Butler University Libraries

Butler University
Digital Commons @ Butler University

Scholarship and Professional Work - Business

College of Business

1998

Where all the Children are Above Average: A Meta analysis of the Performance Appraisal Purpose Affect

J. Jawahar

Chuck R. Williams
Butler University, crwillia@butler.edu

Where all the Children are Above Average: A Metaanalysis of the Performance Appraisal Purpose Affect

By: Jawahar, I. M., Williams, Charles R

More than 40 years ago, Taylor and Wherry (1951) hypothesized that performance appraisal ratings obtained for administrative purposes, such as pay raises or promotions, would be more lenient than ratings obtained for research, feedback, or employee development purposes. However, research on appraisal purpose has yielded inconsistent results, with roughly half of such studies supporting this hypothesis and the other half refuting it. To account for those differences, a meta-analysis of performance appraisal purpose research was conducted with 22 studies and a total sample size of 57,775. Our results support Taylor and Wherry's hypothesis as performance evaluations obtained for administrative purposes were, on average, one-third of a standard deviation larger than those obtained for research or employee development purposes. In addition, moderator analyses indicated larger differences between ratings obtained for administrative and research purposes when performance evaluations were made in field settings, by practicing managers, and for real world subordinates. Implications for researchers and practitioners are discussed.

Several models have emphasized the potential influence of performance appraisal purpose (PAP) on leniency and accuracy of ratings (e.g., DeCotiis & Petit, 1978; DeNisi, Cafferty, & Meglino, 1984). Based on Wherry's work (Taylor & Wherry, 1951), it is generally predicted that ratings obtained for administrative purposes, such as pay increases, promotions, or retention, are likely to be more lenient and less accurate than those obtained for research, feedback, or employee development purposes. This hypothesis is based on the premise that raters bias ratings obtained for some purposes versus others. Raters may intentionally bias administrative ratings to avoid providing negative feedback (Fisher, 1989), to avoid negative consequences associated with harsh but accurate ratings (e.g., no pay increase), to obtain positive consequences (e.g., pay increase), or to motivate a poor performer (Murphy & Cleveland, 1991). In contrast to administrative ratings, the confidential nature of research ratings is likely to encourage managers to honestly record their "true" evaluations of subordinate's work performance. Likewise, when ratings are obtained for training or employee development purposes, ratings may be less lenient as managers are likely to be motivated to help employees accurately identify and correct performance deficiencies.

In theory, reliable, valid appraisal data that differentiate above average performers from average and below average performers should help managers make sound personnel decisions, such as pay increases, promotions, and terminations. However, the PAP effect suggests that managers will be reluctant to make these important distinctions in performance. A recent review supports this view. Bretz, Milkovich, and Read (1992,p. 333) concluded that the "norm in U.S. industry is to rate employees at the top end of the scale." Unfortunately, when this occurs, appraisal ratings are of dubious value to managers faced with difficult and important personnel decisions. Indeed, the pervasiveness of appraisal leniency brings to mind humorist Garrison Keillor's fictitious town of Lake Wobegone in which, he jokes, "all of the children are above average."

Because performance appraisals are the most commonly used criterion measure in selection research, the PAP effect is likely to be of concern to personnel researchers as well. If the PAP effect occurs as predicted, then appraisal ratings collected for administrative purposes could suffer from serious criterion contamination relative to appraisal ratings collected for research or development purposes. Some view the criterion contamination problem as so severe that they recommend "extreme caution" when using administrative ratings for selection validation (Ilgen, Barnes-Farrell, & McKellin, 1993). According to Gatewood and Feild (1994), the criterion contamination associated with administrative ratings should result in smaller validity coefficients relative to the validity coefficients found with research ratings. Two studies by McDaniel and his associates provide mixed evidence on this issue. In a meta-analysis of the selection validity of employment interviews, McDaniel, Whetzel, Schmidt, and Maurer (1994) found that the average validity coefficient was .47 for research criteria versus .36 for administrative criteria. In general, this pattern held for job-related interviews (.50 vs. .39), structured interviews (.51 vs. .37), but not for unstructured interviews (.38 vs. .41). By contrast, in a recta-analysis of selection validities for the Wonderlic Personnel Test and the Otis Intelligence Test, Rogers and McDaniel (1994) found just the opposite. The average selection validity was .57 for research criteria and .83 for administrative criteria. However, Rogers and McDaniel (1994) attributed the latter result to second order sampling error resulting from the small number of studies (k = 10) and small total sample size (n = 581) in their administrative criteria subgroup.

Previous Research on PAP Effects

Contrary to widespread confidence in the PAP effect, much of the research investigating the influence of PAP on appraisal leniency and accuracy has been inconsistent. Some studies support the PAP effect (Aleamoni & Hexner, 1980; Beckner, Highhouse, & Hazer, 1995; Bernardin & Orban, 1990; Driscoll & Goodwin, 1979; Farh, Cannella, & Bedeian, 1991; Farh & Werbel, 1986; Gmelch & Glasman, 1977; Harris, Smith, & Champagne, 1995; Jawahar, 1994; Pritchard, Peters, & Harris, 1973; Taylor & Wherry, 1951; Veres, Field, & Boyles, 1983; Waldman & Thornton, 1988; Williams, DeNisi, Blencoe, & Cafferty, 1985; and Zedeck & Cascio, 1982). However, many others do not (e.g., Berkshire & Highland, 1953; Bernardin, 1978; Bernardin & Cooke, 1992; Centra, 1976; Hollander, 1965; McIntyre, Smith, & Hassett, 1984; Meier & Feldhusen, 1979; Murphy, Balzer, Kellam, & Armstrong, 1984; Sharon & Bartlett, 1969; Shore, Adams, & Tashchian, 1995).

Evidence from previous meta-analytic studies (Kraiger & Ford, 1985; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Murphy, Herr, Lockhart, & Maguire, 1986; Rogers, & McDaniel, 1994; Tett, Jackson, & Rothstein, 1991) has also been mixed. Furthermore, with one exception (Murphy et al., 1986), previous meta-analytic studies have not been based on experimental investigations in which appraisal purpose was manipulated to discern its direct effects on the leniency of performance ratings. Instead, prior meta-analytic studies were based on indirect evidence in which the size of a relationship between two variables was compared for studies using administrative ratings against those using research ratings. In other words, previous studies examined whether the PAP effect acted as a moderator. For example, Kraiger and Ford's examination of appraisal purpose contrasted the correlation between race and performance appraisal ratings in studies using administrative ratings (k = 18, n = 6,955, r = .17) with studies using research ratings (k = 37, n = 8,259, r = .16). They found no difference. Tett et al. (1991) used a similar between studies approach to determine whether selection validities for personality measures were larger when performance appraisal criterion data were gathered for research purposes (k = 70, n = 10,644, r = .15) than for administrative purposes (k = 23, n = 2,334, r = .18). The difference was small, in the opposite direction, and nonsignificant. And as discussed above, McDaniel et al. (1994) and Rogers and McDaniel (1994) also used a between studies approach with the results, respectively, supporting the PAP effect for employment interview validities and refuting it for cognitive ability tests.

Only Murphy et al.'s (1986) meta-analysis contained experimental studies (see Table 5, p. 657) that directly manipulated performance appraisal purpose. However, it was based on only 9 studies that were used for just one analysis: to determine if performance appraisal purpose moderated the difference in appraisal ratings between "paper people" (k = 2, d = 1.73, n = not listed) and those based on behavioral observations of "real people" (k = 7, d = .26, n = not listed). By contrast, the broad purpose of our meta-analysis was to determine whether performance appraisal purpose directly affects the leniency or level of performance appraisal ratings. This is very different from asking whether appraisal purpose moderates the relationship between appraisal ratings and race, personality, employment interviews, cognitive ability, and paper versus real people. Given the practical and research importance of performance appraisal ratings, this study attempted to account for the inconsistent results in PAP research by conducting a meta-analysis of 22 studies (n = 57,775) that directly manipulated appraisal purpose to determine its direct effect on appraisal leniency.

Moderators

Contrary to common belief, the results of PAP research have been relatively inconsistent from study to study. In part, this inconsistency may be a function of sampling error in individual studies. If so, the metaanalysis will account for those differences. However, inconsistency may also be due to real differences across studies, such as those related to the research setting (field vs. lab, student vs. organizational raters, paper people vs. video vs. behavior observation, upward vs. downward appraisal, and ratings vs. decisions) or measurement issues (forced choice scales vs. graphic rating scales vs. behaviorally anchored rating scales, and multiple purposes vs. specific purposes).

One explanation for the inconsistent findings may be the research setting itself. Typically, raters in lab studies are provided with performance information "all-at-once" whereas raters in field studies obtain information over an extended period of time. Raters and ratees do not generally interact in lab studies whereas they do so on an ongoing basis in field studies. In addition, ratings in lab studies have little or no real consequences for either raters or ratees, whereas ratings in field studies generally lead to significant consequences (e.g., pay increases, promotions, etc.) and may affect the relationship between raters and ratees (Ilgen & Favero, 1985). For these reasons, we expect ratings in field studies to be more lenient than ratings in lab studies. Indeed, reviews of the performance appraisal literature (e.g., Bernardin & Villanova, 1986; Kraiger & Ford, 1985; Murphy & Cleveland, 1991) generally indicate larger leniency effects in ongoing organizations than in laboratory investigations. Accordingly, we expect the PAP effect to be larger in field studies than in lab studies.

Second, the PAP effect may depend on whether ratings were provided by college students or organizational raters. In comparison to managers, students typically have limited experience with actual performance appraisals and are therefore unlikely to understand the political and the affective nature of performance appraisals in organizations (Longenecker, Sims, & Gioia, 1987). For instance, "real" managers sometimes inflate ratings to reward subordinates they like, to encourage loyalty, or to promote their personal agendas. Sometimes managers deflate ratings to punish rebellious subordinates despite otherwise good performance (see Longenecker et al., 1987). For these reasons, we expect the PAP effect to be larger in studies that collected ratings from organizational raters than in studies that collected ratings from student raters.

Third, the source or direction of appraisal may moderate purpose effects. Studies have used subordinate evaluations (upward), peer evaluations, and supervisor evaluations (downward) to investigate the influence of appraisal purpose on leniency of ratings. From an information processing perspective, raters at different levels in an organization may have more or fewer opportunities to observe ratee performance. In addition, raters at different levels may weight performance dimensions differently and come to different conclusions regarding how well ratees perform their jobs. Furthermore, in most companies, downward appraisals are used for administrative purposes, whereas upward (subordinate) appraisals are generally used for feedback and management development purposes (Bernardin, 1986). Accordingly, we expect the PAP effect to be larger for downward appraisal than for upward appraisal.

Fourth, the PAP effect may be moderated by performance appraisal rating formats, such as graphic rating scales (GRS), forced-choice scales (FCS), or behaviorally anchored rating scales (BARS). Greene, Bernardin and Abbott (1985) compared the quality of ratings obtained on different rating formats and concluded that although formats make little difference when ratings are obtained for research purposes they may yield significant differences when used to obtain administrative ratings. Likewise, the few studies that obtained ratings on both GRS and FCS (e.g., Taylor & Wherry, 1951) and GRS and mixed standard formats (e.g., Bernardin & Orban, 1990) found graphic rating formats to be less resistant to appraisal purpose effects. Finally, it should be remembered that FCS were introduced into the performance appraisal literature to specifically foil deliberate attempts to inflate ratings (Berkshire & Highland, 1953), whereas BARS were introduced to improve rating accuracy (Landy & Farr, 1980). Thus, we expect the PAP effect to be larger for GRS than for FCS or BARS.

Fifth, the kind of dependent measure used, ratings versus decisions, may moderate the relationship between appraisal purpose and leniency. In general, we have argued that raters inflate appraisal ratings when those ratings are instrumental for obtaining positive outcomes (e.g., pay increases) or when they want to avoid negative outcomes, such as being confronted by a subordinate who is dissatisfied with a low performance rating (Murphy & Cleveland, 1991). However, when managers are asked to make personnel decisions (e.g., pay raises or retention or promotions) directly without first providing ratings, the tendency to be lenient should be even stronger. For instance, actual pay increase decisions (i.e., awarding a larger or smaller pay increase) are likely to have more impact on the rater-ratee relationship (and thus more impact on raters) than appraisal ratings which may or may not be closely linked to the size of one's pay increase. Hence, we expect the PAP effect to be larger for personnel decisions than for performance ratings.

Sixth, the PAP effect may vary depending on whether appraisal ratings were used for single or multiple purposes. For instance, in Centra's (1976) study, subjects in the administrative condition were informed that tenure, salary, and promotion decisions would be contingent upon the ratings they assigned to ratees. Clearly, from a research perspective, combining specific purposes in this manner confounds appraisal purpose. However, from the appraisal rater's perspective, combining what are typically separate decisions into one package of multiple decisions may encourage raters to be even more lenient, now that their rating decisions have even greater consequences for ratees. Therefore, we expect the PAP effect to be larger for multiple purpose ratings than for single purpose ratings.

Finally, we re-examined Murphy, Herr, Lockhart, and Maguire's (1986) hypothesis that performance appraisal effect sizes are larger in studies when raters appraise paper people than when they appraise ratees after observing their behavior (i.e., behavior observation). Murphy et al. (1986) reasoned that performance appraisal effect sizes would be larger for paper people because, in contrast to behavior observation, the performance information presented to raters in paper people studies is less ambiguous and contains less irrelevant or distracting information; thus, making it easier for raters to detect true differences in performance across paper people ratees. Overall, Murphy et al. found that effect sizes were on average larger for paper people (d = .42) than for behavior observation (d = .31). Murphy, et al. extended their analyses to investigate if appraisal purpose moderated this paper people effect. Indeed, administrative ratings were significantly higher than research ratings (i.e., the PAP effect) in paper people studies (d = 1.74) than in studies that involved behavior observation (d = .26). We reanalyzed Murphy et al.'s hypothesis using a much larger number of studies and a much larger total sample size. In addition, we extended their analysis by examining the PAP effect for videotaped subjects, in addition to paper people and behavioral observation.

Method

Data Collection

In contrast to previous meta-analytic studies that examined the indirect, moderating effects of performance appraisal purpose (e.g., Kraiger & Ford, 1985), our meta-analysis was based on experimental studies that manipulated appraisal purpose to determine its direct effect on leniency of performance appraisal ratings.

Studies were identified for possible inclusion in the meta-analysis: (a) through a traditional literature search of the performance appraisal literature, (b) by scanning the abstracts of articles published in ***Personnel Psychology***, the Journal of Applied Psychology, the Academy of Management Journal, and Organizational Behavior and Human Decision Processes; (c) through electronic searches in Infotrac, First Search, and Dissertations Abstracts; and (d) through two-rounds of announcements, 6 weeks apart, on Academy of Management electronic discussion lists, such as RMNET and HRNET. In those announcements, we encouraged authors to contact us if they had conducted unpublished performance appraisal purpose research that was presented at a conference, printed as a proceedings, sitting in a file drawer, presently submitted for journal review, now in-press, or which had been completed for a master's thesis or doctoral dissertation.

These procedures generated an initial list of 50 possible studies for inclusion. Studies were excluded from the meta-analysis if they did not manipulate appraisal purpose, if they did not collect ratings for at least one administrative purpose and either research or employee development purpose, if they did not measure appraisal leniency, or if they did not report sufficient data to allow a determination of the study sample size or effect size. These criteria resulted in the exclusion of several well known studies associated with the performance appraisal purpose literature. For example, the best known of these, Zedeck and Cascio (1982), could not be included because it analyzed the average standard deviation of appraisal ratings, rather than the level or average level of appraisal ratings which is used when measuring performance appraisal leniency. On the other hand, we also excluded a number of studies measuring leniency (e.g., McIntyre et al., 1984) because they did not provide the information (i.e., means and standard deviations, t test, f test, etc.) needed to calculate an effect size. These search efforts, in conjunction with our study inclusion and exclusion criteria, yielded a total usable sample of 22 studies (k = 22) and a total usable sample size (T) of 57, 775.

Analysis

General meta-analysis procedures. We used the following general procedures when conducting our meta-analysis. First, we calculated the average effect size, d, across studies. The effect size d, is used when cumulating effect sizes from experiments. It is computed by subtracting the mean of the control group (i.e., ratings obtained for research purposes, in our case) from the mean of the treatment group (i.e., ratings obtained for administrative purposes), and then dividing that difference by the standard deviation of the control group (Glass, 1977), or by the pooled standard deviation from the control and treatment groups (Hunter & Schmidt, 1990). Thus, like a z-score, d, indicates in standard deviation units how large a difference exists between experimental conditions.

Next, the observed variance in effect sizes was determined and then corrected by subtracting the variance attributable to sampling error. Because meta-analysis reduces but does not eliminate sampling error, 95% confidence intervals were placed around the mean effect size to determine whether those estimates were different from zero. The stability of mean effect sizes was determined using a "fail-safe N" formula provided by Hunter and Schmidt (1990,p. 513). The fail-safe N formula estimates the number of missing or unpublished studies with null results (i.e., d = 0) necessary to reduce the average effect size to a critical value of d, that we specified a priori, as -.01.[1] The larger the absolute size of the fail-safe N, the greater the stability of the estimate of the mean population effect size (Carson, Schriesheim, & Kinicki, 1990).

Both the average d and the estimated population standard deviation were then corrected for unreliability by dividing by the square root of the reliability of the dependent variable. Because reliability data were inconsistently reported, we were unable to build an artifact distribution to estimate reliability. Therefore, we used a reliability estimate of .60 derived by Pearlman (1979) to make the correction for unreliability. This estimate has been used in other meta-analyses involving performance ratings (McDaniel et al., 1994).

Tests for theorized moderators were conducted if study artifacts did not account for more than 75% of the variance across studies and if the 95% credibility intervals surrounding the mean corrected d included zero. The overlap between moderator subgroup mean effect sizes was tested using a critical ratio developed by Hunter and Schmidt (1990,pp. 436-438).

Outlier sample sizes in meta-analysis. Because 92% of the total sample size in our meta-analysis was attributable to just three studies, Berkshire and Highland (1953; n = 4,053), Pritchard, Peters, and Harris, (1973; n = 34,504), and Taylor and Wherry (1951; n = 14,712), we were concerned about the possible biasing influence of sample-size outliers.

Meta-analytic results are easily skewed by individual studies with particularly large (or small) sample sizes or effect sizes. Osburn and Callender (1992) found that when there are outlier sample sizes, as in our study, the traditional meta-analytic practice of sample-size weighting individual study effect sizes can result in underestimation of the standard error of the mean used to compute confidence intervals. Similarly, when there are outlier sample sizes, sample-size weighted variances will typically underestimate the observed variance in study correlations, producing an underestimation of the true population variance which then leads to an underestimation of credibility intervals.

We used two methods to guard against the possible influence of outliers in our results. First, we used Huffcutt and Arthur's (1995) sample-adjusted meta-analytic deviance (SAMD) statistic to test for meta-analysis outliers. Second, following Osburn and Callender's (1992) recommendations, we conducted an unweighted meta-analysis. Because study effect sizes are not weighted by sample size in this procedure, the mean effect size produced is much more stable because it cannot be unduly affected by studies with overly small or large sample sizes. For instance, in a meta-analysis of performance and voluntary turnover, Williams and Livingstone (1994) used the unweighted procedure and found that the inclusion or exclusion of a large sample-size outlier (n = 3,986) did not affect the size of the unweighted mean correlation. However, the same outlier changed the size of the sample-size weighted correlation by 33%! Furthermore, because sample-size weighted variances are affected much more strongly by outliers than sample-size weighted means (Hunter & Schmidt, 1990), Osburn and Callender (1992) also recommend using an unweighted variance.

In all, an unweighted meta-analysis is a much more conservative approach than the traditional, sample-size weighted meta-analytic procedure. However, as a safeguard, the results of each unweighted analysis will be compared to the results obtained from a parallel sample-size weighted meta-analysis. Any difference in the estimates obtained from the unweighted and the sample-size weighted approaches will be highlighted and discussed.

Results

Statistical Outliers and Primary Meta-Analysis

First, the effect size for each of the 22 studies (available from the study authors) was computed. Next, we used Huffcutt and Arthur's (1995) sample-adjusted meta-analytic deviance (SAMD) statistic to test for the presence of outliers. The first step in using the SAMD statistic is to compute the average sample-size weighted effect size after removing the effects of a particular study, d_i. The more the average effect size changes when a study is removed from the sample, the greater the likelihood of that study being an outlier. Because the sample-size weighted mean was .40, removing most studies had little or no effect on the sample-size weighted mean. Only the removal of the Pritchard et al. (1973) study produced a sizable change, from .40 to .31. The second step in using the SAMD statistic is to compare a particular study's effect size, d_i, to the average sample-size weighted effect size that has been calculated without including that d_i. The larger the raw difference between those numbers, the greater the likelihood of a statistical outlier. The largest difference, -.80, was associated with Bernardin's (1978) effect size of -.40. The final step in calculating the SAMD statistics is to account for the influence of sample-size outliers. The larger a study's sample size compared to other studies, the greater the effect it has on sample-size weighted meta-analytic results.

Values of the SAMD statistic, ranged from -9.77 to a high of 11.13, with a mean of - 1.05 and a standard deviation of 3.96. Because only the magnitude and not the direction of the difference matters, and because the SAMD statistic does not have a "critical value" to indicate statistical significance, Huffcutt and Arthur (1995) recommended that SAMD statistics be interpreted by rank-ordering the absolute value of SAMD estimates to form a Scree plot, similar to that used in exploratory factor analysis. The slope of the scree plot for our data, which is available from the study authors, indicated that three studies have much larger SAMD values than the others in our sample. Those outlier studies were Pritchard et al. (1973,absolute value SAMD = 11.13) because of its extraordinary sample size (n = 34,504), Berkshire and Highland (1953,absolute value SAMD = 9.77) because of its small effect size (d = .10) and large sample size (n = 4,053), and Sharon and Bartlett (1969,absolute value SAMD = 7.89) because of its small negative effect size (d = -.09) and moderately large sample size (n = 1,046).

The standard treatment for statistical outliers is to remove them from the analysis. However, because of the moderate number of studies in our primary meta-analysis (k = 22) and the even smaller number of studies in some of our moderator subgroups, we were reluctant to eliminate the statistical outliers from our analysis without good reason. Eliminating these studies might produce more accurate meta-analytic estimates, but it could also significantly increase second-order sampling error, especially in moderator subgroups which typically contain one-half to one-third the number of studies found in the primary meta-analysis. So we followed Huffcutt and Arthur's (1995) advice to conduct a follow-up investigation to determine if the outlier studies were different in terms of samples, measures, or other relevant characteristics from the other studies in our analysis. For example, in a re-examination of their own meta-analysis of employment interviews (Huffcutt & Arthur, 1994), Huffcutt and Arthur (1995) found that several outlier studies artificially inflated their results by computing sample statistics on the top and bottom criterion groups, rather than the entire range of scores. By more closely examining their data, they were able to determine why some studies in their meta-analysis were outliers. When we closely examined the Pritchard, et al. (1973), Berkshire and Highland (1953), and Sharon and Bartlett (1969) studies, we did not find any statistical, sample, or procedural anomalies to explain why these studies were statistical outliers in our distribution of studies. Thus, unlike Huffcutt and Arthur (1995), we could not discern any meaningful differences that would give us good reason to exclude outlier studies from the analysis.

So, rather than eliminate them, we minimized their influence by conducting an unweighted meta-analysis in which, contrary to standard meta-analysis techniques, neither the mean nor variance are weighted by sample size (Osburn & Callender, 1992). As discussed above, one of the main advantages of this approach is that the meta-analytic estimates it produces are much more stable. For example, as mentioned above, eliminating the Pritchard et al. (1973,n = 34,504) study from the analysis reduced the sample-size weighted mean d from .40 to .31. By contrast, removing the Pritchard et al. study from the unweighted analysis only reduced the unweighted mean d from .25 to .24. Thus, unweighted meta-analytic estimates are much less likely to change as new study data become available over time. Second, Osburn and Callender's (1992) monte carlo studies found that sample-size weighted means and variances are most likely to be misestimated when there are large samplesize outliers and fewer than 50 studies in a meta-analysis. With 22 studies overall, and with three of those studies accounting for 92% of the total sample size, this was likely to occur with our data. In all, we felt that the unweighted meta-analysis presented the fewest tradeoffs. Given the characteristics of our data, it should produce more accurate and stable meta-analytic estimates, and, it does not require elimination of studies.

Table 1 shows the results of the primary unweighted meta-analysis. Because the confidence intervals (.12 to .38) did not include zero, it can be concluded that there is a nonzero effect size. In fact, the average d corrected for unreliability was .32. This indicates that performance appraisal ratings are nearly one-third of a standard deviation larger when ratings are obtained for administrative purposes than for research purposes. Furthermore, this estimate is very stable because the fail-safe N indicates that it would take 528 missing studies with null effects to reduce the mean d to the critical value of -.01.

The small amount of variance attributable to sampling error, just 2%, plus 95% credibility intervals (-.46 to 1.11) that included zero, clearly indicate a large amount of true variance across studies. Therefore, we tested for the hypothesized moderators.

Moderator Analyses

We predicted that research setting issues (field vs. lab, student vs. organizational raters, paper people vs. video vs. behavior observation, upward vs. downward appraisal, and rating performance vs. making a decision) and measurement issues (FCS vs. GRS vs. BARS, and multiple purposes vs. specific purposes) would moderate the effects of performance appraisal purpose. Three steps were involved in placing individual studies into moderator subgroups. First, we created a coding sheet that identified the decision rules to be used in placing studies into one moderator group or another. Second, the study co-authors independently coded each individual study. Independent assessment resulted in initial agreement of 94%. Finally, joint discussion by the co-authors was used to resolve remaining disagreements.

Field versus laboratory studies. As predicted, the performance appraisal purpose effect (i.e., administrative ratings being more lenient than research ratings) was larger in field settings (mean corrected d = .41) than in laboratory settings (mean corrected d = .09). The confidence intervals shown in Table 1, and the critical ratio (Z = 1.94, p < .05) indicate that the field and laboratory setting subgroup means and distributions are significantly different. However, the smaller effect size for laboratory settings should be viewed as tentative, given its smaller fail-safe N, 36, and that it was based on only 6 studies.

Student raters versus organizational raters. As predicted, the performance appraisal purpose effect was larger with organizational raters (mean corrected d = .50) than with student raters (mean corrected d = .22). The confidence intervals shown in Table 1, and the critical ratio (Z = 1.82, p < .05) show that the organizational rater and student rater subgroup means and distributions are significantly different. However, the larger effect size for organizational raters should be viewed as somewhat tentative given that it was based on just eight studies. But, with a fail-safe N of 304 studies, there is some justification for assuming that this mean effect size could be stable.

Paper people versus behavioral observation versus video. Previous research results concerning appraisal stimuli have been mixed. We found that the performance appraisal purpose effect was larger when ratings were based on live observation of ratees (mean corrected d = .41) rather than paper people (mean corrected d = .15) or videotapes (mean corrected d = -.04). However, the confidence intervals shown in Table 1, and the critical ratios (observation vs. paper people, Z = 1.18, p < ns; observation vs. video, Z = 3.20, p < .05; paper people vs. video, Z = .89, ns) show that the only significant difference was between live observation and video. But because the paper people subgroup was based on only 4 studies and the video subgroup was based on just 2 studies, these results should be viewed as tentative because they could easily change with the inclusion of additional research studies.

Upward versus downward appraisal. The performance appraisal purpose effect was larger for downward appraisal (mean corrected d = .40) than for upward appraisal (mean corrected d = .09). The confidence intervals shown in Table 1, and the critical ratio (Z = 1.82, p < .05) show that the downward appraisal subgroup mean and distribution are significantly larger than those of the upward appraisal subgroup. Again, however, both sets of subgroup results should be viewed as somewhat tentative, given that they were based on just 8 and 11 studies. On the other hand, with a fail-safe N of 330 studies for downward appraisal, there is justification for assuming that this mean effect size has some stability.

Ratings versus personnel decisions. The performance appraisal purpose effect was nearly the same when making performance ratings (mean corrected d = .34) as when making personnel decisions (mean corrected d = .32). The overlapping confidence intervals shown in Table 1, and the critical ratio (Z = .07, p < ns) show that there is little difference between these moderator subgroups. However, because the decision subgroup was based on just 3 studies, these results could change as more study data become available.

Graphic rating scales (GRS) versus forced choice scales (FCS) versus behaviorally anchored rating scales (BARS). The performance appraisal purpose effect was larger when GRS (mean corrected d = .34) were used than when FCS (mean corrected d = .18) or BARS (mean corrected d = .03) were used. However, the confidence intervals shown in Table 1, and the critical ratios (GRS vs. FCS, Z = .75, ns; GRS vs. BARS, Z = .56, ns; and, FCS vs. BARS, Z = .27, ns) show a significant amount of overlap which indicates that these subgroup means and distributions are not significantly different. However, because the forced choice subgroup was based on only 3 studies, while the BARS subgroup was based on just 2 studies, these results could easily change in the future with the inclusion of additional research studies.

Multiple purposes versus single purpose. Although not significant, the performance appraisal purpose effect was slightly larger when ratings were obtained for multiple purposes (mean corrected d = .36) than when they were obtained for a single purpose (mean corrected d = .31). The overlapping confidence intervals shown in Table 1, and the critical ratio (Z = .91, p < ns) show a significant overlap between these moderator subgroups. However, because the multiple subgroup was based on just 7 studies, these results could change as more study data become available.

Unweighted versus weighted meta-analysis. Out of the 22 primary and secondary meta-analyses conducted in this study, there were only three minor differences between the unweighted and weighted metaanalysis (available from the authors) procedures. The first, mentioned above, was that the sample-size weighted average d was .40, while the unweighted average d was .25. However, these estimates, .31 and .24, respectively, were much closer when the Pritchard et al. (1973,n = 34,504) study was removed from the analysis. The second minor difference was that the weighted analysis indicated (Z = 1.65, p < .05) a significant difference between paper people and behavior observation, whereas the unweighted analysis indicated a smaller nonsignificant difference between these subgroups (Z = 1.18, ns). Again, however, this difference is not unexpected. Because the unweighted procedure produces a larger variance and, thus, wider confidence intervals, it is less likely than the weighted procedure to indicate differences between moderator subgroup distributions. The third difference, which was expected, was that the unweighted analysis produced effect sizes that were approximately one-third smaller than those obtained from the unweighted analysis. Finally, it is important to note that these small differences did not affect the basic conclusions. Except for the paper people versus behavior observation contrast, the weighted and unweighted meta-analysis techniques produced the same conclusions about the primary meta-analysis and moderators.

Discussion

The major impetus for this meta-analysis was the inconsistency of prior research investigating the effects of appraisal purpose on leniency of performance appraisal ratings. In contrast to previously inconsistent results, we found that performance appraisal ratings obtained for administrative purposes were nearly one-third of a standard deviation larger than those obtained for research or employee development purposes. Thus, these results clearly support performance appraisal models which predict that appraisal purpose influences appraisal leniency (e.g., DeCotiis & Petit, 1978; DeNisi, Cafferty, & Meglino, 1984; Taylor & Wherry, 1951).

However, the moderator analyses also showed that the PAP effect varied in consistent ways. Although the type of dependent measure (ratings vs. decisions), appraisal purpose (single vs. multiple), and rating format (GRS vs. FCS vs. BARS) did not moderate the PAP effect, we found that research setting, type of rater, type of appraisal stimulus, and direction/source of appraisal did moderate the PAP effect in a clearly identifiable and understandable pattern. Specifically, administrative ratings were more lenient than research ratings when managers (not students) in real organizations (not lab settings) rated real (not paper people or videotaped people) subordinates (not superiors). Simply put, administrative ratings will be much more lenient when those ratings are "for keeps" (Taylor & Wherry, 1951).

Because of the small number of studies in some of the moderator subgroups, we want to reiterate that many of these results should be viewed as tentative because they could change as additional data become available for future inclusion in a PAP meta-analysis. A case in point, is our reanalysis of Murphy et al.'s (1986) finding that the PAP effect was larger for paper people than for behavior observation. However, we found just the opposite, namely, that the PAP effect was much larger for behavior observation than for paper people. This difference is most likely due to the fact that Murphy et al.'s meta-analysis contrasted 2 paper people studies with 7 behavior observation studies, whereas our study contrasted 4 paper people studies with 16 behavior observation studies. Obviously, some of our moderator subgroup results could change in the same way as new data become available. However, we also believe that our results merit an acceptable degree of confidence. Our use of the more conservative unweighted meta-analytic procedure means that our results are more stable, and thus less likely to change as new data become available. Finally, like the results from individual studies, meta-analytic results cannot be fully understood until they are placed in a larger nomological net. Here, we have even more confidence in our moderator results because they not only fit so well together but are also consistent with prior research documenting the pervasiveness of leniency in organizations (Bretz et al., 1992; Kane, Bernardin, Villanova, & Peyrefitte, 1995; Longenecker et al., 1987). The pattern of results obtained in this meta-analysis clearly indicates that administrative ratings are more lenient than research ratings, and are even more lenient when managers in ongoing organizations evaluate the performance of actual subordinates.

Although our results clearly indicate that the PAP effect is sizeable in organizational settings, there is little in them that is positive for researchers or practitioners. For personnel researchers, the implications are straightforward: Avoid using administrative ratings or at the very least use them with extreme caution. Not only are administrative ratings artificially high, they are also likely to suffer from systematic range restriction and serious criterion contamination. This means that whenever possible, researchers should expend the additional time and money to obtain new appraisal data that are collected just for "research purposes."

By contrast, it is less clear what actions personnel practitioners should take to deal with the PAP effect. Ostensibly collecting data for "research purposes," and then turning around and using those same data to make administrative decisions is a one-time solution at best and is at worst dishonest. One place to start would be to decrease the discomfort that raters and ratees have with the appraisal process (Bernardin & Orban, 1990; Villanova, Bernardin, Dahmus, & Sims, 1993). For example, encouraging raters to provide feedback at frequent, regular intervals throughout the appraisal period might reduce rater and ratee discomfort. Another potential solution would be to increase raters' motivation to be accurate. Ilgen and Knowlton (1980) and Larson (1984) argue that the problem is not getting managers to recognize poor performance, but rather getting them to rate poor performance accurately. Research by Klimoski and Inks (1990) and Mero and Motowidlo (1995) suggests that holding raters accountable for the ratings they provide to their subordinates may be another way to reduce leniency. Unfortunately, there is little research on how to reliably accomplish any of these goals in field settings with "real" managers and subordinates. Thus, all of these potential solutions would appear to be fruitful.

In conclusion, the results of this study make it clear that the central question for future research is not whether there are serious problems with appraisals collected for administrative purposes, but rather, what should and can be done to reduce those problems. Regardless of the potential solutions researchers may choose to address these problems, the best indication of progress will be a smaller difference between appraisals conducted for administrative purposes and appraisals conducted for research purposes.

We thank our colleagues for helping us identify studies for possible inclusion in this meta-analysis, and for willingly furnishing us with information essential for computing effect sizes. Without their help, our analysis would have been restricted to fewer than 12 studies and an overall sample size of just 2,000.

An abbreviated version of this paper was awarded the 1997 SHRM Research Award, presented in memory of Dale Yoder, Ph.D. and Herbert G. Heneman, Jr., Ph.D.

[1] When calculating the fail-safe N, setting the critical value to 0 results in division by zero in the fail-safe N formula. Thus, we set the critical value to -.01 instead.

TABLE 1 Meta-Analysis Results

Legend for Chart:

A - Analysis
B - N
C - k
D - Mean observed d +/- 95% conf. intervals
E - Mean corrected d +/- 95% cred. intervals
F - sigma^2
G - sigma$^{2,\ sub\ e}$
H - % sigma^2 due to sampling error
I - Z
J - FSN

A	B	C	D	E	F	G	H	I	J
Primary									
Unweighted	57,775	22	.12 </= .25 </= .38	-.46 </= .32 </= 1.11	.0978	.0015	2%		528
Moderators									
Laboratory	1,061	6	-.13 </= .07 </= .27	-.42 </= .09 </= .60	.0631	.0229	36%		36
vs.								1.94[*]	
Field	56,714	16							

	.17 </= .32 </= .47	-.37 </= .41 </= .1.20
	.0975	.0011
	1%	496
Student	3,955	14
	.00 </= .17 </= .34	-.57 </= .22 </= 1.01
	.1110	.0143
	13%	224
vs.		1.82[*]
Org. rarer	53,820	8
	.23 </= .39 </= .55	-.08 </= .50 </= 1.08
	.0532	.0006
	1%	304
Paper people	793	4
	-.17 </= .12 </= .41	-.51 </= .15 </= .82
	.0904	.0204
	23%	44
vs.		1.18
Live observation	56,714	16
	.17 </= .32 </= .47	-.37 </= .41 </= 1.20
	.0975	.0011
	1%	496
Paper people	793	4
	-.17 </= .12 </= .41	-.51 </= .15 </= 82
	.0904	.0204
	23%	44
vs.		.89
Video	268	2
	-.18 </= -.03	</= .12 -.04
	.0117	.0303
	100%	8
Live observation	56,714	16
	.17 </= .32 </= .47	-.37 </= .41 </= 1.20
	.0975	.0011
	1%	496
vs.		3.20[*]
Video	268	2
	-.18 </= -.03	</= .12 -.04
	.0117	.0303
	100%	8
Upward	3,033	8
	-.13 </= .07 </= .27	-.57 </= .09 </= .75
	.0793	.0106
	13%	48

vs.		1.82[*]
Downward	54,490	11
	.14 </= .31 </= .48	-.32 </= .40 </= 1.13
	.0829	.0008
	1%	330
Perf. rating	57,505	20
	.12 </= 2.25 </= .40	-.48 </= .34 </= 1.15
	.1058	.0014
	1%	500
vs.		.07
Decision	590	3
	.03 </= .25 </= .47	-.01 </= .32 </= .66
	.0381	.0207
	54%	72
Graphic	57,353	19
	.13 </= .26 </= .39	-.37 </= .34 </= 104
	.0792	.0013
	2%	475
vs.		.75
Forced	19,811	3
	-.15 </= .14 </= .43	-.45 </= .18 </= .82
	.0637	.0006
	1%	39
Graphic	57,353	19
	.13 </= .26 </= .39	-.37 </= .34 </= 104
	.0792	.0013
	2%	475
vs.		.56
BARS	360	2
	-.80 </= .02 </= .84	-1.43 </= .03 </= 1.48
	.3528	.0225
	6%	2
Forced	19,811	3
	-.15 </= .14 </= .43	-.45 </= .18 </= .82
	.0637	.0006
	1%	39
vs.		.27
BARS	360	2
	-.80 </= .02 </= .84	-1.43 </= .03 </= 1.48
	.3528	.0225
	6%	2
Single	55,657	15
	.06 </= .24 </= .42	-.57 </= .31 </= 1.19

	.1220	.0011
	0.9%	345
vs.		.91
Multiple	2,118	7
	.10 </= .28 </= .46	-.16 </= .36 </= .89
	.0563	.0134
	24%	189

<u>REFERENCES</u>

Studies indicated by an asterix were included in the meta-analysis.

* Aleamoni LM, Hexner PZ. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. Instructional Science, 9, 67-84.

* Beckner D, Highhouse S, Hazer JT. (1995). Peer-rater motivation: A field experiment of rating purpose and accountability effects on inflation and avoidance. Unpublished manuscript.

* Berkshire JR, Highland RW. (1953). Forced-choice performance rating--A methodological study. ***PERSONNEL PSYCHOLOGY***, 6, 355-378.

* Bernardin HJ. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin HJ. (1986). Subordinate appraisal: A valuable source of information for managers. Human Resource Management, 25, 421-439.

Bernardin HJ, Abbott J, Cooper D. (1985,August). The effects of appraisal purpose and rater training on rating characteristics. Paper presented at the 1985 Annual Academy of Management Meetings, San Diego.

* Bernardin HJ, Cooke DK. (1992). Effects of appraisal purpose on discriminability and accuracy of ratings. Psychological Reports, 70, 1211-1215.

* Bernardin HJ, Orban JA. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. Journal of Business and Psychology, 5, (2), 197-211.

Bernardin HJ, Villanova P. (1986). Performance appraisal. In Locke EA (Ed.), Generalizing from laboratory to field settings (pp. 43-62). Lexington, MA: Lexington.

Bretz RD, Milkovich GT, Read W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. Journal of Management, 18; 321-352.

Carson KP, Schriesheim CA, Kinicki AJ. (1990). The usefulness of the "fail-safe" statistic in meta-analysis. Educational and Psychological Measurement, 50, 233-243.

* Centra JE. (1976). The influence of different directions on student ratings of instruction. Journal of Educational Measurement, 13, 277-282.

DeCotiis T, Petit A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.

DeNisi AS, Cafferty TP, Meglino BM. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

* Driscoll LA, Goodwin WL. (1979). The effects of varying information about the use and disposition of results on university students' evaluations of faculty and courses. American Educational Research Journal, 16, 25-37.

* Farh JL, Cannella AA, Bedeian AG. (1991). The impact of purpose on rating quality and user acceptance. Group and Organizational Studies, 16, 367-386.

* Farh JL, Werbel JD. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. Journal of Applied Psychology, 71, 527-529.

Fisher CD. (1989). Current and recurrent challenges in HRM. Journal of Management, 15, 157-180.

Gatewood RD, Feild HS. (1994). Human resource selection. Ft. Worth, TX: Dryden.

Glass GV. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Greene L, Bernardin HJ, Abbott J. (1985). A comparison of rating formats after corrections for attenuation. Educational and Psychological Measurement, 45, 503-515.

* Gmelch WH, Glasman NS. (1977). The effects of purpose on student evaluation of college instructors. Educational Research Quarterly, 2, 45-55.

* Harris MM, Smith DE, Champagne D. (1995). A field study of performance appraisal purpose: Research versus administrative based ratings. Personnel Psychology. 48, 151-160.

Hollander EP. (1965). Validity of peer nominations in predicting a distant performance criterion. Journal of Applied Psychology, 49, 443-438.

Huffcutt AI, Arthur W Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. Journal of Applied Psychology, 79, 184-190.

Huffcutt AI, Arthur W Jr. (1995). Development of a new outlier statistic for meta-analytic data. Journal of Applied Psychology, 80, 327-334.

Hunter JE, Schmidt FL. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Beverly Hills, CA: Sage.

Ilgen DR, Barnes-Farrell JL, McKellin DB. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? Organizational Behavior and Human Decision Processes, 54, 321-368.

Ilgen DR, Favero JL. (1985). Limits in generalization from psychological research to performance appraisal processes. Academy of Management Review, 10, 311-321.

Ilgen DR, Knowlton WA. (1980). Performance attributional effects on feedback from superiors. Organizational Behavior and Human Performance, 25, 441-456.

* Jawahar IM. (1994). Purpose of appraisal revisited: An examination of the relationship between purpose and characteristics of performance ratings. Unpublished dissertation, Oklahoma State University.

Kane JS, Bernardin HJ, Villanova P, Peyrefitte J. (1995). Stability of rater leniency: Three studies. Academy of Management Journal, 34 (4), 1036-1051.

Klimoski R, Inks L. (1990). Accountability forces in performance appraisal. Organizational Behavior and Human Decision Processes, 45, 194-208.

Kraiger K, Ford JK. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

Landy FS, Farr JL. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Larson JR Jr. (1984). The performance feedback process: A preliminary model. Organizational Behavior and Human Performance, 33, 42-76.

Longenecker CO, Sims HP, Gioia DA. (1987). Behind the mask: The politics of employee appraisal. The Academy of Management Executive, 1, 183-193.

McIntyre RM, Smith DE, Hassett CE. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 149-156.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer SD. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of Applied Psychology, 79, 599-616.

* Meier RA, Feldhusen JF. (1979). Another look at Dr. Fox: Effect of stated purpose of evaluation, lecturer expressiveness, and density of lecture content on student ratings. Journal of Educational Psychology, 71, 339-345.

Mero NP, Motowidlo SJ. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. Journal of Applied Psychology, 80, 517-524.

* Murphy KR, Balzer WK, Kellam KL, Armstrong JG. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76 45-54.

Murphy KR, Cleveland JN. (1991). Performance appraisal: An organizational perspective. Boston: Allyn and Bacon.

Murphy KR, Herr BM, Lockhart MC, Maguire E. (1986). Evaluating the performance of paper people. Journal of Applied Psychology, 71, 654-661.

Osburn HG, Callender J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. Journal of Applied Psychology, 77, 115-122.

Pearlman K. (1979). The validity of tests used to select clerical personnel: A comprehensive summary and evaluation (TS-79-1). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB 80.102650).

* Pritchard RD, Peters LH, Harris AF. (1973). The effects of confidentiality on the distribution of naval performance appraisals. Prepared for Psychological Sciences Division of the Office of Naval Research (Contract No. N00014-67-A-0226-0018).

Rogers PD, McDaniel MA. (1994,August). Criterion purpose as a moderator of employment test validity. Paper presented at the Annual American Psychological Association Convention, Los Angeles, CA.

* Sharon A, Bartlett CJ. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 22, 251-263.

* Shore TH, Adams JS, Tashchian A. (1995,August). Effects of self-appraisal information, appraisal purpose, and feedback target on performance appraisal ratings. Paper presented at the 1995 Annual Academy of Management Meetings, Vancouver, Canada.

* Taylor EK, Wherry RJ. (1951). A study of leniency in two rating systems. Personnel Psychology, 4, 39-47.

Tett RP, Jackson DN, Rothstein M. (1991). Personality measures as predictors of job performance: A meta-analytic review. Personnel Psychology, 44, 703-742.

* Veres JG, Feild HS, Boyles WR. (1983). Administrative versus research performance ratings: An empirical test of rating data quality. Public Personnel Management Journal, 12, 290-298.

Villanova P, Bernardin J, Dahmus SA, Sims RL. (1993). Rater leniency and performance appraisal discomfort. Educational and Psychological Measurement, 53, 789-799.

* Waldman DA, Thornton GC III. (1988). A field study of rating conditions and leniency in performance appraisal. Psychological Reports, 63, 835-840.

Williams CR, Livingstone LP. (1994). Another look at the relationship between performance and voluntary turnover. Academy of Management Journal, 37, 269-298.

* Williams KJ, DeNisi AS, Blencoe AG, Cafferty TP. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. Organizational Behavior and Human Decision Processes, 35, 314-339.

Zedeck S, Cascio WF. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.