collected would be used for research purposes only. However, the subordinate-supervisor relationship was somewhat contentious and there was general distrust in organizational leadership. It is possible that supervisors in this environment were less likely to rate accurately because of their subordinates’ membership in a union and because of the strong organizational distrust. It may be interesting to examine the effects of rater characteristics, and the influence of rater goals, in an environment that is not unionized, and where raters feel more secure assigning ratings reflective of actual performance.

I examined rater variables such as cognitive ability and conscientiousness as research suggests they are important predictors of performance. I also analyzed the effect of additional
rater demographics such as tenure and number of subordinates. Future researchers could examine the effect of other potentially important rater characteristics. For example, it may be interesting to analyze the effect of rater gender on the gender-performance relationship of the ratees to determine if the raters assign inflated (i.e., less accurate) ratings to subordinates of the same gender; these results may be even more pronounced in industries that are traditionally male dominated such as manufacturing.

Research suggests that, when assigning ratings, raters may be trying to accomplish a myriad of other goals and may not be trying to provide an accurate assessment of their subordinates’ performance (Murphy and Cleveland 1995). In this study, I attempted to use rater characteristics (and their ratings) as proxies for rater goals. For example, I surmised that those raters higher in Adaptability may have the goal of motivating their employees or keeping the relations within the team smooth. Future research could use alternative methods to ascertain the actual goals of the rater. Perhaps an item could be included on the performance rating measure that allows raters to select or list their goal when assigning the rating(s). Then the rater goals could be evaluated directly, as a level 2 predictor, for their effect on performance ratings and validity coefficients.
References


Table 1

Means, Standard Deviations, and Intercorrelations of Subordinate Study Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Attention to Detail</td>
<td>5.87</td>
<td>1.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Positive Attitude</td>
<td>6.12</td>
<td>1.95</td>
<td>.40**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Process Monitoring</td>
<td>5.81</td>
<td>2.17</td>
<td>.52**</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Qualitative Problem Solving</td>
<td>5.74</td>
<td>2.09</td>
<td>.18**</td>
<td>.12**</td>
<td>.25**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Quantitative Problem Solving</td>
<td>6.46</td>
<td>1.47</td>
<td>.33**</td>
<td>.18**</td>
<td>.34**</td>
<td>.44**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Responsibility</td>
<td>6.03</td>
<td>2.54</td>
<td>.43**</td>
<td>.38**</td>
<td>.07</td>
<td>.10**</td>
<td>.11*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Risk Reliability</td>
<td>6.73</td>
<td>2.02</td>
<td>.38**</td>
<td>.51**</td>
<td>.03</td>
<td>-.04</td>
<td>.05</td>
<td>.36**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Safety</td>
<td>7.67</td>
<td>1.63</td>
<td>.19**</td>
<td>.21**</td>
<td>-.11*</td>
<td>.02</td>
<td>-.02</td>
<td>.26**</td>
<td>.28**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Teamwork</td>
<td>6.07</td>
<td>1.14</td>
<td>.24**</td>
<td>.27**</td>
<td>.15**</td>
<td>.60**</td>
<td>.21**</td>
<td>.19**</td>
<td>.20**</td>
<td>.14**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Conscientiousness</td>
<td>6.14</td>
<td>1.66</td>
<td>.37**</td>
<td>.57**</td>
<td>-.04</td>
<td>-.05</td>
<td>.02</td>
<td>.39**</td>
<td>.55**</td>
<td>.36**</td>
<td>.18**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Perceptual Speed</td>
<td>6.09</td>
<td>1.48</td>
<td>.49**</td>
<td>.27**</td>
<td>.79**</td>
<td>.23**</td>
<td>.31**</td>
<td>.16**</td>
<td>.20**</td>
<td>.04</td>
<td>.19**</td>
<td>.27**</td>
<td></td>
</tr>
<tr>
<td>12. Average Job Performance</td>
<td>4.9</td>
<td>.93</td>
<td>.13**</td>
<td>.07</td>
<td>.10*</td>
<td>.06</td>
<td>.10*</td>
<td>.03</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.08</td>
<td>.08</td>
</tr>
</tbody>
</table>
Table 2

*Means, Standard Deviations, and Intercorrelations of Supervisor Study Variables*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Interpreting Information</td>
<td>7.4</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Business Planning</td>
<td>4.04</td>
<td>.52</td>
<td>-.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Number of Subordinates</td>
<td>6.02</td>
<td>6.45</td>
<td>.04</td>
<td>-.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Length of Supervision</td>
<td>.79</td>
<td>1.05</td>
<td>.07</td>
<td>-.06</td>
<td>-.28*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Promotability</td>
<td>.26</td>
<td>.44</td>
<td>-.16</td>
<td>-.31**</td>
<td>.65**</td>
<td>-.26*</td>
<td></td>
</tr>
<tr>
<td>6. Adaptability</td>
<td>3.93</td>
<td>2.46</td>
<td>.11</td>
<td>.49**</td>
<td>-.03</td>
<td>.08</td>
<td>-.01</td>
</tr>
</tbody>
</table>

*Note. Promotability was coded so they received a one if they had been promoted between the time they assigned the rating to their subordinate and the data was analyzed, zero if they were still at the same level. Length of Supervision was coded such that 0 = 3-6 months, 1 = 6-9 months, 2 = 9-12 months, 3 = 12-15 months, and 4 = 16 months or longer.*
Table 3

*Estimated Variance Components (Random Effects)*

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>$X^2$</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Attention to Detail</td>
<td>2.05</td>
<td>2</td>
</tr>
<tr>
<td>2. Conscientiousness</td>
<td>1.56</td>
<td>2</td>
</tr>
<tr>
<td>3. Process Monitoring</td>
<td>2.21</td>
<td>2</td>
</tr>
<tr>
<td>4. Positive Attitude</td>
<td>2.77</td>
<td>2</td>
</tr>
<tr>
<td>5. Qualitative Problem Solving</td>
<td>.21</td>
<td>2</td>
</tr>
<tr>
<td>6. Quantitative Problem Solving</td>
<td>.36</td>
<td>2</td>
</tr>
<tr>
<td>7. Responsibility</td>
<td>.76</td>
<td>2</td>
</tr>
<tr>
<td>8. Risk Reliability</td>
<td>.92</td>
<td>2</td>
</tr>
<tr>
<td>9. Safety</td>
<td>6.36*</td>
<td>2</td>
</tr>
<tr>
<td>10. Teamwork</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td>11. Perceptual Speed</td>
<td>1.99</td>
<td>2</td>
</tr>
</tbody>
</table>

*Note.  * $p < .05.*
## Supervisor Competencies Predicting the Intercept

<table>
<thead>
<tr>
<th>Supervisor Competency</th>
<th>$\gamma$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpreting Information</td>
<td>-.12**</td>
<td>.04</td>
</tr>
<tr>
<td>Business Planning</td>
<td>.04</td>
<td>.16</td>
</tr>
<tr>
<td>Number of Subordinates</td>
<td>-.00</td>
<td>.01</td>
</tr>
<tr>
<td>Length of Supervision</td>
<td>.06</td>
<td>.07</td>
</tr>
<tr>
<td>Promotability</td>
<td>-.32</td>
<td>.20</td>
</tr>
<tr>
<td>Adaptability</td>
<td>.08*</td>
<td>.03</td>
</tr>
</tbody>
</table>

*Note.* $^* p < .05$, $^{**} p < .01$
Table 5

*Effect of Supervisor-level Competencies on Level One Slope*

<table>
<thead>
<tr>
<th>Level 1 Predictor</th>
<th>Interpreting Information $\gamma$</th>
<th>Interpreting Information $SE$</th>
<th>Business Planning $\gamma$</th>
<th>Business Planning $SE$</th>
<th>Number of Subordinates $\gamma$</th>
<th>Number of Subordinates $SE$</th>
<th>Length of Supervision $\gamma$</th>
<th>Length of Supervision $SE$</th>
<th>Promotability $\gamma$</th>
<th>Promotability $SE$</th>
<th>Adaptability $\gamma$</th>
<th>Adaptability $SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process Monitoring</td>
<td>-.01 .01</td>
<td>.04 .04</td>
<td>-.00* .00</td>
<td>.03 .02</td>
<td>-.04 .04</td>
<td>.00 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative Problem Solving</td>
<td>-.00 .01</td>
<td>.01 .03</td>
<td>.00 .00</td>
<td>-.00 .02</td>
<td>.04 .03</td>
<td>-.00 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative Problem Solving</td>
<td>-.01 .01</td>
<td>-.03 .06</td>
<td>-.00 .00</td>
<td>-.01 .03</td>
<td>.02 .05</td>
<td>-.01 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceptual Speed</td>
<td>-.03 .02</td>
<td>.05 .06</td>
<td>-.00 .00</td>
<td>.03 .03</td>
<td>-.06 .05</td>
<td>-.00 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention to Detail</td>
<td>.01 .02</td>
<td>.04 .06</td>
<td>-.01 .00</td>
<td>.08* .04</td>
<td>-.08 .06</td>
<td>.00 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>-.03** .01</td>
<td>.05 .04</td>
<td>-.00 .00</td>
<td>.04 .02</td>
<td>-.03 .04</td>
<td>.01 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsibility</td>
<td>-.00 .01</td>
<td>.07 .04</td>
<td>-.01* .00</td>
<td>.03 .02</td>
<td>-.07 .04</td>
<td>-.01 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Safety</td>
<td>-.04* .02</td>
<td>-.01 .08</td>
<td>-.00 .00</td>
<td>.05 .04</td>
<td>-.07 .07</td>
<td>.01 .02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Attitude</td>
<td>-.01 .01</td>
<td>.01 .04</td>
<td>-.00* .00</td>
<td>.03 .02</td>
<td>-.05 .04</td>
<td>-.00 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teamwork</td>
<td>.01 .01</td>
<td>.05 .05</td>
<td>-.00 .00</td>
<td>.01 .03</td>
<td>-.06 .05</td>
<td>-.02 .01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* *$^* p < .05, **$ p < .01