A FIELD STUDY OF THE EFFECTS OF RATING PURPOSE ON THE QUALITY OF MULTISOURCE RATINGS

GARY J. GREGURAS
Department of Psychology
Louisiana State University

CHET ROBIE
Department of Business
Wilfrid Laurier University

DEIDRA J. SCHLEICHER
Department of Psychology
University of Tulsa

MAYNARD GOFF III
Personnel Decisions International

Using a field sample of peers and subordinates, the current study employed generalizability theory to estimate sources of systematic variability associated with both developmental and administrative ratings (variance due to items, raters, etc.) and then used these values to estimate the dependability (i.e., reliability) of the performance ratings under various conditions. Results indicated that the combined rater and rater-by-ratee interaction effect and the residual effect were substantially larger than the person effect (i.e., object of measurement) for both rater sources across both purpose conditions. For subordinates, the person effect accounted for a significantly greater percentage of total variance in developmental ratings than in administrative ratings; however, no differences were observed for peer ratings as a function of rating purpose. These results suggest that subordinate ratings are of significantly better quality when made for developmental than for administrative purposes, but the same is not true for peer ratings.

The use of organizational interventions targeted at evaluating and developing managerial performance has surged over the past several years and continues to grow in popularity (Church & Bracken, 1997). One such intervention is multisource rating systems in which performance ratings are collected from multiple relevant viewpoints (e.g., peers, subordinates). These performance ratings may be used for a variety of different purposes including: between-person (e.g., administrative decisions such as salary administration), within-person (e.g., identifying individual strengths and weaknesses for developmental purposes),
or system-maintenance uses (e.g., evaluating personnel systems; Murphy & Cleveland, 1995).

Originally, multisource feedback systems were used almost exclusively for developmental purposes (London & Smither, 1995). However, the emerging trend is for organizations to also use these systems for administrative purposes (Dalessio, 1998; London, 2001). For example, of the 756 companies responding to a survey by Bohl (1996), 22% reported having multisource rating systems and, of these, more than 90% of the organizations reported using their systems to make administrative decisions. Likewise, Timmreck and Bracken (1995) reported that 28% of their respondents used multisource ratings as input into performance appraisal decisions, and 45% were moving toward using it as a part of the performance appraisal process, although in followups some organizations had decided not to move in that direction (Timmreck & Bracken, 1996, 1997).

The trend toward using multisource rating systems for administrative purposes appears to be based on (a) the assumption that such systems provide more complete and better quality information than that gathered from a single rater source (Dalessio, 1998; London, 2001), and (b) a desire to get a greater return on the investment involved in the development and implementation of these rating systems (Ghorpade, 2000; Waldman, Atwater, & Antonioni, 1998). Somewhat ironically, however, if such ratings used for administrative purposes were shown to be lower in quality than those for developmental purposes, organizations may actually be obtaining less (rather than more) return on their investments because of the resultant poorer quality administrative decisions. Unfortunately, there has been little research on how rating purpose actually affects the quality of multisource ratings (London & Smither, 1995; Sulsky & Keown, 1998). Accordingly, the current paper assesses the effects of rating purpose on the quality of peer and subordinate ratings. Using generalizability theory (GT), the current study estimates multiple sources of systematic variability associated with the ratings (variance due to items, raters, etc.) and then uses these values to estimate the dependability (i.e., reliability) of the performance ratings.

This study contributes to the existing literature investigating multisource ratings and rating purpose in several ways. First, this study investigates the effects of rating purpose on rating quality in a field setting with an employee sample rather than in the more typical laboratory setting with a student sample. Second, this study investigates the effects of rating purpose on peer and subordinate ratings. The majority of existing studies examining the effects of rating purpose have focused on student or supervisor ratings; this is unfortunate given the increased use of peers and subordinates as raters (e.g., Bernardin, Dahmus, &
Redmon, 1993). Third, GT is used to simultaneously estimate multiple sources of error variance of peer and subordinate ratings, whereas previous studies have primarily used classical test theory to investigate the effects of rating purpose on rating quality. As Murphy and DeShon (2000b) note, GT is the appropriate approach for estimating the dependability of performance ratings, yet this approach has largely been neglected by organizational researchers. The current study is in line with recent recommendations made by several authors to apply GT to individual and organizational phenomena (e.g., Gerhart, Wright, McMah- han, & Snell, 2000; Murphy & DeShon, 2000a, 2000b). Finally, the observed variance estimates obtained from the generalizability theory analyses are then used as input to project dependability estimates under multiple measurement conditions (i.e., differing numbers of raters, differing numbers of items, different rating purposes). These projected estimates may guide practitioners when designing and implementing multisource rating systems for different rating purposes.

Theoretical Influence of Rating Purpose

There are a number of reasons why rating purpose may influence the quality of observed ratings. First, rating purpose likely impacts a rater’s motivation (e.g., political agendas, Longenecker, Gioia, & Sims, 1987; personal beliefs, Jawahar & Williams, 1997; rating goals, Murphy & Cleveland, 1995). For example, raters may bias administrative ratings to (a) avoid the potential ramifications of negative evaluations (Bernardin, 1981; Fisher, 1989; Longenecker et al., 1987); (b) secure desirable outcomes for themselves or others (Jawahar & Williams, 1997); or (c) motivate poor performers (Murphy & Cleveland, 1995). In contrast, for developmental purposes, raters may be more motivated to help employees accurately identify strengths and weaknesses.

Second, different rating purposes may conflict with the rater’s typical role in the organization (Cleveland, Murphy, & Williams, 1989; Dorfman, Stephan, & Loveland, 1986; Stephan & Dorfman, 1989). Although peers and subordinates may consider providing feedback to other employees a part of their jobs, in contrast, they may feel uncomfortable impacting administrative decisions that typically are made by supervisors. In addition, having subordinates or peers provide information that influences administrative decisions violates the typical organizational hierarchy and, as Murphy and Cleveland (1995) note, such violations may represent a serious problem for the effectiveness of such rating systems.

Third, rating purpose is thought to be an important contextual factor that interacts with both the capabilities and the cognitive processes of the rater (Cleveland et al., 1989; Murphy & Cleveland, 1995). Specifically,
research suggests that different purposes may require different levels and types of cognitive processing (Bernardin & Beatty, 1984; DeNisi, Cafferty, & Meglino, 1984; Ilgen & Feldman, 1983; Landy & Farr, 1980). Different types of information processing and integration may be required because raters may use different performance standards or frames of reference for different rating purposes (Murphy & Cleveland, 1995). That is, raters who provide administrative ratings likely make evaluations of the target’s performance relative to other raters, whereas raters who provide developmental ratings likely make evaluations about the target’s absolute level of performance. GT also distinguishes between relative and absolute rating decisions. This distinction is important because error variance is conceptualized and calculated (see Method section for differences in calculations) differently and different conditions (e.g., differing numbers of raters or items) may be required to minimize error (i.e., increase the dependability of ratings) for each type of decision.

Taken together, there are several reasons to expect that rating purpose will affect rating quality; the current study uses GT to assess these potential effects across peer and subordinate ratings. A brief overview of GT is presented below, followed by a review of relevant research that has used GT to assess the quality of performance ratings; finally, hypotheses for the current study are outlined.

**Generalizability Theory**

Generalizability theory is based on analysis of variance and provides a framework for examining the dependability (i.e., reliability) of behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963). In GT, a distinction is usually made between generalizability (G) studies and decision (D) studies (Shavelson & Webb, 1991). The purpose of a G-study is to simultaneously estimate multiple sources of variance (e.g., variance due to ratees, raters, items) within a single, multifaceted experiment. As such, in contrast to classical test theory (which only partitions observed variance into true and random error variance), GT provides more accurate estimates of the dependability of observations (Brennan, 1992, 2000; Nugent & Hankins, 1992). The purpose of the D-study is to use the estimated variance components from the G-study to project reliability estimates under any number of measurement conditions modeled in the G-study (e.g., differing numbers of raters and items). Projecting reliabilities under different measurement conditions is useful because one may readily observe how to improve the dependability (i.e., reliability) of the observations. For a more detailed discussion on GT, and comparisons between
GT and classical test theory, see Brennan (1992), Murphy and DeShon (2000a, 2000b), and Shavelson and Webb (1991).

Few studies have used GT to analyze the dependability of performance ratings. Similar to the current study, these studies generally have investigated variance components associated with the rater, the task, and these components' interactions with the ratee (Clauser, Clyman, & Swanson, 1999). For example, Kraiger and Teachout (1990) investigated the generalizability of performance ratings (made for research purposes) of Air Force jet mechanics across several rating forms and rater sources (i.e., self, peer, and supervisors). Within-rater source analyses revealed that the most variance was attributed to an undifferentiated residual term, followed by the ratee term (i.e., object of measurement or true score variance), with the remaining terms (i.e., those associated with the items, forms, and all interactions) accounting for negligible amounts of variance. Note that these within-source analyses do not include a rater term because only one rater per source was available and, as such, the estimates in Kraiger and Teachout (1990) are analogous to intrarater, rather than interrater, reliabilities.

Webb, Shavelson, Kim, and Chen (1989) investigated the generalizability of job performance ratings of Navy machinist mates across tasks (e.g., job knowledge tests, hands-on performance tests) and rater sources (i.e., self, peer, and supervisor). Ratings in this study were collected as part of a pilot testing phase of a data collection project. For all rater sources, results indicated that the person effect (i.e., object of measurement) accounted for the largest amount of variance and that the second largest effect was the residual variance component. For supervisor and self-ratings, only one rater was available and, as such, the effects attributable to the rater main effect and rater interaction effects could not be computed. However, multiple peer raters were available, and results indicated that the rater-by-ratee interaction effect accounted for a substantial amount (i.e., 23%) of the variance in observed ratings. These results are similar to other studies that have found that rater idiosyncrasies account for a large amount of variability in performance ratings (e.g., Day & Silverman, 1992; Mount, Judge, Scullen, Systma, & Hezlett, 1998; Scullen, Mount, & Goff, 2000).

Greguras and Robie (1998) analyzed the generalizability of developmental ratings made by supervisors, peers, and subordinates. Within-source analyses revealed, across sources, that the largest amount of variance was attributable to an undifferentiated error term, followed by a combined rater main effect and rater-by-ratee interaction effect, followed by the ratee effect (i.e., true score variance). Results further indicated negligible amounts of variance being accounted for by the item effect or the item-by-person interaction effect.
Taken together, the few studies that have used GT to analyze the dependability of performance ratings generally have observed a substantial amount of variance associated with the residual term. When multiple raters were available, Webb et al. (1989) observed a large rater-by-ratee interaction effect for peer ratings, and Greguras and Robie (1998) observed a large rater and rater-by-ratee interaction effect for supervisors, peers, and subordinates. These results are consistent with Murphy and DeShon's (2000a) review of several studies and their conclusion that, in multisource rating systems, rater effects appear to be at least as large and probably larger than the residual variance component. The design of the current study allows for partitioning variance within a source into five components (i.e., variance due to the ratee, item, item-by-ratee interaction, a combined rater main effect and rater-by-ratee interaction, and an undifferentiated error term; see Method section for details). Based on the studies using GT discussed above and Murphy and DeShon's (2002a) review, for both rating purposes the current study hypothesizes:

Hypothesis 1(a-b): For peers (a) and subordinates (b), a substantial amount of variance will be attributed to the undifferentiated error term and the combined rater main effect and rater-by-ratee interaction effect, with less variance being attributed to the ratee effect. Negligible amounts of variance will be accounted for by the item or person-by-item interaction effects.

Using the estimated variance components from the G-study, the dependability (i.e., interrater reliability) of the ratings may be estimated. Numerous reviews and meta-analyses (e.g., Conway & Huffcutt, 1997; Viswesvaran, Ones, & Schmidt, 1996) have investigated the interrater reliability of performance ratings for different rater sources. However, few studies have investigated the effects of rating purpose on interrater reliability. In fact, in their meta-analysis, Conway and Huffcutt (1997) noted that too few coefficients based on administrative ratings were available to investigate the effects of rating purpose on interrater reliability. However, the above discussion on rating purpose suggests that rating purpose may affect the interrater reliability of performance ratings.

Research investigating the effects of rating purpose on rating quality generally has found that ratings made for administrative purposes are more lenient (Dobbins, Cardy, & Truxillo, 1988; Farh, Cannella, & Bedeian, 1991; Farh & Werbel, 1986; Harris, Smith, & Champagne, 1995; Jawahar & Williams, 1997; Longenecker et al., 1987; McIntyre, Smith, & Hasset, 1984; Williams, DeNisi, Blencoe, & Cafferty, 1985), less variable (e.g., Farh et al., 1991), and less accurate (McIntyre et al., 1984) than ratings made for developmental or research purposes. These rater biases and the restricted range associated with administrative ratings likely serve to decrease interrater reliability. Consistent with
this rationale, a study by Farh et al. (1991) provides initial support that interrater reliability will be lower for administrative than for developmental ratings. As such, the current study hypothesizes:

Hypothesis 2(a-b): Peer (a) and subordinate (b) interrater reliability will be lower for administrative ratings than for developmental ratings.

Method

Participants and Administration

Data were collected from managers in a large telecommunications company in which a multisource rating instrument was administered to rate job performance for developmental purposes (in 1995), and 6 months later (in 1996) for administrative purposes. The initial administration included instructions indicating that the ratings would be anonymous and would be used for developmental purposes only. Ratees were expected to generate and execute a developmental plan after they had received their feedback reports following this first administration. At the second administration, instructions indicated that, in addition to the anonymous ratings being used for development, the ratings would be available to ratees’ supervisors as input into administrative decisions (i.e., compensation and promotion). To simplify presentation, we refer to this second administration as being collected for administrative purposes, although ratings actually were made for administrative and developmental purposes combined. Although researchers and practitioners generally refer to “administrative” purposes, in reality, when organizations collect administrative ratings, they generally also use the ratings for developmental purposes (Brutus & Derayeh, 2002; Cleveland et al., 1989).

Our original data set contained 499 ratees; however, only 454 of the ratees were rated for both purposes. These 454 ratees were rated on average by 1.0 supervisor, 5.9 peers, and 3.4 direct reports for developmental purposes (i.e., Time 1) and 1.0 supervisor, 5.4 peers, and 3.2 direct reports for decision-making purposes (i.e., Time 2). Because the GT analyses investigated within-source reliability, at least two raters per ratee within a source were required to estimate the variance components in this study. Given this requirement and the fact that almost every ratee in this study was rated by only one supervisor (98% of the sample for developmental purposes and 99% of the sample for administrative purposes), we were unable to estimate the variance components for supervisors and instead focus on peer and subordinate ratings.

The majority of the ratees (N = 454) were male (65%), with more Whites (92%) than African Americans (6%) or others (2%). The
amount of time the ratees had held their current position was as follows:
less than 1 year (14.6%); 1 to 2 years (14.0%); 3 to 5 years (24.3%); 6 to
10 years (22.1%); and more than 10 years (25.0%). The mean age of
ratees was 44 years ($SD = 5.91$).

Measure

A multirater instrument was designed specifically for the organization under study. The instrument comprised 16 scales (e.g., communicates effectively) with each scale containing 3 items (except Scale 4, which contained 5 items) for a total of 50 items. All items used a 5-point Likert-type scale ranging from $1 = \text{not at all}$ to $5 = \text{to a very great extent}$. Scale names are listed in Table 1.

Overview of Design

Ratings were analyzed separately by source and by scale. Raters and items were treated as random in both the G- and D-studies. Results from each source were then collapsed across rating scales so that we could illustrate the interrater reliability of the average multisource scale across different measurement conditions for each purpose (see Greguras & Robie [1998, pp. 961–962] for more on the analytic rationale for similar analyses). Peer and subordinate ratings were analyzed by scale, using a 2-facet, partially nested design: raters nested within ratees and crossed with items, or an $(r : p) \times i$ design. Raters were nested within ratees because different sets of raters evaluated each ratee, although every rater used all items, and every ratee was rated on all items. This design allows five separate sources of variance to be estimated: (a) variance resulting from the object of measurement (i.e., the ratee, true score variance, $\hat{\sigma}^2_p$), (b) variance resulting from the item ($\hat{\sigma}^2_i$), (c) variance resulting from the interaction of ratee and items ($\hat{\sigma}^2_{pi}$), (d) an undifferentiated variance term that contains both the rater main effect and the person $\times$ rater interaction ($\hat{\sigma}^2_{ri, pr}$), and (e) an undifferentiated variance term that contains the rater $\times$ item interaction, the 3-way interaction of ratee, rater, and item, and any remaining unmeasured sources of error ($\hat{\sigma}^2_{ri, pri, e}$). In computing coefficients for the D-studies, we report the generalizability coefficient ($\hat{\rho}^2$) for relative decisions (i.e., developmental purposes) and the index of dependability ($\hat{\phi}$) for absolute decisions. (i.e., administrative purposes). For relative decisions, each component that interacts with the ratee contributes to the error variance (i.e., $\hat{\sigma}^2_{pi}$, $\hat{\sigma}^2_{ri, pr}$, $\hat{\sigma}^2_{ri, pri, e}$) whereas, for absolute decisions, each component's main effect and its interactions contribute to the error variance ($\hat{\sigma}^2_i$, $\hat{\sigma}^2_{pi}$, $\hat{\sigma}^2_{ri, pr}$, $\hat{\sigma}^2_{ri, pri, e}$).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>3.78</td>
<td>0.68</td>
<td>3.94</td>
<td>0.66</td>
<td>3.80</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>3.39</td>
<td>0.79</td>
<td>3.59</td>
<td>0.75</td>
<td>3.35</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>3.66</td>
<td>0.76</td>
<td>3.84</td>
<td>0.70</td>
<td>3.62</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>3.94</td>
<td>0.61</td>
<td>4.01</td>
<td>0.58</td>
<td>3.82</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>4.05</td>
<td>0.69</td>
<td>4.16</td>
<td>0.67</td>
<td>3.89</td>
<td>0.66</td>
</tr>
<tr>
<td>6</td>
<td>4.05</td>
<td>0.65</td>
<td>4.11</td>
<td>0.63</td>
<td>3.90</td>
<td>0.60</td>
</tr>
<tr>
<td>7</td>
<td>4.07</td>
<td>0.80</td>
<td>4.18</td>
<td>0.68</td>
<td>3.96</td>
<td>0.74</td>
</tr>
<tr>
<td>8</td>
<td>3.60</td>
<td>0.72</td>
<td>3.75</td>
<td>0.67</td>
<td>3.58</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>3.75</td>
<td>0.72</td>
<td>3.89</td>
<td>0.68</td>
<td>3.67</td>
<td>0.62</td>
</tr>
<tr>
<td>10</td>
<td>3.45</td>
<td>0.79</td>
<td>3.70</td>
<td>0.81</td>
<td>3.56</td>
<td>0.64</td>
</tr>
<tr>
<td>11</td>
<td>3.48</td>
<td>0.83</td>
<td>3.66</td>
<td>0.78</td>
<td>3.48</td>
<td>0.64</td>
</tr>
<tr>
<td>12</td>
<td>3.69</td>
<td>0.87</td>
<td>3.82</td>
<td>0.79</td>
<td>3.57</td>
<td>0.66</td>
</tr>
<tr>
<td>13</td>
<td>3.66</td>
<td>0.88</td>
<td>3.80</td>
<td>0.79</td>
<td>3.68</td>
<td>0.67</td>
</tr>
<tr>
<td>14</td>
<td>3.70</td>
<td>0.73</td>
<td>3.83</td>
<td>0.66</td>
<td>3.62</td>
<td>0.61</td>
</tr>
<tr>
<td>15</td>
<td>3.54</td>
<td>0.79</td>
<td>3.71</td>
<td>0.76</td>
<td>3.54</td>
<td>0.60</td>
</tr>
<tr>
<td>16</td>
<td>3.89</td>
<td>0.73</td>
<td>3.98</td>
<td>0.69</td>
<td>3.81</td>
<td>0.64</td>
</tr>
</tbody>
</table>

*Note: All items are rated on a 5-point Likert-type scale ranging from 1 = not at all to 5 = to a very great extent. Scale labels are: (1) empowers and delegates authority, (2) achieves results through others, (3) communicates effectively, (4) focuses on customers and competition, (5) demonstrates high integrity, (6) consistently achieves goals, (7) recognizes the value of teamwork, (8) sees the big picture, (9) deals effectively with ambiguity, (10) takes and supports risks, (11) challenges traditional thinking, (12) displays stamina and resistance, (13) demonstrates technical competence, (14) understands relationships within the organization, (15) is profit and loss focused, and (16) has professional competencies.*
To maximize the number of ratees obtained, we randomly chose two raters for each ratee from each source to estimate the variance components using EMS estimation. This analytic strategy still resulted in a sample size reduction: 452 ratees possessed at least two peer ratings for developmental purposes; 443 ratees possessed at least two peer ratings for administrative purposes; 312 ratees possessed at least two direct report ratings for developmental purposes; and 316 ratees possessed at least two direct report ratings for administrative purposes.

**Results**

Preliminary analyses did not reveal coefficients for any of the scales that deviated more than two standard deviations from the mean of the coefficients across all of the scales. Thus, we retained all 16 scales in the analyses and results presented below are across the 16 scales. Descriptive statistics for each of the scales across rater source and rating purpose are presented in Table 1. Note that for the variance components, significance tests were conducted between rating purposes within rater source by (a) calculating the percentage of total variance that the component explained for each scale, and then (b) conducting t-tests on the means of these percentages.

**Variance Components**

Using GT, multiple sources of variance were estimated for both peer and subordinate ratings. Table 2 presents the results of the G-study analyses collapsed across scales. Consistent with past research (e.g., Farh et al., 1991), the total amount of variance associated with peer ($t[30] = 5.79, p < .05$) and subordinate ($t[30] = 2.39, p < .05$) ratings was significantly larger for developmental ratings than for administrative ratings, suggesting that raters are less willing to make distinctions between ratees when there are important consequences associated with their ratings (e.g., Bernardin, 1981; Jawahar & Williams, 1997; Longenecker et al., 1987). Hypothesis 1(a-b) predicted that for peers and subordinates, respectively, a substantial amount of variance would be attributed to the undifferentiated error term and the combined rater main effect and rater-by-ratee interaction effect, with less variance being attributed to the ratee effect. Negligible amounts of variance were predicted to be accounted for by the item or person-by-item interaction effects. Each of these variance components is discussed below.

For both purposes and rater sources, the largest source of variance modeled in the current study was the variance associated with the combined rater main effect and ratee-by-rater interaction effect. Consis-
TABLE 2

G-Study Estimates: \((r \times p) \times i \text{ Design}\)

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>(\hat{\sigma}^2)</th>
<th>Peers</th>
<th>Subordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person (p)</td>
<td>(\hat{\sigma}_p^2)</td>
<td>.1205</td>
<td>.2162</td>
</tr>
<tr>
<td>Items (i)</td>
<td>(\hat{\sigma}_i^2)</td>
<td>.0217</td>
<td>.0293</td>
</tr>
<tr>
<td>(pi)</td>
<td>(\hat{\sigma}_p^i)</td>
<td>.0303</td>
<td>.0538</td>
</tr>
<tr>
<td>(ri,pr)</td>
<td>(\hat{\sigma}_{ri,pr}^2)</td>
<td>.3954</td>
<td>.4808</td>
</tr>
<tr>
<td>(ri,pr,e)</td>
<td>(\hat{\sigma}_{ri,pr,e}^2)</td>
<td>.3823</td>
<td>.4606</td>
</tr>
<tr>
<td>Total variance</td>
<td></td>
<td>.9502</td>
<td>1.2407</td>
</tr>
<tr>
<td>(\hat{\rho}^2)</td>
<td></td>
<td>.1298</td>
<td>.1182</td>
</tr>
<tr>
<td>(\hat{\phi})</td>
<td></td>
<td>.1475</td>
<td>.1212</td>
</tr>
</tbody>
</table>

Notes: \(n'_i = 1, n'_e = 1\); All estimates are means calculated across the 16 scales. For the variance components, significance tests were conducted between rating purposes within rater source by (1) calculating the percentage of total variance that the component explained for each scale, and then (2) conducting \(t\)-tests on the means of these percentages. Components that are underlined are significantly different between rating purposes within rater source, \(p < .05\). \(\hat{\rho}^2\) = generalizability coefficient for relative decisions; \(\hat{\phi}\) = index of dependability for absolute decisions.

tent with previous research (Greguras & Robie, 1998; Scullen et al., 2000), this combined rater effect accounted for a substantial amount of the total variance in ratings (i.e., 41.6% peers—developmental; 43.3% peers—administrative; 38.8% subordinates—developmental; 46.6% subordinates—administrative). These results indicate that ratee performance was evaluated differently by different raters, but because of the nesting, we cannot conclude whether this was a result of a rater main effect (i.e., some raters being more stringent than others) or the result of a ratee-by-rater interaction effect (i.e., some ratee and rater combinations producing unique results), or both. Further, for subordinates this combined rater effect accounted for a significantly greater amount of total variance in administrative than developmental ratings, \((t[30] = -5.29, p < .05)\). In contrast, for peers these variance components did not account for a significantly different amount of total variance between rating purposes, \((t[30] = -1.05, p > .05)\).

The second largest estimated variance component across rater sources and rating purposes was the undifferentiated error term (see Table 2). The rating model in the current study was quite simple in that it only included two facets (i.e., items and raters). As such, any other source of variability in the observed ratings cannot be estimated with the current design and, instead, is included in this undifferentiated error term. Similar to the combined rater effect, this undifferentiated error term accounted for a substantial amount of the total variance in observed ratings (40.2% peers—developmental; 37.3%
peers—administrative; 37.1% subordinates—developmental; 36.6% subordinates—administrative); these results are consistent with Murphy and DeShon's (2000a) conclusion that rater effects appear to be at least as large and probably larger than the residual variance effect.

The third largest variance component across rating purpose and rater source was the observed true score variance component, \( \sigma_p^2 \) (i.e., variance attributed to the ratee). For peers, the person effect (i.e., true score variance) accounted for a similar percentage of the total variance in administrative ratings (14.8%) and in developmental ratings (12.7%); these estimates were not significantly different from one another (\( t[30] = -1.51, p > .05 \)). However, for subordinates, the person effect accounted for a significantly larger percentage of total variance in developmental (17.4%) than in administrative (12.1%) ratings (\( t[30] = 4.74, p < .05 \)). These results suggest that subordinate ratings are of significantly better quality when made for developmental rather than for administrative purposes and that this effect is not observed for peer ratings.

For both rating sources and rating purposes, little variance was associated with the item effect (\( \sigma_j^2 \)) or the ratee \times item effect (\( \sigma_{ij}^2 \)). These estimates suggest that item variance and the variance due to ratees being differentially rank ordered on items within a scale are relatively small. Finding little variance attributable to these components was expected, given that these scales were developed rationally and factor analytically to be unidimensional. Note that the combined rater effect and the residual effect were both significantly larger than the person effect, which was significantly larger than the item or person-by-item effect (\( p < .05 \)). As such, Hypothesis 1(a-b) are supported.

**Interrater Reliability**

Hypothesis 2(a-b) predicted that peer and subordinate interrater reliability would be lower for administrative ratings than for developmental ratings, respectively. Controlling for the number of raters and items (i.e., 1 rater and 1 item), inspection of the phi and generalizability coefficients in Table 2 reveals that peer interrater reliability estimates between the two rating purposes were not significantly different (\( t[30] = -1.23, p > .05 \)); thus, Hypothesis 2a was not supported. In contrast, subordinates evidenced significantly higher interrater reliability estimates (a difference of .0573) for developmental purposes than for administrative decisions (\( t[30] = 5.04, p < .05 \)); thus, Hypothesis 2b is supported. These findings are consistent with the above findings regarding true score variance. Note, however, that because GT calculates interrater reliability coefficients differently for absolute and relative decisions, it is conceptually and analytically possible to observe significant differences in the
TABLE 3

D-Study Estimates: \((r:p) \times i\) Design

<table>
<thead>
<tr>
<th>No. of raters</th>
<th>No. of items</th>
<th>Peers</th>
<th>Subordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback</td>
<td>Decisions</td>
<td>Feedback</td>
</tr>
<tr>
<td>1</td>
<td>.130</td>
<td>.148</td>
<td>.179</td>
</tr>
<tr>
<td>1</td>
<td>.182</td>
<td>.207</td>
<td>.246</td>
</tr>
<tr>
<td>1</td>
<td>.200</td>
<td>.223</td>
<td>.268</td>
</tr>
<tr>
<td>1</td>
<td>.215</td>
<td>.238</td>
<td>.288</td>
</tr>
<tr>
<td>2</td>
<td>.301</td>
<td>.340</td>
<td>.385</td>
</tr>
<tr>
<td>2</td>
<td>.328</td>
<td>.363</td>
<td>.416</td>
</tr>
<tr>
<td>2</td>
<td>.352</td>
<td>.383</td>
<td>.443</td>
</tr>
<tr>
<td>3</td>
<td>.386</td>
<td>.432</td>
<td>.475</td>
</tr>
<tr>
<td>3</td>
<td>.418</td>
<td>.459</td>
<td>.510</td>
</tr>
<tr>
<td>3</td>
<td>.446</td>
<td>.481</td>
<td>.540</td>
</tr>
<tr>
<td>4</td>
<td>.449</td>
<td>.501</td>
<td>.537</td>
</tr>
<tr>
<td>4</td>
<td>.484</td>
<td>.528</td>
<td>.575</td>
</tr>
<tr>
<td>4</td>
<td>.515</td>
<td>.551</td>
<td>.607</td>
</tr>
<tr>
<td>5</td>
<td>.497</td>
<td>.533</td>
<td>.583</td>
</tr>
<tr>
<td>5</td>
<td>.535</td>
<td>.581</td>
<td>.622</td>
</tr>
<tr>
<td>5</td>
<td>.567</td>
<td>.605</td>
<td>.655</td>
</tr>
<tr>
<td>6</td>
<td>.536</td>
<td>.594</td>
<td>.618</td>
</tr>
<tr>
<td>6</td>
<td>.575</td>
<td>.623</td>
<td>.658</td>
</tr>
<tr>
<td>6</td>
<td>.609</td>
<td>.646</td>
<td>.692</td>
</tr>
<tr>
<td>7</td>
<td>.567</td>
<td>.628</td>
<td>.646</td>
</tr>
<tr>
<td>7</td>
<td>.608</td>
<td>.656</td>
<td>.687</td>
</tr>
<tr>
<td>7</td>
<td>.642</td>
<td>.680</td>
<td>.721</td>
</tr>
<tr>
<td>8</td>
<td>.593</td>
<td>.656</td>
<td>.669</td>
</tr>
<tr>
<td>8</td>
<td>.635</td>
<td>.684</td>
<td>.710</td>
</tr>
<tr>
<td>8</td>
<td>.670</td>
<td>.707</td>
<td>.745</td>
</tr>
<tr>
<td>9</td>
<td>.615</td>
<td>.679</td>
<td>.688</td>
</tr>
<tr>
<td>9</td>
<td>.657</td>
<td>.707</td>
<td>.729</td>
</tr>
<tr>
<td>9</td>
<td>.693</td>
<td>.730</td>
<td>.764</td>
</tr>
</tbody>
</table>

Notes: All estimates are means calculated across the 16 scales. \(r\) = raters; \(p\) = ratees; \(i\) = items. Estimates for feedback purposes are phi coefficients (i.e., reliability coefficients for absolute decisions). Estimates for decision purposes are generalizability coefficients (i.e., reliability coefficients for relative decisions).

percentage of total variance accounted for by the object of measurement (i.e., true score variance) for each rating purpose, yet observe no differences in the dependability of ratings.

D-Study

As noted above, the estimated variance components generated by the G-study can be used to project estimated reliabilities under different measurement conditions. Table 3 contains the results of the D-studies for peers and subordinates for both developmental and administrative purposes. Recall that little variance was associated with the item main effect or ratee-by-item effect. Consequently, estimated reliabilities increased with increases in the number of raters more rapidly than with increases in the number of items employed.
**Discussion**

Organizations use multisource rating systems for a variety of purposes (e.g., Brutus & Derayeh, 2002), yet little is known about how rating purpose affects the quality of multisource ratings (London & Smither, 1995; Sulsky & Keown, 1998). The challenge for researchers and practitioners is to determine under what conditions multisource rating systems for different purposes can exist (Tornow & London, 1998). Using a field sample, the current study used generalizability theory to partition the observed variance of peer and subordinate ratings into different components. These components were then used to estimate peer and subordinate interrater reliability and to project reliability estimates under different measurement conditions. Results indicated that (a) the combined rater and rater-by-ratee interaction effect and the residual effect were substantially larger than the person effect (i.e., object of measurement); (b) for subordinates, the person effect accounted for a significantly greater percentage of total variance in developmental than in administrative ratings (no differences were observed for peers); and (c) subordinates evidenced higher interrater reliability for developmental than for administrative purposes, whereas no such purpose effect was observed for peers. Each of these findings is discussed below.

**Generalizability Theory Analyses**

For peers and subordinates, the largest source of variation was the combined rater main effect and ratee-by-rater interaction effect. These results are consistent with previous research that has found that rater idiosyncratic biases account for the largest percentage of the observed variance in performance ratings (Conway, 1996; Mount et al., 1998; Viswesvaran et al., 1996). Previous studies generally have investigated rater effects for developmental ratings (e.g., Mount et al., 1998; Scullen et al., 2000); however, the results of this study suggest that the finding of a large combined rater effect also generalizes to administrative ratings. Although other studies have reported lower estimates for rater idiosyncratic effects than in this study (e.g., 25%; Conway, 1996), Scullen et al. (2000) note that idiosyncratic rating variance is expected to be larger in nested (as in the current study) than in crossed designs because each rater may exhibit a different degree of variability.

The combined rater effect and rater-by-ratee interaction effect accounted for a significantly greater amount of total variance in administrative than in developmental ratings for subordinates, whereas no purpose effect was observed for peers. Hoyt and Kerns (1999) observed that rater-ratee relationship factors contribute significantly to the rater-by-ratee interaction effect. It may be that subordinates, but not peers, take
these relationship factors more into account when making administrative than developmental ratings. Future research should explore these and other factors that might contribute to this combined rater effect.

The nesting of the current design precluded disentangling the rater main effect from the ratee-by-rater interaction effect. Although this design is commonly encountered with job performance evaluations (Viswesvaran et al., 1996), it does not allow these two rater effects to be estimated separately. However, a recent meta-analysis of 79 studies investigating observer ratings (e.g., ratings of job descriptions, assessments of written essays, evaluations of one’s emotional maturity) by Hoyt and Kerns (1999) provides some insight into the magnitude of, and factors influencing, this combined rater effect. Specifically, in their meta-analysis this combined rater effect accounted for 37% of the variance in ratings, a similar estimate to that in the current study. For a subset of studies in which the rater effect and rater-by-ratee interaction effect could be estimated separately, their results suggested that 8% of the observed variance in ratings was attributable to the rater main effect and 11% of the observed variance was attributable to the rater-by-ratee interaction effect. Hoyt and Kerns (1999) also tested for moderators. For the combined rater main and interaction effects, analyses revealed that ratings contained considerably more bias if raters were required to make inferences rather than explicit judgments (e.g., frequency counts) about performance, if raters were relatively untrained (under 5 hours), if raters did not completely overlap in their observations (i.e., raters potentially observed different behaviors compared to all raters observing the same sample of behaviors), and if raters were acquainted with the ratee. Considering that most, if not all, of these factors generally are present in typical multisource rating systems, it is not surprising that this combined rater effect was the largest estimated variance component in this study.

The percentage of total variance accounted for by the person effect (i.e., object of measurement) in peer ratings was not significantly different between rating purposes. However, for subordinates, the percentage of total variance accounted for by the person effect (and, hence, the interrater reliability of the ratings) was significantly larger for developmental than for administrative ratings. These findings are noteworthy because Scullen et al. (2000) analyzed supervisor, peer, and subordinate ratings for developmental purposes but based their implications on the assumption that the relative magnitudes of the variance components are similar across rating purposes; the current results suggest that meaningful differences exist between rater sources and rating purposes. Future research should investigate why rating purpose affects peer and subordinate ratings differently. For example, perhaps peers feel as if they are in direct competition with the ratee, whereas subordinates may not. Like-
wise, subordinates may fear retribution from the ratee whereas peers may not. Research exploring these or other between-source differences might help explain the differences observed in the current study.

Little variance was associated with the item or person-by-item interaction effects for both rater source and rating purpose conditions. The small amount of variance associated with the item effect was expected given that the performance dimensions were developed factor analytically to be unidimensional. The small person-by-item interaction effects indicate that raters were not very likely to be rank ordered differently as a function of the item.

D-study. Using the variance estimates from the G-study as input, a D-study was conducted. Because little variance was associated with the item, or the person-by-item interaction effects, projected reliability estimates increased much more with increases in the number of raters, rather than number of items; increasing the number of raters can significantly reduce the effects of bias and random error (Cronbach et al., 1972; Scullen et al., 2000). Overall, results suggest that the within-source reliability for conditions normally encountered in practice (e.g., typically 4–6 raters, London & Smither, 1995), regardless of rating purpose, is lower than the conventional criterion of .70 for a minimally acceptable level of reliability (Nunnally, 1978). Given this information, users of multisource rating systems may need to increase or decrease the number of raters depending upon the rater level and rating purpose in order to maximize the efficiency of such systems; the reliability estimates of the D-study can aid in this endeavor.

The current results suggest an interesting irony in the implementation of multisource rating systems. That is, wanting more return on their investments, organizations are increasingly using their multisource rating systems for administrative purposes (Dalessio, 1998; Ghorpade, 2000). However, the information obtained from subordinates for administrative purposes is of lower quality than the information obtained for developmental purposes, thereby likely reducing (rather than increasing) an organization’s actual return on investment.

Order Effect of Rating Purpose

A few comments on the design of the current study are warranted. Similar to a study by Harris et al. (1995), all ratings for a given purpose were made after ratings for a different purpose. As such, this design cannot disentangle the effects due to time and order (Administration 1 and Administration 2) and those due to rating purpose (administrative or developmental). As Harris et al. (1995) note, the first administration of ratings may have a priming effect on the later ratings. However, there
are two primary reasons why we believe this potential confound concern is minimized in the current investigation.

First, we acknowledge that because the first administration was for developmental purposes, ratees would be expected to improve their performance after receiving developmental feedback (e.g., Smither et al., 1995). This improvement in performance might, in part, explain why there is less total variance associated with the administrative ratings than with the developmental ratings (i.e., because of performance improvement most ratees might be clustered at the favorable end of the scale). However, it is not apparent how this potential confound would explain differences in the percentage of total variance accounted for by any given variance component. For example, if purpose did not have an effect, we would expect that the percentage of total variance explained by the person effect to be the same across rating purposes, yet our results suggest otherwise for subordinates. Moreover, it is interesting to note that we observed differences between rater sources (e.g., subordinates were significantly more reliable for developmental than for administrative purposes, but peer interrater reliability did not differ significantly as a function of rating purpose). It is not clear why this effect of order, if indeed a confound, would operate differently for peer than for subordinate ratings. We think these rater-source differences in and of themselves are interesting and future research should attempt to explain the differences.

Second, the design utilized in the current study reflects what typically is done in organizations and therefore has a high degree of external validity. Specifically, several authors suggest (e.g., London, 2001; Dalessio, 1998), and organizations generally adhere to (Timmreck & Bracken, 1997), phasing in multisource feedback systems such that the system is first used for developmental purposes and then is phased in to be used for administrative purposes. As such, one purpose of the current investigation was to investigate differences in rating characteristics as a function of rating purpose under conditions commonly encountered in organizations. Results from this study can inform organizations who plan to shift the purpose of their multisource system from developmental purposes to administrative purposes. As such, assessing the effects of changing from one purpose to another should not be considered a confound per se, but rather, such analyses capture the phenomenon of interest (i.e., the effects on ratings of transitioning from one rating purpose to another).

Limitations

In addition to the possible confounding of rating occasion and purpose discussed above, there are other possible limitations of the current
study to consider. First, only one organization and one instrument were investigated. The generalizability of the observed findings to other instruments or contexts should be established. Second, the current rating model was rather simple and the second largest source of variance across rater sources and rating purposes was the undifferentiated residual effect ($\hat{\sigma}^2_{r_i,pri,e}$). This suggests that factors other than the ones included in this study (e.g., type of organization, opportunity to observe, experience with rating systems) contribute to the variability of peer and subordinate ratings for both rating purposes; Future research should expand the rating model to include more of these factors.

**Future Research**

There are several additional directions for future research. First, studies should extend the current findings by investigating why rating purpose affects the observed ratings, by exploring potential mediators of these effects. Drawing from Wherry's (1952) theory of performance ratings, potential mediators might include whether the rater believes that job performance is controlled by the ratee or the work situation; or whether the rater is aware that the ratings may need to be justified to the ratees (for a complete list of theorems and corollaries, see Wherry & Bartlett, 1982). Second, research might explore the concept of interrater reliability as an index of performance. That is, some researchers have argued that we should not expect raters within a source to agree (e.g., Tornow, 1993). Perhaps reliability might serve as an indication of a target's consistency in performance across projects or employees. Third, research should investigate how several multisource process factors might interact with rating purpose to affect the quality of observed ratings (e.g., rater anonymity, number of raters, whether the raters were chosen or assigned, who has access to the information). Fourth, more attention should be given to the broad categories we use when labeling rating purposes. For example, does the quality of ratings change if the "administrative" purpose is for promotion versus pay increases? Finally, research should explore why rating purpose had different effects on peer and subordinate ratings.

**REFERENCES**


