A Multi-Level Investigation of Overall Job Performance Ratings

Emily C. Johnson
Federal Management Partners

Adam W. Meade
North Carolina State University

Multi-level modeling was used to understand how supervisors assign overall job performance ratings. Results indicated a uniform relationship between task and overall performance ratings across supervisors but significant variability in the relationship between contextual and overall performance ratings. Employee and supervisor attributes were examined to explain this variability.

Job performance is typically conceptualized as “actions and behaviors that are under the control of the individual that contribute to the goals of the organization.” (Rotundo & Sackett, 2002, p. 66). Campbell (1990) asserts that job performance is inherently multidimensional, a view that is shared by the vast majority of performance researchers (e.g., Borman & Motowidlo, 1993; Campbell, Gasser, & Oswald, 1996; Hesketh & Neal, 1999; Motowidlo, Borman, & Schmit, 1997; Murphy & Shiarella, 1997). However, it is also generally accepted that, while multidimensional conceptions of performance are appropriate when conducting research, for decision making in organizations, a unidimensional, or composite criteria is preferred (e.g., Schmidt & Kaplan, 1971). And while the notion that raters will differ in the way they combine information to arrive at an overall rating is far from new (e.g., Naylor & Wherry, 1965), in the job performance domain, research dedicated to explaining this variability is limited.

Performance

Task performance refers to behaviors and activities that support the organization’s technical core, which can involve the execution of technical processes (transforming raw materials into the goods or services provided by the organization) or the maintenance of those processes, for instance by providing raw materials, distributing products, or through planning and coordination functions (Borman & Motowidlo, 1993; Motowidlo, et al., 1997). The term contextual performance was coined by Borman and Motowidlo (1993) who argued that performance measures used in selection research and practice ignored activities such as persisting, helping, and endorsing organizational objectives. They argue that the criterion domain consists of task performance as well as contextual performance, or behaviors that support the broader psychological and social environment in which that technical core must function.

An individual’s overall performance rating can be thought of as a measure of his or her organizational worth (Motowidlo & Van Scotter, 1994). Implicit in this definition is that job performance must include only those behaviors that contribute to the organization’s goals (Campbell, 1990). Accordingly, the rating assigned to an incumbent’s overall job performance will depend not only on the level at which he or she performs certain behaviors, but also on the rater’s beliefs about the goals of the organization and his or her mental model relating job behaviors to those goals. According to role theory, the role expectations that one holds are influenced by characteristics of both the individual and the context in which he or she is embedded (Welbourne, D. E. Johnson, & Erez, 1998). Lievens, Conway, and De Corte (2008) suggest that performance evaluation represents an appraisal of how well a worker’s behavior conforms to the role expectations held by the rater (typically a supervisor). Thus, if a supervisor does not believe that contextual performance behaviors are an important aspect of his or her employees’ job role, these behaviors will not weigh heavily into the supervisor’s assessments.

The current study uses MLM to investigate the influence of various factors on ratings of overall job performance in a sample of law enforcement officers. In particular, we expect that many of these factors will play a role in whether or not supervisors will view contextual performance as role-required and, as such, will affect its weight in supervisors’ evaluations of their employees’ performance.

Level 1 Variables

The relationship between overall performance ratings and both task and contextual
Multilevel Performance

performance has been well-established (Bergeron, 2007; Podsakoff, MacKenzie, Paine, & Bachrach, 2000). Based on the substantial empirical evidence that task performance and contextual performance are distinct constructs (e.g., LePine, Erez, & D. E. Johnson, 2002), both of which contribute to ratings of overall performance (e.g., Borman, White, & Dorsey, 1995; Conway, 1999; Motowidlo & Van Scotter, 1994), we predict:

**Hypothesis 1:** Ratings of both task performance and contextual performance will be positively related to overall performance ratings.

**Research Question 1:** What proportion of the variance in overall job performance ratings can be attributed to task vs. contextual performance?

Few attempts have been made to determine whether the relationships between task and contextual performance and overall performance ratings vary among raters. In one of the few attempts to do so, Rotundo and Sackett (2002) used a policy capturing approach to demonstrate that raters do not share a common policy for weighting performance dimensions. Lievens et al. (2008) also found that organizational culture accounted for significant observed variance in raters’ policies for combining performance dimensions. However, rating the described behavior of a paper person clearly involves a fundamentally different cognitive process compared to the task of rating actual, observed performance of real-world employees (Murphy, Herr, Lockhart, & Maguire, 1986). In the current study, each supervisor provided ratings of multiple, real incumbents.

Because task performance consists of activities and behaviors that are publically recognized by the organization as defining aspects of the job (i.e., what the employee has been hired to do), we expect that the organization’s policy will dictate the relationship between an employee’s task performance and his or her overall job performance.

**Hypothesis 2(a):** The strength of the relationship between task performance and overall performance will not vary across supervisors.

On the other hand, contextual behaviors are less likely than task behaviors to be formally required or clearly linked to rewards (Organ, 1997). As such, within an organization individuals will likely vary in their beliefs about how contextual performance relates to overall performance.

**Hypothesis 2(b):** The strength of the relationship between contextual performance and overall performance ratings will vary across supervisors.

**Other level-1 influences.** Another employee-level variable that might influence ratings of overall performance is work experience. Quiñones, Ford, and Teachout’s (1995) meta-analysis demonstrated that, regardless of the operationalization used, work experience tends to be positively related to overall job performance (estimated $r=.27$).

**Hypothesis 3:** Employee tenure will be positively related to overall performance ratings.

Moreover, J. W. Johnson (2001) suggests that an employee’s level of experience may influence how his or her supervisor weights the dimensions. Organ (1988) theorized that as individuals move up the organizational ranks, the breadth of their roles expands. We predict that role expansion will occur as employee tenure increases, such that an employee’s contextual performance will be weighted more heavily as that employee becomes more experienced.

**Hypothesis 4:** Employee time on the job will moderate the relationship between contextual performance and overall performance ratings, such that the relationship is stronger for those with more (vs. less) time on the job.

**Research Question 2:** Is opportunity to observe related to overall performance ratings?

**Level 2 Variables**

**Supervisor experience.** Research findings regarding supervisor experience have been mixed, with some studies reporting that supervisors with more tenure tend to be more lenient, and others finding no effect (Landy & Farr, 1980).

**Research Question 3:** Will supervisor experience (time in current rank) positively influence overall performance ratings?

**Research Question 4:** Does the relationship between contextual and overall performance ratings depend on how long the supervisor providing those ratings has held the rank of sergeant?

**Method**

**Participants**

The sample consisted of North Carolina State Highway Patrol (NCSHP) troopers and their supervising sergeants selected to participate in a
recent selection system revalidation effort (see E. C. Johnson, DuVernet, & Wilson, 2008).

**Supervisors.** Only those supervising sergeants who provided usable performance ratings for at least two troopers were included in the analyses. Among those sergeants (N = 86), the number of troopers rated ranged from 2 to 13, with a median = 3 (M = 3.72, SD = 2.07). The majority reported that, at the time of rating, they had been in the role of sergeant for over three years. While no demographic information is available for the supervisor sample, there is no reason to expect that they differed substantially from the total sergeant population (N = 222), which was 16.2% African American, 1.8% Native American, 81.5% Caucasian, and 96.4% male at the time of data collection.

**Employees.** Usable performance ratings were available for 324 troopers. Time between being sworn in as a trooper and the date of the evaluation ranged from 1.96 years to 7.81 years (M = 4.85, SD = 1.85). The sample was approximately 98% male, with Caucasians making up 86% of the total and African Americans comprising an additional 9%.

**Measures & Analyses**

An online rating website instructed raters that all ratings were being collected for research purposes only and that no ratings would be seen by any member of the NCSHP. Task Performance scores were computed as the average rating across 13 detailed task statements derived from a comprehensive job analysis. Contextual performance ratings were calculated as the average rating across 19 items (Podsakoff, et al., 1990) designed to assess four of dimensions of citizenship performance proposed by Organ (1988): Altruism, Conscientiousness, Civic Virtue, and Courtesy. Both task and contextual performance scores were grand-mean centered. Supervisors also provided ratings of troopers’ overall performance, by responding to a single item, “After considering everything you know about this Trooper, how would you rate his/her overall performance?” Multilevel models tested and their results are presented in the results section.

**Results**

**Fully Unconditional Model**

The simplest possible (fully unconditional) model, is typically assessed as the first step in MLM analysis (Raudenbush & Bryk, 2002) and contains no predictor variables. The criterion variable in all models individual troopers’ overall performance. The results of the unconditional model indicated significant between-supervisor variance component ($\tau_{10} = .010, z = 3.01, p < .01$), and a significant within-supervisor component ($\sigma^2 = 0.40, z = 11.04, p < .01$). The ICC indicated that 20.3% of variability in ratings of overall performance occurs at the supervisor level, while 79.7% was at the trooper level.

**Level-1 Models**

Level-1 models were tested using the recommended “step-up” approach (Raudenbush & Bryk, 2002), in which variables are added one at a time. The first model included task performance as a predictor of overall performance. The average slope relating trooper task was significant ($t = .96, n(237) = 25.64, p < .01$), providing support for Hypothesis 1. A second model included contextual performance as the only predictor and was significant ($t = .55, n(237) = 9.87, p < .01$). Taken together with the results of the task performance model, this result provides full support for Hypothesis 1.

Research Question 1, which concerns the variance in overall performance ratings attributable to task and contextual performance was tested using the following calculation:

$$R^2_{\text{within}} = \frac{\sigma^2_{\text{uc}} - \sigma^2_{\text{u}}}{{\sigma^2_{\text{u}}}},$$

where $\sigma^2_{\text{uc}}$ is the residual level-1 variance in the unconditional model (i.e., the model with no predictors) and $\sigma^2_{\text{u}}$ is the residual level-1 variance for the model which includes the predictor of interest. The results indicate that task performance ratings alone account for 64.69% of the variance ratings while the variability accounted for by contextual performance alone is 54.02%.

As predicted by Hypothesis 2(a), the slope relating task performance to overall performance did not vary significantly across supervisors ($t_{11} = .004, Z = .42, p = .34$). Conversely, as predicted by Hypothesis 2(b), the results of the contextual performance model indicated that the slope relating contextual performance to overall performance does vary significantly across supervisors ($t_{11} = .11, Z = 2.91, p < .01$). Based on these results, a model was specified in which both task performance and contextual performance are included as level-1 predictors of overall performance, with the contextual performance slope, allowed to vary across level-2 units. The results of this model indicate that, together, task and contextual performance account for 73.52% of the trooper-level variability in overall performance ratings (see Table 1). Level-1 residual variance remained significant ($\sigma^2 = 0.11, z = 9.79, p < .01$), thus, additional potential level-1 predictors of overall performance ratings were investigated.

Hypotheses 3 and 4 stated that trooper tenure would be positively related to overall performance and that tenure will moderate the relationship between contextual performance and
overall performance. Neither hypothesis was supported as trooper tenure did not have a significant main effect on performance ratings ($\gamma_{00} = .00$, $t(234) = .72$, $p = .47$) nor moderate the relationship between contextual performance and overall performance ($\gamma_{40} = .00$, $t(234) = .6$, $p = .55$).

To answer Research Question 2, which asks whether a supervisor’s opportunity to observe a trooper will be related to overall performance ratings, a model was tested in which trooper tenure was removed from the above model and opportunity to observe was added. The results of this model indicate that opportunity to observe has no appreciable effect on overall performance ratings ($\gamma_{00} = .01$, $t(235) = .34$, $p = .74$).

**Level-2 Models**

**Intercepts-as-outcomes model.** Research Question 3 specified a cross-level main effect, namely, that a supervisor’s amount of experience (operationalized as self-reported time in current rank) would be positively related to ratings of overall performance. The results of this model indicated that supervisor experience does not significantly affect overall performance ratings ($\gamma_{01} = -.01$, $t(84) = -.47$, $p = .64$).

**Slopes-as-outcomes model.** Research Question 4 asked whether the relationship between contextual performance and overall performance depends on supervisor experience. The results of this model indicated that supervisor experience does not moderate the relationship between ratings of contextual and ratings of overall performance ($\gamma_{21} = -.01$, $t(235) = -.36$, $p = .72$).

**Discussion**

The results indicated that, within a single organization, the relationship between task performance and overall performance ratings is generally consistent across raters. We reasoned that supervisors would share a common set of expectations regarding the task requirements of the job and, as such, evaluate task performance as relating to overall performance in largely the same way. The finding that the task performance-overall rating slope did not vary significantly across raters is supportive of this line of reasoning.

On the other hand, the relationship between contextual performance and overall performance varied significantly across raters. Contextual performance behaviors are far less likely than task behaviors to be formally acknowledged by the organization as job requirements (Organ, 1997). Consequently, supervisors receive less information from the organization regarding how and to what extent their employees’ contextual performance is related to successful job performance. The finding of variable slopes relating contextual performance to overall performance ratings at the supervisor level suggests that, unlike task performance, beliefs about the contribution of contextual performance are idiosyncratic to individual supervisors. Thus, it may be the case that, in the absence of information from the organization, evaluations of employee job performance are informed by the rater’s own implicit theories regarding contextual performance. Taken together, these findings suggest an interesting distinction between the two dimensions of performance. Namely, perhaps the value of task behaviors, in terms of their contribution to the organization’s overall goals, is understood across members of the organization, while the perceived value of contextual behavior is more idiosyncratic to individual organizational members.

For researchers, the notion that different individuals may use performance information differently when rating overall performance has some serious implications for the use of overall performance ratings as criteria in test validation research. Murphy and Shiarella (1997) demonstrated that the validity of selection tests can vary considerably depending on how performance dimensions are weighted to form an overall performance composite. In addition, procedures such as meta-analysis assume that measures of overall performance assess the same construct across studies. In the absence of research that identifies specific characteristics of raters and work context that might be expected to affect overall evaluations, researchers should proceed with caution when attempting to draw conclusions from research which relies on overall performance ratings as a criterion measure.

Within organizations, interventions might be designed with the goal of developing the same type of shared understanding as to how contextual behaviors relate to employees’ performance of their job roles as is currently apparent regarding task behaviors. This would require increased formalization of the contextual performance requirements of jobs, through job descriptions and appraisal tools. Providing rater training on the new
instruments would increase the salience of such behaviors and enhance the likelihood that supervisors evaluate their employees using the same standards for successful performance. Alternatively, the use of raters might be circumvented entirely in determinations of overall performance through the use of statistical procedures for combining dimension ratings.

An additional aim of the current study was to highlight the potential benefits of the MLM approach for job performance research. This methodology allows multiple levels of influences on job performance ratings to be investigated, thus providing a more complete picture of the performance appraisal process. MLM can serve as an invaluable tool for furthering our understanding of performance appraisal as a process that is influenced not only by the employee being evaluated, but also by the employee’s supervisor and the performance context. Moreover, hierarchical models allow that ratings of different employees made by the same supervisor may be more similar than ratings of different employees made by different supervisors, by estimating separate residuals at the employee and supervisor level of analysis. When collecting performance data in organizations, departmental structure often simply does not allow each supervisor to rate only a single employee (Judge & Ferris, 1993). When employees are nested within supervisors or work contexts, MLM provides a more statistically sound approach to analyzing performance data, as compared to OLS regression, which assumes independence of observations (Raudenbush & Bryk, 2002).

**Limitations and Future Directions**

Future research should address rater and contextual characteristics that influence the contextual performance-overall performance relationship. The finding of significant level-2 variance in slopes indicates that this is a promising direction for research. Although the current study attempted to explain this variance, none of the hypotheses were supported. One reason for this failure to find a relationship may be that the variables included were simply not the correct variables. For example, supervisor time in rank fails to capture specific experiences that may have important influences on how supervisors come to think about the role of contextual performance. To this end, measuring work environment demands more directly or variables such as supervisor personality or leadership style may provide interesting results.

Finally, it remains to be seen whether the conclusions drawn from the present study are specific to the law enforcement population or whether they can be supported across work environments. Because the notion that contextual performance tends to be less clearly linked to organizational rewards than is task performance (Organ, 1997) is not a law enforcement-specific, it seems likely that the slope relating contextual performance to overall performance varies across raters in a variety of organizational settings. Similarly, to the extent to which organizations send clear messages about the task demands of a job, one would expect to find the same sort of consistency in task-overall performance slopes observed in the present sample. However, further research is needed to investigate the generalizability of the current study’s findings.

**Conclusions**

Recent studies suggest that raters are guided by unique policies regarding how different aspects of performance should be combined to produce an overall rating (e.g., Lievens et al., 2008; Rotundo & Sackett, 2002). The current study builds on existing research by using a MLM approach to show that the relationship between contextual performance and overall performance is more profoundly affected by the individual rater differences than is the relationship between task performance and overall performance. Additionally, the MLM approach used in the current study overcame many limitations of policy capturing studies, which involve rating the performance of hypothetical employees (see Murphy et al., 1986), by allowing the effects of real organizational forces to be investigated. Research is needed to examine how individual and contextual factors influence the weight given to contextual performance as well as to determine whether there are organizational factors (such as socialization or culture) that increase the likelihood that raters will perceive the contribution of task behaviors to overall job performance similarly.

**References**


**Author Contact Info:**

**Emily Johnson**  
E-mail: ecjohnso@gmail.com

**Adam W. Meade**  
Department of Psychology  
North Carolina State University  
Campus Box 7650  
Raleigh, NC 27695-7650  
Phone: 919-513-4857  
Fax: 919-515-1716  
E-mail: awmeade@ncsu.edu
Table 1

Results of Task and Contextual Performance Model

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>df</th>
<th>t Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $\gamma_{00}$</td>
<td>5.002</td>
<td>0.026</td>
<td>85</td>
<td>195.26</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Task performance, $\gamma_{10}$</td>
<td>0.891</td>
<td>0.043</td>
<td>236</td>
<td>20.49</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Contextual performance, $\gamma_{20}$</td>
<td>0.156</td>
<td>0.041</td>
<td>236</td>
<td>3.81</td>
<td>.0002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>Z Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{00}$</td>
<td>0.019</td>
<td>0.010</td>
<td>1.94</td>
<td>0.0259</td>
</tr>
<tr>
<td>$\tau_{10}$</td>
<td>0.004</td>
<td>0.008</td>
<td>0.44</td>
<td>0.6592</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>0.035</td>
<td>0.014</td>
<td>2.58</td>
<td>0.005</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.107</td>
<td>0.011</td>
<td>9.79</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Notes.* $\tau_{00} =$ unconditional variance in the level-1 intercepts; $\tau_{10} =$ unconditional covariance between the level-1 slopes and intercepts; $\tau_{11} =$ unconditional variance in the level-1 slopes. $\sigma^2 =$ residual level-1 variance.