The Relative Importance of Task, Citizenship, and Counterproductive Performance to Global Ratings of Job Performance: A Policy-Capturing Approach

Maria Rotundo
University of Toronto

Paul R. Sackett
University of Minnesota, Twin Cities Campus

A review of research on job performance suggests 3 broad components: task, citizenship, and counterproductive performance. This study examined the relative importance of each component to ratings of overall performance by using an experimental policy-capturing design. Managers in 5 jobs read hypothetical profiles describing employees' task, citizenship, and counterproductive performance and provided global ratings of performance. Within-subjects regression analyses indicated that the weights given to the 3 performance components varied across raters. Hierarchical cluster analyses indicated that raters' policies could be grouped into 3 homogeneous clusters: (a) task performance weighted highest, (b) counterproductive performance weighted highest, and (c) equal and large weights given to task and counterproductive performance. Hierarchical linear modeling indicated that demographic variables were not related to raters' weights.

For decades, industrial and organizational psychologists encouraged researchers to develop the criterion and identified key issues involved in this process (Astin, 1964; Dunnette, 1963; Jenkins, 1946; Nagle, 1953; Patterson, 1946; Thorndike, 1949). Today, progress has been made in defining the criterion and understanding its underlying structure (Campbell, 1990; Campbell, McCloy, Oppler, & Sager, 1993; Campbell, McHenry, & Wise, 1990; Murphy, 1989). This article presents a review of these models indicating that there are three distinct groups of behaviors that constitute the domain of job performance. These behaviors include task, citizenship, and counterproductive performance. We examined the relative importance of the three components to supervisory ratings of overall job performance by using an experimental policy-capturing design.

There are a variety of issues that research on task, citizenship, and counterproductive performance attempts to resolve. Researchers and practitioners are interested in identifying which individual characteristics predict whether employees engage in these behaviors. A considerable amount of research has investigated antecedents and consequences of task, citizenship, and counterproductive behavior (e.g., Giacalone & Greenberg, 1997; Hollinger & Clark, 1983; Konovsky & Organ, 1995; Murphy & Shiarella, 1997; Ones, Viswesvaran, & Schmidt, 1993; Organ & Lingl, 1995; Organ &

Ryan, 1995; C. A. Smith, Organ, & Near, 1983). However, interest in predicting these behaviors stems from the assumption that managers and organizations value these behaviors. Therefore, another issue in which researchers and practitioners are interested is identifying which aspects of employees' performance are rewarded by managers in performance evaluations. A limited amount of research has investigated the role that these components play in performance evaluations (e.g., Conway, 1999; MacKenzie, Podsakoff, & Fetter, 1991; Motowidlo & Van Scotter, 1994). The research that does exist focused on task and citizenship performance and did not examine counterproductive behavior as a separate construct. In this study, we sought to achieve a better understanding of which aspects of employees' performance are valued most by managers.

Theoretical Development

Seminal papers dating back to the 1970s defined job performance in terms of actions and behaviors rather than the results of these actions (Campbell, 1990; Murphy, 1989; P. C. Smith, 1976). Definitions also focused on behaviors that affect the goals of the organization and are under the control of the individual, with the latter condition precluding behaviors constrained by the environment. On the basis of these ideas, job performance is conceptualized as those actions and behaviors that are under the control of the individual and contribute to the goals of the organization.

A number of behaviors fall under this broad definition of job performance. Therefore, it is important to determine the structure underlying the criterion. Some attempts have been made to model the entire domain of job performance (Borman & Brush, 1993; Borman & Motowidlo, 1993; Campbell, 1990; Campbell et al., 1990; Katz & Kahn, 1978; Murphy, 1989; Welbourne, Johnson, & Erez, 1998). These attempts are summarized in Table 1. Another

Table 1
A Summary of Various Efforts to Describe the Domain of Job Performance

Reference	Component	Description
Katz & Kahn (1978)	Role performance in system	Meeting or exceeding the quantitative and qualitative standards of performance
	Innovative or spontaneous behaviors	Facilitate the achievement of organizational goals, cooperating, protecting the organization
	Joining and staying with the organization	Low turnover and absenteeism
Murphy (1989)	Task performance	Accomplishment of duties and responsibilities
	Interpersonal relations	Cooperating, communicating, exchanging job-related information
	Destructive or hazardous behaviors	Violating security and safety, destroying equipment, accidents
	Downtime behaviors	Substance abuse, illegal activities
Campbell (1990)	Job-specific task proficiency	Core technical tasks
	Non-job-specific task proficiency	Tasks not specific to a given job
	Written and oral communication proficiency	Preparing written materials or giving oral presentations
	Demonstrating effort	Exerting extra effort, willing to work under adverse conditions
	Maintaining personal discipline	Avoid negative or adverse behaviors (e.g., substance abuse)
	Facilitating peer and team performance	Support and assist peers, reinforce participation
	Supervision and leadership	Influence, setting goals, rewarding and punishing
	Management and administration	Organize people and resources, monitor progress, problem solve
Borman & Motowidlo (1993)	Task performance	Formally recognized as part of the job and contribute to the organization's technical core
	Contextual performance	Discretionary, not necessarily role-prescribed, contribute to social and psychological environment
Borman & Brush (1993)	Technical activities	Planning, demonstrating technical proficiency, administration
	Leadership and supervision	Guiding, directing, motivating, coordinating
	Interpersonal dealings	Communicating, maintaining a good organizational image and working relationships
	Useful personal behavior	Working within the guidelines and boundaries of the organization
Welbourne, Johnson, & Erez (1998)	Job	Doing things specifically related to one's job description
	Career	Obtaining the necessary skills to progress through one's organization
	Innovator	Creativity and innovation in one's job and the organization as a whole
	Team	Working with coworkers and team members, toward success of the firm
	Organization	Going above the call of duty in one's concern for the firm

group of researchers have focused on specific performance components (Brief & Motowidlo, 1986; Crino, 1994; George & Brief, 1992; Hollinger & Clark, 1982; Hunt, 1996; Organ, 1988, 1997; Raelin, 1994; Robinson & Bennett, 1995; Robinson & Greenberg, 1998; Van Dyne, Cummings, & Parks, 1995). These efforts are summarized in Table 2. A review of this literature indicates that job performance can be described by three broad performance components.

Task Performance

Each conceptualization of performance presented in Table 1 makes reference to a group of behaviors involved in the completion of tasks. Katz and Kahn (1978) defined *role performance in system* as meeting or exceeding the quantitative and qualitative standards of performance. Welbourne et al. (1998) defined *job role* as the quantity and quality of work output. Murphy (1989) defined *task performance* as the accomplishment of duties and responsibilities associated with a given job. Campbell (1990) used the terms *job-specific* and *non-job-specific task proficiency* to describe actions and behaviors engaged in for the purpose of completing technical tasks. Borman and Motowidlo (1993) defined *task performance* as activities that are formally recognized as part of the job and that contribute to the organization's technical core. Borman and Brush (1993) used the term *technical activities* to describe behaviors that demonstrate technical proficiency.

These conceptualizations of task performance include two central features. They require that activities contribute to the technical core (i.e., the process by which raw materials are converted into the products in which the organization specializes; Borman & Motowidlo, 1993, p. 92) and be formally recognized as part of the job. The notion of *contributing to the technical core* is an important feature that helps distinguish this performance component from the others. However, requiring that behavior be formally recognized as part of the job makes it difficult to compare task performance across organizations because formal requirements vary from one organization to another. For the purposes of the present study, *task performance* includes behaviors that contribute to the production of a good or the provision of a service. However, the definition is not restricted to include only those behaviors that are listed in the job description.

Citizenship Performance

The domain of task performance explains an important part of the criterion. However, researchers believe that it is not sufficient to study only task-related behaviors. Another group of activities that are not necessarily task-related but that contribute to the organization in a positive way are included in Tables 1 and 2. *Demonstrating effort*, *facilitating peer and team performance* (Campbell, 1990), *altruism*, *conscientiousness* (Organ, 1988), *organizational role* (Welbourne et al., 1998), and *affiliative–promotive* behavior (Van Dyne et al., 1995) have

Table 2
A Summary of Various Efforts to Conceptualize Citizenship and Counterproductive Performance

Reference	Component	Behavioral category
Brief & Motowidlo (1986)	Prosocial organizational behavior	Assisting coworkers with job-related matters
		Showing leniency
		Providing services or products to consumers in organizationally consistent ways
		Providing services or products to consumers in organizationally inconsistent ways
		Helping consumers with personal matters unrelated to organizational services or products
		Complying with organizational values, policies, and regulations
		Suggesting procedural, administrative, or organizational improvements
		Objecting to improper directives, procedures, or policies
		Putting forth extra effort on the job
		Volunteering for additional assignments
		Staying with the organization despite temporary hardships
		Representing the organization favorably
		Assisting coworkers with personal matters
Organ (1988)	Organizational citizenship behavior	Altruism
		Conscientiousness
		Sportsmanship
		Courtesy
		Civic virtue
George & Brief (1992)	Organizational spontaneity	Helping coworkers
		Protecting the organization
		Making constructive suggestions
		Developing oneself
		Spreading goodwill
Raelin (1994)	Professional deviant–adaptive	Work scale (e.g., unethical practices, absenteeism, work-to-rule, bootlegging)
		Self scale (e.g., flaunting of external offers, rationalization, alienation, apathy)
		Career scale (e.g., premature external search, external performance emphasis)
Van Dyne, Cummings, & Parks (1995)	Extrarole behavior	Affiliative–promotive (e.g., helping and cooperative behaviors)
		Challenging–promotive (e.g., constructive expression of challenge)
		Challenging–prohibitive (e.g., criticism of situation to stop inappropriate behavior)
		Affiliative–prohibitive (e.g., unequal power or authority)
Robinson & Bennett (1995)	Employee deviance	Property deviance
		Production deviance
		Political deviance
		Personal aggression
Hunt (1996)	Generic work behaviors	Adherence to confrontational rules
		Industriousness
		Thoroughness
		Schedule flexibility
		Attendance
		Off-task behavior
		Unruliness
		Theft
		Drug misuse

been defined as showing perseverance and helping and supporting peers. *Interpersonal relations* (Murphy, 1989), *interpersonal dealings and communication* (Borman & Brush, 1993), and *courtesy* (Organ, 1988) describe those behaviors related to cooperating, communicating, and exchanging job-related information. *Civic virtue* (Organ, 1988), *making constructive suggestions*, *spreading goodwill* (George & Brief, 1992), and *endorsing, supporting, and defending organizational objectives* (Borman & Motowidlo, 1993) describe behaviors related to participating in the political life of the organization and promoting a favorable organizational image.

A number of researchers have defined and explicated this group of behaviors. However, a review of Tables 1 and 2 indicates that the end product of these efforts is separate taxonomic structures with partially overlapping definitions and behavioral categories. These muddied waters can be explained in part by variation among researchers on the defining features of

citizenship performance. For example, when trying to differentiate between task and citizenship performance, researchers often use the following criteria: whether the behavior is inrole, part of the job description, or rewarded (e.g., Katz & Kahn, 1978; Van Dyne et al., 1995). These criteria are problematic because the same behavior may be inrole or rewarded in one organization and extrarole in another (Schmidt, 1993). Regardless of the situation, the behavior itself is the same. When a decision is made about whether to include a behavior in the citizenship domain, an emphasis should be placed on the behavior itself rather than the context in which the behavior occurs. Therefore, we conceptualized this domain in a manner that is different from previous research in this area. We did not rely on whether behavior is part of a job description or rewarded when defining this domain and when classifying behaviors. On the basis of these ideas, the working definition of *citizenship performance* for this study was behavior that contributes to the

goals of the organization by contributing to its social and psychological environment.

Counterproductive Performance

An increasing concern among organizations is counterproductive employee behavior (e.g., Hollinger, Slora, & Terris, 1992). Research attention has been devoted to defining this construct and to determining its underlying structure in an effort to predict counterproductive behavior. The conceptualizations of job performance presented in Tables 1 and 2 include a group of behaviors that detract from the goals of the organization. *Personal deviance* (Robinson & Bennett, 1995), *downtime behaviors* (Murphy, 1989), and *maintaining personal discipline* (Campbell, 1990) describe behaviors related to substance abuse or poor self-discipline. *Destructive–hazardous behaviors* (Murphy, 1989) and *property deviance* (Robinson & Bennett, 1995) describe behaviors that destroy company property or equipment. *Political deviance* (Robinson & Bennett, 1995), *personal aggression* (Robinson & Bennett, 1995), and *unruliness* (Hunt, 1996) describe negative actions that harm coworkers. *Compliance* (Borman & Motowidlo, 1993; Brief & Motowidlo, 1986; Organ, 1988) and *useful personal behavior* (Borman & Brush, 1993) describe behaviors related to following rules and regulations. Like the citizenship domain, this area is becoming burdened with numerous definitions and conceptualizations of employee deviance. Building on Robinson and Greenberg's (1998) and Robinson and Bennett's (1995) definitions, we defined *counterproductive performance* as voluntary behavior that harms the well-being of the organization.

Hypotheses

These three broad performance components emerged from a review of the past 20 years of research on the structure of job performance. The first component, subsequently labeled *task performance*, includes actions and behaviors related to the production of a good or the provision of a service. The second component, subsequently labeled *citizenship performance*, describes actions and behaviors that further the goals of the organization by contributing to its social and psychological environment. The third component, labeled *counterproductive performance*, includes voluntary actions and behaviors that harm the well-being of the organization or its members. Although these three components describe the domain of job performance, it is still unclear how managers use information about each one when they form overall impressions of employees. In this study, we investigated the relative importance of the three performance components to supervisory ratings of overall job performance.

A limited amount of research provides guidance with regard to formulating hypotheses about the relative importance of task, citizenship, and counterproductive performance. The evidence that does exist is restricted to task and citizenship performance. Conway (1999) found support for the importance of contextual performance (job dedication and interpersonal facilitation) and task performance (technical performance and leadership) in ratings of overall performance. Technical performance, job dedication, and interpersonal facilitation had significant path coefficients to overall performance: .48, .31, and .21, respectively. Motowidlo and Van Scotter (1994) found that task performance explained 13% of the variance in overall performance above contextual performance, whereas contextual performance explained 11% of the variance above task performance. The results of this research indicate that task performance has a larger influence on ratings of overall performance than citizenship performance but that citizenship performance also explains rating variance.

> *Hypothesis 1*: Task performance will be given a significant weight in ratings of overall performance.
>
> *Hypothesis 2*: Citizenship performance will be given a significant weight in ratings of overall performance.
>
> *Hypothesis 3*: Task performance will be given a larger weight than citizenship performance.

Some researchers have raised the possibility that citizenship performance is valued more for effective task performers than for less effective task performers (Kiker & Motowidlo, 1999; Orr, Sackett, & Mercer, 1989; Werner, 1994). This interaction has been supported in the literature (Kiker & Motowidlo, 1999; Werner, 1994). Therefore, it is predicted that task performance moderates the relationship between citizenship performance and overall job performance; citizenship performance will have a stronger influence on ratings of overall performance when task performance is high compared with when it is low. One may also argue the opposite effect for counterproductive performance. That is, counterproductive performance may have a weaker influence on overall performance when task performance is high compared with when it is low. Or, the interaction between task and counterproductive performance may operate in yet another way such that the weight given to task performance decreases as counterproductive performance increases. A similar pattern could be argued for the relationship between counterproductive and citizenship performance. Although these interactions have not been examined in prior research and their nature and direction are unclear, we examined all possible interactions in an exploratory manner and specified a formal hypothesis only for the interaction between task and citizenship performance.

> *Hypothesis 4*: Task performance will moderate the relationship between citizenship performance and overall job performance; citizenship performance will receive a larger weight in ratings of overall performance when task performance is high compared with when it is low.

The wealth of literature developing in the area of integrity testing and employee deviance indicates that organizations are concerned about counterproductive behavior in the workplace because of the obvious negative consequences of these behaviors for the organization and its members (Crino, 1994; Hollinger & Clark, 1982; Robinson & Bennett, 1995). Therefore, it is predicted that information about employees' deviant behavior will influence managers' evaluation of their overall performance.

> *Hypothesis 5*: Counterproductive performance will be given a significant weight in ratings of overall performance.

The relative importance of task, citizenship, and counterproductive performance can also depend on the job or the organization's culture or strategic goals regarding job performance. Some will argue that organizations have their own culture or goals regarding the importance placed on task, citizenship, and counterproductive

performance. Some organizations may value employees who are high task performers and have little interest in the degree to which they engage in citizenship or counterproductive performance, whereas other organizations may place equal value on all three components. This logic argues for differences from one organization to another and suggests that raters may demonstrate policies that are consistent with their organization's culture or strategic goals. However, another argument could be made for job differences. Some researchers would argue that more weight might be given to counterproductive performance in retail settings, because there is a greater concern about theft and customer relations in these settings. One could make a similar argument for nursing because of the consequences of deviant behaviors in a hospital setting. However, these arguments for job differences conflict with the previous arguments made for organizational culture differences. For example, if job and organization differences do exist and if raters from the same job represent two different organizations with different strategic goals regarding performance, the raters could demonstrate different policies even though they represent the same job. Therefore, it is difficult to formally hypothesize job and organization differences, although these differences are very interesting. Although we did not specify hypotheses for job or organization differences, we examined these relationships in the data.

Policy Capturing

In this study, we used policy capturing to determine the relative importance of task, citizenship, and counterproductive performance. A typical policy-capturing study involves (a) presenting raters with a series of profiles in which the independent variables of interest are manipulated, (b) obtaining raters' judgments about the dependent variable, and (c) using multiple regression analyses to compute the relative importance of each independent variable. The end product is a statistical equation or "captured rating policy" for each rater that represents an expression of how the rater combines and weights the information contained in each profile to arrive at a decision or judgment.

Although there may be some concern about the construct validity of the captured policies obtained from experimental policy-capturing designs, it is an appropriate methodology for achieving the goal of this study; it provides a means to investigate the way in which raters use information about the different components of job performance to produce a judgment about overall job performance. An effort was also made to address some of the shortcomings of traditional experimental policy-capturing designs. First, the behaviors were scaled to ensure that they varied over equivalent standard deviation units in their respective populations in an attempt to avoid making an unfair comparison about the relative importance of the three performance components (Cooper & Richardson, 1986). Second, the number of hypothetical profiles included in the survey was limited to avoid problems of boredom and fatigue. However, this did not compromise the power of the study because the profile-to-cue ratio remained within the 10:1 ratio recommended by Nunnally (1978). Third, to increase the generalizability of our results and to address the common critique of policy-capturing studies that they are often conducted on small samples of raters in one job, we examined the relative importance of the three components in five different jobs across 15 different organizations.

Method

Sample

The study was designed to compare the rating strategy of managers in various jobs. Jobs were selected to vary in complexity, to represent more than one occupation or field, and to ensure a sufficiently large number of managers for the job in a given organizational setting. On the basis of these three criteria, five jobs were chosen: accountant, nurse, administrative assistant, retail cashier, and machine operator.

A total of 123 accounting managers, 271 senior administrators, 200 head cashiers or store supervisors, 75 production supervisors or managers, and 104 nursing managers in 15 different organizations across the United States and Canada were invited to participate in the study, for a potential sample of 773. The 15 organizations were identified from alumni records at a large midwestern university. A total of 520 participants agreed to complete the survey, yielding a response rate of 67%. However, 16 surveys were unusable (i.e., some pages were left blank or the rater used anchors outside of the 5-point Likert scale), resulting in a final sample of 504. Demographic information for the participants is summarized in Table 3 by job.

Profile Development

Three major steps were involved in the development of the hypothetical job performance profiles: (a) the compilation of a separate list of behaviors for each performance component, (b) the scaling of the behaviors to ensure that they were manipulated with equivalent strength, and (c) the creation of the profiles.

Table 3
Mean Sample Characteristics by Job

Variable	Accountant	Administrative assistant	Retail cashier	Machine operator	Nurse
Age (in years)	36.8 (8.1)	47.3 (8.2)	38.5 (8.4)	42.1 (7.6)	44.9 (7.0)
Work experience (in years)	14.1 (8.4)	24.4 (8.1)	15.9 (9.2)	21.7 (8.6)	21.2 (7.0)
% White	95.1	92.9	95.1	88.6	96.6
% male	60.7	8.3	40.5	77.3	1.1
% with a bachelor's degree	72.1	39.7	31.6	22.7	74.6
% rated performance[a]	98.4	92.3	86.1	95.5	95.5
Response rate (%)	52.0	59.0	80.0	61.0	88
n	61	155	158	44	86

Note. When applicable, the standard deviation is included in parentheses.
[a] Represents the percentage of respondents who had previous experience rating job performance.

Step 1. A separate list of behaviors was generated for each performance component. Regarding the task behaviors, each job was matched with its respective code in the 4th edition of the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1991). Behaviors were chosen from the job descriptions that fell within the working definition of task performance and were reviewed by subject matter experts for accuracy. These subject matter experts were individuals who were employed in the respective jobs. They reviewed the task statements to verify that they reflected what their jobs actually entailed. The following are some examples of task behaviors: prepares balance sheets (accountant), receives money and issues change to customers (retail cashier), and submits the health care plan of individual patients for periodic review (nurse).

For the citizenship and counterproductive performance components, a list of behaviors that fell within the working definition of each of these components was compiled from the literature. It was possible to use the same citizenship and counterproductive behaviors for each job because their working definitions were not restricted to behaviors that contributed to the production of a good or the provision of a service. Examples of these behaviors included attends functions that promote the organization (citizenship) and fights or argues with coworkers (counterproductive). This process produced a total of 8–9 behaviors for each performance component.

Step 2. The purpose of the second step was to scale the behaviors to ensure distributional equivalence (Cooper & Richardson, 1986). Distributional equivalence means that the levels of the three performance components vary over equivalent spreads in their respective populations. Each of the 8 behaviors from Step 1 were modified to reflect high, medium, and low levels of performance (e.g., correctly prepares balance sheets, never complains about the organization to coworkers), producing a total of 25 behaviors for each performance component.

A survey that consisted of 25 task, 25 citizenship, and 25 counterproductive performance items was created for each job. The surveys were identical with the exception of the task behaviors. The surveys were distributed to a sample of 85 business graduate students. This sample had similar backgrounds and experience to persons working in the field who rated job performance (e.g., $M = 13.5$ years of work experience [$SD = 8.5$ years] and $M = 35.6$ years of age [$SD = 8.5$ years]) and were employed in a number of different jobs (e.g., accounting manager, chief financial officer, administrative director, chief nursing officer, operations director, product manager, facilities engineer). Each student received a survey for one of the five jobs. In the survey, the student was provided with the working definition of each performance component and was asked to read each behavioral statement and rate the level of performance it reflected using a 7-point Likert scale (e.g., the Likert scale for task performance was anchored by 1 = *low task performance* and 7 = *high task performance*).

Item-level statistics were computed for each behavior. The item had to satisfy three criteria to be selected for use in the final survey. First, there had to be relatively high agreement among the sample as to the level of performance the item reflected (Landy, Rastegary, Thayer, & Colvin, 1991). A low standard deviation provides evidence of agreement. Therefore, any item with a standard deviation higher than 1.5 was eliminated. Second, the aggregate mean and variance for the final set of behaviors had to be similar across performance components. This step ensured that the performance components as a whole were manipulated with equivalent strength. Third, the performance components had to be approximately normally distributed. Items were eliminated until these three conditions were satisfied. This procedure was repeated for each job. The final set of behaviors included 16 items in each performance component. See Table 4 for descriptive information on the final set of behavioral items by job.

Step 3. The final step involved the creation of the hypothetical profiles. One behavior was randomly selected without replacement from each set of task, citizenship, and counterproductive behaviors, producing 16 unique profiles. This step was repeated to obtain another set of 16 profiles, yielding a total of 32 unique profiles satisfying the recommended profile-to-cue ratio (Nunnally, 1978). Although each behavior appeared twice in the final survey, no 2 profiles appeared twice (with the exception of 2 repeat profiles that were included to assess reliability), yielding a total of 34 profiles. Although we did not use a completely crossed design in which all possible combinations of cues and performance levels were reflected in the profiles, the procedure of randomly sampling (without replacement) one behavior from the list of task, citizenship, and counterproductive behaviors resulted in independent performance components. The correlations among the performance components ranged from .00 to .05. A few sample profiles are included below. However, it is important to note that the 34 profiles included a larger variety of task, citizenship, and counterproductive behaviors than are provided in the few examples.

> *Retail cashier*: Rob rarely ever makes errors when receiving money and issuing change to customers in payment for goods. He always attends functions that promote the well-being of the store. He never gossips about other cashiers.

> *Accountant*: Chris sometimes spreads false rumors about coworkers. He always makes constructive suggestions about how to improve the organization. He sometimes makes errors when preparing profit and loss statements.

> *Machine operator*: Michelle always demonstrates knowledge about other machines. She sometimes helps other coworkers check the set-up or inspect the finished product for flaws. She never blames other coworkers for her mistakes.

The performance components were randomly ordered within each profile to ensure that primacy or recency effects were not confounded with the importance weights. For example, if information about task performance was always presented first in the profiles and the results showed that raters

Table 4

Descriptive Statistics for the Pool of Task, Citizenship, and Counterproductive Performance Behaviors Included in the Final Survey

Job	Task performance		Citizenship performance		Counterproductive performance	
	M	*SD*	*M*	*SD*	*M*	*SD*
Accountant	3.64	1.78	3.69	1.86	3.57	1.96
Administrative assistant	3.51	1.97	3.78	2.09	3.39	2.01
Retail cashier	3.82	2.01	3.89	2.06	3.46	2.10
Machine operator	3.54	2.05	3.86	2.03	3.59	2.15
Nurse	3.61	2.11	3.86	2.21	3.67	2.18

Note. The items were measured on a 7-point scale (e.g., for task performance, 1 = *low task performance* and 7 = *high task performance*).

gave task performance the largest weight, one could argue that order was confounded with a characteristic of the profile.

Procedure

Managers who were responsible for rating the performance of employees in the given jobs were asked to participate in the study. The primary investigator attended monthly managerial meetings in which she introduced the study and verbally invited the managers to participate in the study and complete the survey during the meeting. When it was not possible for the researcher to attend the monthly meetings, each participating organization provided the researcher with a mailing list. The surveys were mailed to all managers in the targeted job at the given organization. A business-reply envelope addressed to the researcher was included with each survey. In both administrations, a cover letter, which explicitly stated that participation was voluntary and that all responses were anonymous, accompanied the survey.

Participants were asked to complete a survey, which required approximately 20–30 min to complete and consisted of two sections. In the first section, subjects were asked to read each of the 34 hypothetical profiles and to rate the overall job performance of the employee depicted within the profile using a 5-point Likert scale (1 = *low overall performance* and 5 = *high overall performance*). Information about the tenure and work experience of the employees was included in the introduction to Section 1 and was held constant across profiles and jobs (e.g., 3 years of work experience and tenure of 1 year). The second section of the survey requested demographic and background information (e.g., age, education, race, work experience, gender, occupational title, occupation, and prior experience rating job performance).

Analyses

In an effort to obtain as much information as possible about participants' decision strategies, we conducted multiple analyses. The first two sets focused on the policies of individual raters. The last set investigated whether raters could be clustered into homogeneous groups on the basis of their weights.

Captured rating policy. In the second set of analyses, a multiple regression equation was computed for each participant to assess the effects of the linear combination of the three performance components on ratings of overall performance. The overall performance ratings were regressed on the scaled values for the three performance components (see the *Profile Development* section). Because the performance components were orthogonal, the standardized regression coefficients can be interpreted as relative weights; the squared multiple correlation coefficient can be interpreted as a measure of the consistency with which the rater used the captured policy.

Hierarchical linear modeling. This study was a mixed experimental design in that both within-subjects (i.e., three performance components) and between-subjects (i.e., demographic information) factors were examined. The previous analyses focused on within-subjects factors. This set of analyses focused on the between-subjects factors to examine whether individual-differences characteristics (i.e., demographic and background variables) were related to raters' policies. Hierarchical linear modeling was used to assess interindividual differences. The Level 1 analyses consisted of the estimation of separate regression equations for each rater in which ratings of overall performance were regressed on the three performance components. The unstandardized intercept and slopes from these regressions served as the dependent variables for the second level of analyses (assuming that there was sufficient variation in the slopes and intercept). The independent variables in the second level of analyses included the background and demographic variables that were measured in the second section of the survey (age, gender, work experience, and education).

Hierarchical cluster analyses. To determine whether raters could be grouped into a smaller number of homogeneous clusters according to their relative weights, we conducted a fourth set of analyses in which a sequential agglomerative hierarchical cluster procedure was used. The raters' relative weights on the three performance components served as input into the procedure.

This study compared four different algorithms: complete linkage, average linkage between groups, average linkage within groups, and Ward's method. The fusion coefficient was used as the criterion for determining the most interpretable cluster solution (i.e., chose the number of clusters that produced a large increase in the fusion coefficient). Hierarchical cluster analysis was conducted separately for each job.

Pooled data. The data were also pooled across jobs to estimate a policy for the group of raters and to examine interaction effects. It was not possible to include interaction terms in the within-subject regressions because of low power. In an attempt to control for the effect of the individual rater, we created a dummy variable for each rater and included it in the regression analyses as a control variable.

Results

Two repeat profiles were included in the survey to assess reliability. Although it was not possible to compute a reliability coefficient for each rater, the performance scores from the repeat profiles were used to compute an overall reliability coefficient across raters and jobs. The scores were standardized before they were correlated to eliminate elevation differences between raters. This reliability coefficient was .75.

Captured Rating Policy

The purpose of these analyses was to estimate a regression equation for each rater. The results were summarized by job and are presented in Table 5 (a table that contains the 504 individual within-subjects regression equations can be obtained from Maria Rotundo). Table 5 reveals that there was wide variation in the relative weights on the three performance components. However, on average, information about all three performance components was taken into consideration when ratings of overall performance were made. Furthermore, on average, greater weight was given to information about task and counterproductive performance than to information about citizenship performance. The mean standardized regression coefficients and incremental multiple correlation squared provide evidence to support this pattern of weights. Raters in this study took the task seriously and were consistent, on average, when they rated the profiles. This pattern of findings generalized across jobs.

It could be argued that raters' policies are influenced by their organization's culture. For example, the 61 accounting managers in this study represented three different organizations. In an effort to rule out the possibility that the accounting managers' weights covaried with the organization that they represented, we conducted separate one-way analyses of variance within each job. There were no significant differences between organizations on the weights given to task, citizenship, and counterproductive performance.

Pooled Data

In an attempt to examine whether raters combined information about task, citizenship, and counterproductive performance in a multiplicative manner, we pooled the data across raters and jobs and estimated a regression model including the four interaction terms. The performance components were centered on their means

Table 5
Mean Standardized Regression Coefficients and Multiple Correlations Squared by Job

Variable	Accountant	Administrative assistant	Retail cashier	Machine operator	Nurse	Overall
Task performance						
M	.54	.55	.44	.53	.49	.50
SD	.08	.12	.11	.10	.12	.12
Range	.37–.72	.20–.79	.02–.65	.28–.74	.31–.65	.02–.79
% significant coefficients	97	98	98	95	98	
Mean % variance explained	30	32	20	29	25	
Citizenship performance						
M	.29	.20	.30	.43	.40	.29
SD	.08	.09	.10	.06	.14	.13
Range	.11–.50	.01–.46	.01–.51	.28–.56	.17–.61	.01–.61
% significant coefficients	85	40	78	93	93	
Mean % variance explained	10	4	12	20	17	
Counterproductive performance						
M	−.48	−.55	−.55	−.47	−.50	−.52
SD	.11	.13	.09	.10	.13	.11
Range	−.25–−.65	−.21–−.80	−.31–−.79	−.27–−.64	−.27–−.77	−.21–−.79
% significant coefficients	95	96	99	95	98	
Mean % variance explained	24	32	31	23	26	
Overall job performance[a]						
M	2.36	2.37	2.25	2.29	2.22	2.30
SD	1.04	1.08	1.13	1.01	1.10	1.08
R^2						
M	.63	.67	.63	.72	.66	.66
SD	.09	.09	.09	.08	.10	.10
Range	.46–.80	.40–.87	.34–.82	.47–.87	.50–.86	.34–.87
n	61	155	158	44	86	504

[a] Represents the mean rating of overall job performance across raters.

before the interaction terms were computed (Cohen & Cohen, 1983). Rater was included as a control variable in the pooled hierarchical regression. The results are presented in Table 6. All interactions were significant except the three-way interaction. These interactions were plotted and are illustrated in Figures 1A–C. Figure 1A shows the task by citizenship performance interaction. A one-unit increase in citizenship performance produced a larger increase in overall performance when task performance was high compared with when it was low. This pattern is

Table 6
Standardized Regression Coefficients Including Interaction Terms When the Data Were Pooled Across Jobs

Variable	β	ΔR^2
Rater	−.041–.025	.08**
Task	.473**	.25**
Citizenship	.262**	.08**
Counterproductive	−.474**	.23**
Interactions		.03**
Task × Citizenship	.071**	
Task × Counterproductive	−.159**	
Citizenship × Counterproductive	−.043**	
Three-way interaction	−.007	
Total R^2		.67

Note. The individual effect of each rater was removed by including rater as a control variable in the pooled hierarchical regression. The range of values on the 504 regression coefficients for rater are presented in the first row. $N = 16,128$ (504 raters × 32 profiles).
** $p < .01$.

consistent with a reinforcement or synergistic interaction effect. The opposite pattern occurred in the task by counterproductive performance interaction (see Figure 1B). A one-unit increase in counterproductive performance produced a larger decrease in overall performance when task performance was high compared with when it was low. Regarding the citizenship by counterproductive performance interaction (Figure 1C), a one-unit increase in counterproductive performance produced a larger decrease in overall performance when citizenship performance was high compared with when it was low. Although the interactions reached statistical significance, the plots indicate that they were not practically significant.

Hierarchical Linear Modeling

The purpose of these analyses was to determine whether four demographic variables were related to raters' relative weights. The HLM 5 hierarchical linear modeling (Raudenbush, Bryk, & Congdon, 2001) program was used to estimate two models: the first modeling the relationships between overall performance and task, citizenship, and counterproductive performance at the individual level and the second modeling how the relationships at the individual level varied as a function of age, gender, work experience, and education. Hierarchical linear modeling was conducted separately within each of the five jobs (see Table 7). There was significant variation in the intercepts and slopes to justify the second level of analyses in all jobs and for most of the performance components, except where indicated in Table 7. There were only seven significant relationships. These are summarized below.
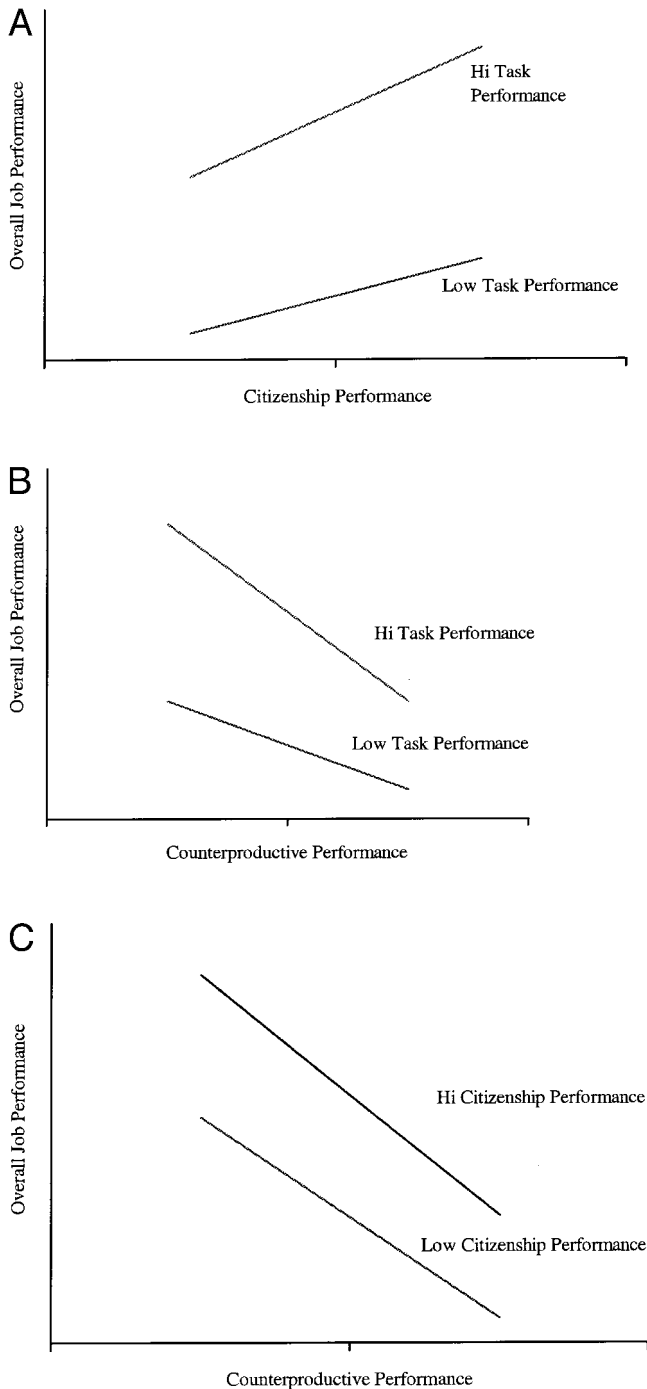
Figure 1. (A) Two-way interaction between task and citizenship performance predicting ratings of overall job performance when the data were pooled across jobs. (B) Two-way interaction between task and counterproductive performance predicting ratings of overall job performance when the data were pooled across jobs. (C) Two-way interaction between citizenship and counterproductive performance predicting ratings of overall job performance when the data were pooled across jobs. Hi = High.

Gender significantly predicted weights on counterproductive performance (for accounting managers, $\gamma = -.043, p < .05$) and task performance (for retail managers, $\gamma = .029, p < .05$). Female accounting managers placed greater weight on counterproductive performance than did male accounting managers. Male retail managers placed greater weight on task performance than did female retail managers. The coefficients on age were significant for task ($\gamma = .002, p < .05$) and counterproductive ($\gamma = -.003, p < .05$) performance in the retail sample. Older retail managers placed greater weight on task performance and less weight on counterproductive performance than did younger retail managers. Older senior administrators ($\gamma = -.006, p < .01$) and nursing managers ($\gamma = -.013, p < .01$) gave significantly lower overall performance ratings compared with their younger counterparts (significant Level 2 intercepts). Work experience significantly predicted weight on counterproductive performance for senior administrators ($\gamma = .003, p < .01$). Senior administrators with more work experience gave higher weight to counterproductive performance than did similar persons with less work experience. Furthermore, only 7 of a possible 64 relationships reached a conventional level of significance. In general, these analyses did not reveal a consistent pattern of relationships across demographic variables or jobs.

Hierarchical Cluster Analysis

The purpose of this set of analyses was to determine whether raters could be grouped together on the basis of the relative weights obtained from the within-subjects regression equations. Although four different algorithms were used to determine which clusters to merge at each iteration, all four produced similar cluster solutions. Therefore, the results for the average linkage between-groups method are presented because validation research has found that this method produces superior cluster recovery (Blashfield, 1976; Kuiper & Fisher, 1975) and satisfies algorithmic properties recommended by Kaufman and Rousseeuw (1990).

The largest increase in the fusion coefficient occurred at a two-cluster solution. Each rater was classified into one of the two clusters identified by the hierarchical clustering procedure. To interpret the meaning of the clusters, the standardized regression coefficients were averaged within each cluster. The results are presented in Table 8.

There were obvious differences between the clusters in terms of the relative importance of the three performance components. One cluster included raters who gave the most weight to task performance and less to citizenship and counterproductive performance. A second cluster included raters who gave almost equal weight to task and counterproductive performance and less weight to citizenship behaviors. A third cluster included raters who gave the most weight to counterproductive performance followed by task and citizenship performance. These findings indicate that raters' relative weights could be grouped into three homogeneous clusters.

Independent-samples *t* tests were run to test whether the mean standardized regression coefficients for each performance component in Cluster 1 were significantly different from those in Cluster 2. The mean regression weight for task performance in Cluster 1 was significantly larger ($p < .01$) than the mean regression weight for task performance in Cluster 2 in all of the jobs: for accountant, $t(59) = 6.16$; for administrative, $t(153) = 9.81$; for

Table 7
Summary of Hierarchical Linear Models Predicting Level 1 Intercepts and Slopes From Demographic and Background Variables in the Field Sample

Variable	Level 1 intercept	Level 1 slope		
		Task	Citizenship	Counterproductive
Accountant				
Gender	−.005			−.043*
Age	−.012			−.000
Work experience	−.006			.003
Bachelor's degree	−.003			−.000
Administrative assistant				
Age	−.006**	.001	−.001	−.001
Work experience	−.001	−.001	.001	.003**
Some college	.005	−.001	.004	−.002
Bachelor's degree	−.003	.001	.001	.002
Retail cashier				
Gender	.007	.029*	−.001	−.010
Age	−.004	.002*	−.001	−.003*
Work experience	.000	−.002	.001	.001
Bachelor's degree	−.010	.000	−.004	.003
Machine operator				
Gender	−.045	.004		.017
Age	−.016	−.001		.003
Work experience	.003	.001		.000
Bachelor's degree	−.050	.001		.001
Nurse				
Age	−.013**	.002	−.001	
Work experience	.008	−.002	.000	
Bachelor's degree	−.007	−.001	−.001	
Master's degree	−.008	.002	−.001	

Note. The first row represents the dependent variables in the hierarchical linear models. The first column represents the independent variables. The values in the remaining columns and rows represent the coefficients from the analyses. There was significant Level 1 variance to justify Level 2 analyses, except where indicated by empty cells.
* $p < .05$.   ** $p < .01$.

retail cashier, $t(156) = 15.10$; for machine operator, $t(42) = 6.24$; and for nurse, $t(84) = 12.17$. There was no significant difference between the clusters on the mean regression weights for citizenship performance, except for the administrative assistant job, $t(153) = -5.36$, $p < .01$. The mean regression weight for counterproductive performance in Cluster 2 was significantly larger ($p < .01$) than the mean regression weight for counterproductive performance in Cluster 1 in all of the jobs: for accountant, $t(59) = 8.63$; for administrative assistant, $t(153) = 13.99$; for retail

cashier, $t(156) = 9.82$; for machine operator, $t(42) = 9.18$; and for nurse, $t(84) = 9.29$.

In an effort to determine whether the clusters were defined primarily by raters from one organization, we used chi-square analyses. For example, a finding that Cluster 1 included significantly more raters from Organization A than from Organization B would suggest that Cluster 1 was defined by raters from Organization A. Within-job chi-square analyses revealed that there were no significant differences between the clusters in terms of the

Table 8
Results of Hierarchical Cluster Analysis by Job and When the Data Were Pooled Across Jobs

Job	Cluster 1				Cluster 2				Cluster 3			
	Task	Citizen	Counter	n	Task	Citizen	Counter	n	Task	Citizen	Counter	n
Accountant	.63	.26	−.34	16	.51	.30	−.52	45				
Administrative assistant	.67	.12	−.40	43	.50	.22	−.61	112				
Retail cashier	.51	.29	−.51	99	.33	.31	−.62	59				
Machine operator	.60	.44	−.40	24	.46	.41	−.56	20				
Nurse	.56	.41	−.43	39	.41	.39	−.58	47				
Pooled data	.67	.19	−.36	54	.30	.22	−.68	33	.50	.31	−.53	417

Note. Each row represents the average weight on task, citizenship (citizen), and counterproductive (counter) performance for the group of raters within each cluster. The last row represents the average weight for each cluster when the data were pooled across jobs.

Table 9
Percentage of Raters Within Each Cluster Pattern by Job and When the Data Were Pooled Across Jobs

Pattern	Accountant	Administrative assistant	Retail cashier	Machine operator	Nurse	Pooled cluster
Task performance dominated	26	28		55	45	11
Counterproductive performance dominated		72	37	45	55	6
Equal weights to task and counterproductive performance	74		63			83

organizations represented therein. Furthermore, there were no significant differences between the two clusters on the demographic and background variables, with the exception of the administrative assistant job, for which there were significantly more raters with a bachelor's degree in Cluster 1 than in Cluster 2, $\chi^2(2, N = 154) = 8.88$, $p < .05$.

Summary

There are three important observations to note from the results of the cluster analyses. First, raters could be grouped into three homogeneous clusters on the basis of their weights. These results are summarized in Table 9. The results did not support a one-cluster solution in which the relative weights were the same for all raters. Second, the raters' policies did not cluster according to the organizations they represented. Third, raters in the clusters did not differ significantly in terms of various demographic and background variables.

Discussion

A review and integration of key theoretical and empirical research surrounding the criterion revealed that job performance can be described by three broad groups of behaviors. The goal of this study was to provide empirical evidence to establish the relative importance of these three components to supervisory ratings of overall job performance. The results revealed that although all three components influence ratings of overall job performance, raters demonstrate unique implicit rating policies that can be grouped into three distinct clusters.

Hypotheses 1, 2, 3, and 5 proposed that raters would place significant weight on information about an employee's task, citizenship, and counterproductive performance when they rated overall performance and that task performance would receive a larger weight than citizenship performance. These hypotheses were supported. On average, raters gave the most weight to task and counterproductive performance and less but significant weight to the citizenship component, as demonstrated in Table 5. In fact, task and counterproductive performance together explained anywhere from 51% to 64% of the variance in the overall performance ratings, whereas citizenship performance explained between 4% and 20% of the variation. The finding that task performance explained a larger proportion of the variation in overall performance ratings relative to citizenship performance has been reported in previous studies (Conway, 1999; Motowidlo & Van Scotter, 1994). Although job performance has traditionally been described as the quantity and quality of work, over the past 30 years researchers have drawn attention to other facets of performance. The findings from this study lend support to a conceptu-

alization of job performance that goes beyond task elements to one that includes citizenship and counterproductive behaviors.

The results of the within-subjects regression analyses presented in Table 5 indicate that managers did not share a common policy for combining information about task, citizenship, and counterproductive performance; there was considerable variation in the relative weights. Subsequent analyses revealed that raters' policies could be grouped into three distinct clusters. One cluster included raters for whom information about task performance carried the most weight. This cluster emerged in all of the jobs except for retail cashiers. A group of managers felt that the successful completion of these actions and behaviors on the job was most important and attended to information about these behaviors when they formed overall impressions about employees. A second cluster included raters who gave the most weight to information about employees' counterproductive performance. These raters believed more strongly in the importance of acting with integrity and conducting oneself in a professional manner than in contributing to the bottom line or being an organizational citizen. The third cluster included raters who placed the most weight on both task and counterproductive performance and less weight on citizenship performance. Raters who fell in this cluster believed in the importance of completing tasks and maintaining personal and professional integrity. Together, the results from the within-subjects regression analyses and the cluster analyses indicate that raters did not share a common policy for rating employees. Knowledge that raters' policies fall into one of three general patterns provides a basis for speculating about why raters have different policies or which individual-differences factors are related to their policies. This study examined whether demographic and background information were related to the three clusters and found no significant pattern of relationships. However, personality characteristics or leadership styles may predict raters' implicit policies. For example, one may speculate that individuals who believe strongly in getting the job done, have a bottom-line focus, or are high in need for achievement are more likely to give larger weight to task performance. These raters may also possess a leadership style that is consistent with initiating structure as opposed to consideration for others (Fleishman, 1953). In contrast, raters who value low counterproductive behavior may be described as being high in integrity. Future research can investigate whether various personality characteristics account for the differences in raters' policies.

Hypothesis 4 proposed an interaction between task and citizenship performance such that citizenship performance would be valued more for high task performers than low task performers. This hypothesis was partially supported. We pooled the data across raters to examine this relationship, producing a total of 16,128 data points. The interaction reached statistical significance. However,

with such a large sample size, most relationships will reach a conventional level of significance. Although significance was found, the plots indicate that a one standard deviation increase in task performance resulted in a marginal increase in the weight given to citizenship performance. This finding does not provide strong support for the practical significance of the interactions. We tested all other two-way and three-way interactions and found a similar pattern. A reviewer raised the possibility of examining the interactions within each rater. However, a proper test of within-rater interactions would require approximately 70 data points or profiles for each rater, and our survey included a total of 32 unique profiles.

Another relationship in which we were interested was whether a different pattern of policies emerged in the five jobs or the 15 organizations. The raters in this study were accounting, nurse, retail, production, or administrative managers. One possibility is that the job the rater represents influences his or her implicit policy about the relative importance of task, citizenship, and counterproductive performance. One could speculate that accounting managers pay more attention to task behaviors, whereas retail and nurse managers place more emphasis on counterproductive behaviors because of the negative effects of these behaviors in a retail or hospital setting. Although we examined whether there were significant differences in the raters' implicit policies by job, the differences were not significant. Therefore, the pattern of relative weights did not covary with job. Still another possibility is that raters' policies are influenced by the organization's culture or strategic goals regarding the degree to which task, citizenship, and counterproductive performance are valued. We tested whether raters from each organization demonstrated a significantly different pattern of weights than raters from the other organizations and found no support for differences between organizations. The finding that the data in the present study do not support general assertions that the relative importance of the three performance components depends on the job or the organization does not mean that these differences do not exist. These differences may not have emerged in our data because raters' implicit policies were so strong that they dominated or superseded any job or organization differences. It would be interesting to investigate whether any job or organization policies emerge after raters have been trained to rate according to the organization's predetermined policy.

Implications and Future Directions

The finding that raters in this study did not share a common policy for weighting task, citizenship, and counterproductive performance has implications for understanding and estimating the reliability of supervisory ratings. The most common way to measure job performance is a supervisor's or manager's rating of an employee's overall performance (Cascio, 1991; Cleveland, Murphy, & Williams, 1989). The reliability of these ratings has traditionally been estimated by correlating the scores from two separate raters who assess the same group of employees. Research consistently shows that when the reliability of supervisory ratings is computed using this approach, raters do not always agree about the performance of individuals (Rothstein, 1990). Viswesvaran, Ones, and Schmidt (1996) conducted a meta-analysis of the reliability of job performance ratings and reported a mean corrected interrater reliability of .52 for overall performance. In a recent debate about

the appropriate method for estimating the reliability of supervisory ratings, Murphy and De Shon (2000) questioned these traditional methods and argued that interrater correlations violate assumptions underlying classical true score theory and should not be used to estimate the reliability of supervisory ratings. To use interrater correlations as reliability estimates, raters must be considered parallel forms of the same rating instrument, and disagreement between raters should be due to random measurement error. However, Murphy and De Shon argued that systematic rater effects provide one explanation for why raters might disagree in their ratings of overall performance and why raters should not be viewed as parallel tests. These rater effects include systematic differences in (a) what raters observe, (b) access to information other than observations of performance, (c) expertise in interpreting what is observed, and (d) evaluating what is observed. The finding in the present study that raters did not share a common policy for the relative importance of task, citizenship, and counterproductive performance is an example of a systematic difference in terms of how raters evaluate what is observed and may explain why raters generally demonstrate low agreement when rating the overall performance of employees. Furthermore, Murphy and De Shon argued that although these systematic rater effects contribute to measurement error, it is not random measurement error. Therefore, two of the three assumptions underlying classical true score theory are violated if interrater correlations are used to estimate the reliability of supervisory ratings. In the present study, the mean interrater reliabilities were higher than the value reported in Viswesvaran et al.'s study: .77 for accountant, .72 for administrative assistant, .71 for retail cashier, .79 for machine operator, and .71 for nurse. However, a number of features in the present study are atypical of traditional performance appraisal settings. For example, the raters in this study read the same descriptions about employees' performance, potentially decreasing systematic differences in what was observed, access to information other than observations of performance, and expertise in interpreting what was observed. These features may explain why the interrater reliabilities in the present study that were computed using the traditional approach are larger than the value reported by Viswesvaran et al. However, the interrater reliabilities are still far from perfect. Thus, the fact that raters demonstrate their own implicit rating policy about the relative importance of task, citizenship, and counterproductive performance provides one example of a systematic rater effect that explains why raters do not always agree about the performance of employees.

The finding of variation in raters' weights also raises concerns about the use of overall performance ratings in test validation research. Imagine a situation in which a test that predicts task performance is validated against supervisor ratings of overall job performance. Suppose that an applicant scores high on the test, is selected for the job, performs poorly on task performance, and engages in no counterproductive performance. Suppose further that his or her supervisor holds a policy that weighs counterproductive performance high and task performance low. These circumstances produce a high rating of overall performance and give the appearance that the test predicts well when in reality it does not. The finding that raters vary in terms of their implicit policies of the relative importance of task, citizenship, and counterproductive performance lends support to earlier discussions (Dunnette, 1963; Guion, 1961; Schmidt & Kaplan, 1971) and recent calls

(Hough & Oswald, 2000; Murphy & Shiarella, 1997) to validate tests against multiple criteria as opposed to ratings of overall performance. Murphy and Shiarella found that the validity of a selection test battery varied as a function of the relative weights given to both the predictors and the criteria. In the example above, if a supervisor was asked to rate the employee's task, citizenship, and counterproductive performance and the test was validated using the task performance rating as the criterion, a different picture of the validity of the test would emerge (assuming no halo bias on the part of the rater).

The results of this research also have practical implications for employees and organizations. One of the most important findings of this study is that experienced managers or supervisors demonstrated reliable differences in their pattern of weights. Therefore, global ratings of performance rely to a great extent on raters' implicit policies of what really matters. Two employees may engage in the same behaviors on the job yet receive different ratings of performance depending on who their rater is. Therefore, it is important for employees to become aware of their raters' implicit policies, especially if organizational rewards are determined on the basis of global ratings of performance. The finding that raters demonstrate different policies may reinforce the need for organizations to communicate to managers what behaviors are important or valued on the job or to make explicit the organization's rating scheme. These efforts are extremely important if organizations want to ensure that the desired behaviors are reinforced and that the productive employees are rewarded.

It is important to keep in mind that the pattern of results obtained from the within-subjects regressions applies to situations in which there is equal variance on the three performance components. This constraint was intentionally built into the profiles to ensure that the criterion of distributional equivalence was satisfied so as to be able to make fair comparisons about the relative importance of the three performance components. However, in organizational settings, there may be less variance in counterproductive performance in the workplace. Although employees engage in various forms of deviant behavior, the more extreme forms may be less prevalent (Hollinger et al., 1992; Lehman, Holcom, & Simpson, 1990; McGurn, 1988). In an effort to depict a situation in which the base rate of occurrence of counterproductive behaviors is low, we report here an additional set of analyses in which a series of weights were applied to the hypothetical profiles and the analyses were repeated on the pooled data using weighted least squares regression. Profiles that included more extreme examples of counterproductive behaviors (e.g., at least one standard deviation above the mean) were given a weight of 1, whereas a range of weights (5, 10, 50, and 100) for all other profiles was examined. As the weight given to the less extreme forms of counterproductive behavior increased (i.e., from 5 to 100), the standardized regression coefficients for task performance increased, whereas the coefficients for citizenship and counterproductive performance decreased (see Table 10). The results from the weighted regression analyses indicate that in situations in which there is less variance on counterproductive performance relative to the other performance components, the weight given to this performance component decreases.

Table 10
Standardized Regression Coefficients From Weighted Least Squares Regression When the Data Were Pooled

Weight[a]	Task performance	Citizenship performance	Counterproductive performance
1	.478**	.279**	−.495**
5	.548**	.277**	−.375**
10	.566**	.274**	−.335**
50	.585**	.268**	−.291**
100	.588**	.267**	−.284**

[a] Represents the weight applied to profiles in which less extreme forms of counterproductive behavior were depicted.
** $p < .01$.

Study Strengths and Limitations

An experimental policy-capturing design was chosen as the appropriate method for this study. An attempt was made to eliminate as many as possible of the threats to external validity that plague a typical experimental policy-capturing design (steps are detailed in the Method section). However, it is important to recognize a limitation of this study. A decision was made to specify independent performance components. Although there is some evidence to suggest that the three performance components are intercorrelated (Campbell et al., 1990; Conway, 1996, 1999; Motowidlo & Van Scotter, 1994; Van Dyne & LePine, 1998), a number of these researchers have attributed these higher values to common halo shared by raters. Therefore, we specified independent components. However, this is a limitation only if true correlations between the performance components are meaningfully different from zero.

Theoretical Implications and Conclusion

A review of the job performance literature indicates that job performance can be conceptualized as task, citizenship, and counterproductive performance. The main objective of this study was to provide insight into how raters use information about these three performance components when they rate overall performance. The results from this study indicate that all three components matter but to different degrees. Although task and counterproductive performance dominate, the extent to which they do depends on who is providing the rating. Furthermore, the analyses conducted within each of the five jobs yielded a similar pattern of findings. Therefore, the relative importance of the three performance components does not vary by job per se but by rater. The present study contributes to the theory of job performance by providing empirical evidence that establishes the relative importance of task, citizenship, and counterproductive performance.

References

Astin, A. (1964). Criterion-centered research. *Educational and Psychological Measurement, 24,* 807–822.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin, 83,* 377–388.

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

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review, 11,* 710–725.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43,* 313–333.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74,* 130–135.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Conway, J. M. (1996). Additional evidence for the task–contextual performance distinction. *Human Performance, 9,* 309–330.

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84,* 3–13.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71,* 179–184.

Crino, M. D. (1994). Employee sabotage: A random or preventable phenomenon? *Journal of Managerial Issues, 6,* 311–330.

Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology, 47,* 251–254.

Fleishman, E. A. (1953). The measurement of leadership attitudes in industry. *Journal of Applied Psychology, 37,* 153–158.

George, J. M., & Brief, A. P. (1992). Feeling good–doing good: A conceptual analysis of the mood at work–organizational spontaneity relationship. *Psychological Bulletin, 112,* 310–329.

Giacalone, R. A., & Greenberg, J. (1997). *Antisocial behavior in organizations.* Thousand Oaks, CA: Sage.

Guion, R. M. (1961). Criterion measurement and personnel judgments. *Personnel Psychology, 14,* 141–149.

Hollinger, R. C., & Clark, J. P. (1982). Formal and informal social controls of employee deviance. *Sociological Quarterly, 23,* 333–343.

Hollinger, R. C., & Clark, J. P. (1983). *Theft by employees.* Lexington, MA: Heath.

Hollinger, R. C., Slora, K. B., & Terris, W. (1992). Deviance in the fast-food restaurant: Correlates of employee theft, altruism, and counterproductivity. *Deviant Behavior: An Interdisciplinary Journal, 13,* 155–184.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future—remembering the past. *Annual Review of Psychology, 51,* 631–664.

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49,* 51–83.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10,* 93–98.

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations.* New York: Wiley.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Kiker, D. S., & Motowidlo, S. J. (1999). Main and interaction effects of task and contextual performance on supervisory reward decisions. *Journal of Applied Psychology, 84,* 602–609.

Konovsky, M. A., & Organ, D. W. (1995). Dispositional and contextual determinants of organizational citizenship behavior. *Journal of Organizational Behavior, 17,* 253–266.

Kuiper, F. K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics, 31,* 777–783.

Landy, F. J, Rastegary, H., Thayer, J., & Colvin, C. (1991). Time urgency: The construct and its measurement. *Journal of Applied Psychology, 76,* 644–657.

Lehman, W. E. K., Holcom, M. L., & Simpson, D. D. (1990). *Employee health and performance in the workplace: A survey of municipal employees of a large southwest city.* Fort Worth: Texas Christian University, Institute of Behavioral Research.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance. *Organizational Behavior and Human Decision Processes, 50,* 123–150.

McGurn, J. (1988, March 7). Spotting the thieves who work among us. *Wall Street Journal,* p. A16.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79,* 475–480.

Murphy, K. R. (1989). Dimensions of job performance. In R. Dillon & J. Pellingrino (Eds.), *Testing: Applied and theoretical perspectives* (pp. 218–247). New York: Praeger.

Murphy, K. R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873–900.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50,* 823–854.

Nagle, B. F. (1953). Criterion development. *Personnel Psychology, 6,* 271–289.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Lexington Books.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10,* 85–97.

Organ, D. W., & Lingl, A. (1995). Personality, satisfaction, and organizational citizenship behavior. *Journal of Social Psychology, 135,* 339–350.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48,* 775–802.

Orr, J. M., Sackett, P. R., & Mercer, M. (1989). The role of prescribed and nonprescribed behaviors in estimating the dollar value of performance. *Journal of Applied Psychology, 74,* 34–40.

Patterson, C. H. (1946). On the problem of the criterion in prediction studies. *Journal of Consulting Psychology, 10,* 277–280.

Raelin, J. A. (1994). Three scales of professional deviance within organizations. *Journal of Organizational Behavior, 15,* 483–501.

Raudenbush, S., Bryk, T., & Congdon, R. (2001). *HLM 5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38,* 555–572.

Robinson, S. L., & Greenberg, J. (1998). Employees behaving badly: Dimensions, determinants, and dilemmas in the study of workplace deviance. In C. L. Cooper & D. M. Rousseau (Eds.), *Trends in organizational behavior* (Vol. 5, pp. 1–30). Chichester, England: Wiley.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 497–515). San Francisco: Jossey-Bass.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419–434.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68,* 653–663.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago: Rand McNally.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

U.S. Department of Labor. (1991). *Dictionary of occupational titles* (4th ed.). Washington, DC: Author.

Van Dyne, L., Cummings, L. L., & Parks, J. M. (1995). Extra-role behaviors: In pursuit of construct and definitional clarity (a bridge over muddied waters). In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 17, pp. 215–285). Greenwich, CT: JAI Press.

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal, 41,* 108–119.

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

Welbourne, T. M., Johnson, D. E., & Erez, A. (1998). The role-based performance scale: Validity analysis of a theory-based measure. *Academy of Management Journal, 41,* 540–555.

Werner, J. M. (1994). Dimensions that make a difference: Examining the impact of in-role and extra-role behaviors on supervisory ratings. *Journal of Applied Psychology, 79,* 98–107.