Pulakos, E. D. (2009). *Performance management: A new approach for driving business results*. Malden, MA: Wiley-Blackwell. doi:10.1002/9781444308747

Selvarajan, T. T., & Cloninger, P. A. (2012). Can performance appraisals motivate employees to improve performance? A Mexican study. *The International Journal of Human Resource Management, 23*(15), 3063–3084. doi:10.1080/09585192.2011.637069

Steelman, L. A., & Rutkowski, K. A. (2004). Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology, 19*(1), 6–18. doi:10.1108/02683940410520637

Taylor, M. S., Tracy, K. B., Renard, M. K., Harrison, J. K., & Carroll, S. J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly, 40*, 495–523. doi:10.2307/2393795

Time To Change the Bathwater: Correcting Misconceptions About Performance Ratings

C. Allen Gorman
East Tennessee State University and GCG Solutions

Christopher J. L. Cunningham
The University of Tennessee at Chattanooga and Logi-Serve

Shawn M. Bergman
Appalachian State University and B&F Associates

John P. Meriac
University of Missouri–St. Louis

Recent commentary has suggested that performance management (PM) is fundamentally "broken," with negative feelings from managers and employees toward the process at an all-time high (Pulakos, Hanson, Arad, & Moye, 2015; Pulakos & O'Leary, 2011). In response, some high-profile organizations have decided to eliminate performance ratings altogether as a solution to the growing disenchantment. Adler et al. (2016) offer arguments both in support of and against eliminating performance ratings in organizations. Although both sides of the debate in the focal article make some strong arguments both for and against utilizing performance ratings in organizations, we believe there continue to be misunderstandings, mischaracterizations,

C. Allen Gorman, Department of Management and Marketing, East Tennessee State University, and GCG Solutions, Limestone, Tennessee; Christopher J. L. Cunningham, Department of Psychology, The University of Tennessee at Chattanooga, and Logi-Serve, Farmington Hills, Michigan; Shawn M. Bergman, Department of Psychology, Appalachian State University, and B&F Associates, Boone, North Carolina; John P. Meriac, Department of Psychology, University of Missouri–St. Louis.

Correspondence concerning this article should be addressed to C. Allen Gorman, Department of Management and Marketing, East Tennessee State University, 128 Sam Wilson Hall, P.O. Box 70625, Johnson City, TN 37614. E-mail: gormanc@etsu.edu

and misinformation with respect to some of the measurement issues in PM. We offer the following commentary not to persuade readers to adopt one particular side over another but as a call to critically reconsider and reevaluate some of the assumptions underlying measurement issues in PM and to dispel some of the pervasive beliefs throughout the performance rating literature.

Measurement Issues in Performance Ratings

As noted by Adler et al., measurement issues have been pervasive in the PM literature since its inception. Understandably, some scholars have argued that the overwhelming focus on measurement issues in the academic literature has rendered PM research essentially useless to PM practitioners (DeNisi & Pritchard, 2006; Fletcher, 2001). Unfortunately, however, PM critics continue to rely on overgeneralized conclusions regarding measurement issues that are based on outdated, unsupported, and misinterpreted research, and these unsubstantiated generalizations have become accepted as truth in our science. Below, we separate fact from fiction with respect to three key areas in PM: (a) rating formats, (b) rater training, and (c) rater (dis)agreement and the reliability of PM ratings.

Rating Formats

The most frequently cited article put forth as "evidence" of the failure of rating format interventions is Landy and Farr's (1980) watershed article in which they famously called for a moratorium on rating format design research and concluded that interventions designed to improve performance rating formats were, at best, minimally successful. However, what is less often communicated is that Landy and Farr's (1980) conclusions regarding the lack of usefulness of rating format research were based almost entirely on the presence of psychometric "*errors*" in performance ratings (DeNisi, 1996). As Colquitt, Murphy, and Ollander-Kane (as cited in Adler et al.) aptly note in their own criticism of PM ratings, psychometric "errors" only represent one type of rating property, and they have also been repeatedly criticized as poor indicators of rating quality (Balzer & Sulsky, 1992; Fisicaro, 1988; Murphy, 2008; Murphy & Balzer, 1989; Murphy, Jako, & Anhalt, 1993; Nathan & Tippins, 1990). Consequently, the tenuous evidence base regarding rating errors calls into question the conclusions of an entire body of rating format research dismissed by Landy and Farr (1980).

More recently, other psychometric indices have been used to evaluate rating quality and have provided a much clearer picture regarding the value of rating formats. Specifically, research that has used more appropriate indices of rating quality, such as predictor validity and rater reactions, has actually yielded favorable results (Bartram, 2007; Benson, Buckley, & Hall,

1988; Borman et al., 2001; Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Roch, Sternburgh, & Caputo, 2007; Tziner, 1984; Wagner & Goffin, 1997). Hoffman, Gorman, Blair, Meriac, Overstreet, and Atchley (2012), for example, found that a new rating format they termed "frame-of-reference (FOR) scales" resulted in an improved factor structure compared with a standard multisource rating instrument and rating accuracy levels comparable with those from a FOR training program. Moreover, recognizing recent advances in technology, the expanding criterion domain, and the creation of new forms of work, Landy (2010) himself officially lifted the 30-year moratorium on rating format design research. Thus, we suggest that rumors of the death of performance rating formats have been greatly exaggerated.

Rater Training

Colquitt et al. (in Adler et al.) state that rater training has been unsuccessful in substantially improving ratings in organizations. We should point out that we agree wholeheartedly that rater *error* training has been a tremendous disappointment as an intervention for improving ratings. Research has shown that although rater error training results in fewer leniency and halo errors, it inadvertently lowers levels of rating accuracy (Bernardin & Pence, 1980; Borman, 1979; Landy & Farr, 1980), and rater error training essentially creates a meaningless redistribution of ratings and is practically useless in terms of improving rating quality (Borman, 1979; Smith, 1986). As noted above, though, this is not surprising given the inherent limitations of psychometric "errors" as indicators of rating quality.

However, we disagree with the assertion made in the focal article that behavior-based rater training (e.g., FOR training) is a disappointing rating intervention. Although there is a relative lack of evidence that rater training improves actual ratings in field settings (cf. Noonan & Sulsky, 2001), for several reasons, we suggest that rater training is an understudied rating intervention that has a great deal of potential for improving ratings in organizations. For example, in a popular industrial–organizational (I-O) psychology textbook, Levy (2010) noted that rater training has become more common in organizations such as the Tennessee Valley Authority, JP Morgan Chase, Lucent Technologies, and AT&T.

Moreover, meta-analytic reviews have found impressive effect sizes for the impact of FOR training on improving rating quality ($d = 0.83$ in Woehr & Huffcutt, 1994, and $d = 0.50$ in Roch, Woehr, Mishra, & Kieszczynska, 2011). Finally, in a recent exploratory survey of for-profit companies, Gorman, Meriac, Ray, and Roddy (2015) found that 61% of the 101 organizations surveyed reported that they use a behavior-based approach (such as FOR training) to train raters, and companies that utilized behavior-focused rater training programs generated higher revenue than those that provide rater

error training or no training at all. More research is clearly needed on this topic, but claiming that rater training is ineffective is premature. Thus, we agree that rater error training should not be considered a viable rater training option but hasten to note that a lack of research in organizational settings does not automatically equate to a failed intervention in the case of FOR and other behavior-based training interventions.

Rater (Dis)agreement and the Reliability of Performance Ratings

Colquitt et al. (in Adler et al.) also suggest that disagreement among raters in PM is a major problem that supports the abandonment of ratings altogether, and they support this assertion using the ubiquitous .52 interrater reliability estimate often cited as a general estimate of the reliability of performance ratings (Viswesvaran, Ones, & Schmidt, 1996). There are at least two problems with this argument: (a) disagreement among raters may reflect true variance, instead of error, and (b) .52 is potentially an underestimate of the reliability of performance ratings. We address each of these issues below.

We agree that "raters do not show the level of agreement one might expect from . . . two different forms of the same paper-and-pencil test" (Adler et al., p. 225). The question of why one would or should expect similar levels of agreement between multiple raters, however, must be asked. Would we actually want to see perfect agreement among multiple raters or sources? From a classical test theory perspective, for example, differences between a true score and an actual score on a paper-and-pencil test are considered error (Nunnally & Bernstein, 1994). But how do we know what the true score is when it comes to job performance?

Research evidence suggests that rating source disagreement may be due more to differences in the performance constructs being rated than differences between sources (Woehr, Sheehan, & Bennett, 2005). Moreover, proponents of the ecological validity perspective have long recognized that performance ratings are based on functionally and socially adaptive judgments that likely represent true sources of variance rather than error (Hoffman, Lance, Bynum, & Gentry, 2010; Lance, Hoffman, Gentry, & Baranik, 2008; Lance & Woehr, 1989). In fact, as Hoffman et al. (2010) aptly noted, why would we gather performance ratings from different sources if we expected them to all completely agree? Thus, we suggest that the assumption that rater disagreement is indicative of a problem with PM is based on a faulty premise.

We completely agree that a reliability estimate around .50 is "hardly the level one would expect if ratings were in fact good measures of the performance of the ratees" (Adler et al., p. 225). However, the oft-cited estimate of .52 is likely a biased underestimate of the reliability of performance (LeBreton, Scherer, & James, 2014). The argument over whether inter- or intrarater correlations are more appropriate measures of reliability (Murphy &

DeShon, 2000; Ones, Viswesvaran, & Schmidt, 2008; Schmidt, Viswesvaran, & Ones, 2000) notwithstanding, there are several other reasons to believe this to be a downwardly biased estimate. First, job performance is a dynamic and multidimensional criterion (Austin & Villanova, 1992), and low reliability is often indicative of a dynamic and multidimensional construct (Nunnally & Bernstein, 1994). Second, it is well-known that performance ratings are skewed to the positive end of the distribution. For example, when using a seven-point scale, 80% of ratings are often a 6 or 7 (Murphy & Cleveland, 1995). This problem becomes even more pronounced if the ratings are used for administrative decisions (Jawahar & Williams, 1997). Thus, restriction of range severely attenuates the observed reliability in job performance ratings (LeBreton, Burgess, Kaiser, Atchley, & James, 2003).

Finally, job performance ratings are rarely modeled in the extant research with training or format as a factor, but research has demonstrated that interventions such as rater training can improve reliability estimates of performance ratings. Lievens (2001), for example, found that a schema-driven training condition produced interrater reliability estimates of at least .80 or greater in a sample of both students and managers across three performance dimensions. In addition, using variance components analysis, Gorman and Jackson (2012) reported that rater idiosyncrasies accounted for a large amount of variance in a control training condition but a negligible amount of variance in a FOR training condition. Thus, ratings from trained raters are much less influenced by idiosyncratic error than ratings made by untrained raters. Hence, in situations where raters are left to their own devices without proper training and well-developed rating instruments, the reliability of job performance ratings may actually be much lower.

Some Additional Considerations

The above issues aside, we also agree with many of the other points made by the authors, including the notion that the overall process must be considered, including the consequences of ratings. Dissatisfaction with the ultimate outcomes of management decisions (e.g., raises, promotions, or terminations) would simply shift the criticism from performance ratings to other elements of the process. Performance judgments and comparisons between employees will inevitably be made, whether we call them "ratings" or something else (Meriac, Gorman, & Macan, 2015). In addition, the social context of PM is, and should remain, an important consideration in the PM process. Without proper management support, accountability (London, Smither, & Adsit, 1997), and an environment supporting the effective use of performance ratings and feedback (e.g., Steelman, Levy, & Snell, 2004), even highly reliable ratings are unlikely to work as expected. However, the abandonment of ratings is unlikely to facilitate effective PM.

Conclusion

In this commentary, we suggested that, as evidenced in the focal article, there are several myths and urban legends surrounding the measurement of performance ratings that have been perpetuated and passed down in the PM literature. Specifically, we argued that (a) premature conclusions have been reached regarding performance rating formats based on outdated research using improper criteria, (b) behavior-focused rater training programs hold great promise as interventions to improve the quality of ratings in organizations but deserve much more research attention in field settings, and (c) rater agreement is an unrealistic goal, but nevertheless, estimates around .50 are likely downward estimates of the reliability of job performance ratings. We further urge readers to consider the research evidence critically for themselves before accepting foregone conclusions regarding the measurement and ultimate value of performance ratings. As I-O scientists and practitioners, a shared understanding of the measurement issues involved in PM must be a priority before we can begin a dialogue on the merits of abandoning a process fundamental to many of our human resource activities.

References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(2), 219–252.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*(6), 836–874.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975–985.

Bartram, D. (2007). Increasing validity with forced-choice criterion measure formats. *International Journal of Selection and Assessment, 15*, 263–272.

Benson, P. G., Buckley, M. R., & Hall, S. (1988). The impact of rating scale format on rater accuracy: An evaluation of the mixed standard scale. *Journal of Management, 14*, 415–423.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60–66.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410–421.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965–973.

DeNisi, A. S. (1996). *Cognitive approach to performance appraisal: A program of research*. New York, NY: Taylor Francis.

DeNisi, A. S., & Pritchard, R. D. (2006). Performance appraisal, performance management, and improving individual performance: A motivational framework. *Management and Organization Review, 2*, 253–277.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology, 73*, 239–244.

Fletcher, C. (2001). Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology, 74*, 473–487.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology, 11*, 23–33.

Gorman, C. A., & Jackson, D. J. R. (2012, April). A generalizability theory approach to understanding frame-of-reference rater training effectiveness. In A. Gibbons (Chair), *Inside assessment centers: New insights about assessors, dimensions, and exercises*. Symposium presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Gorman, C. A., Meriac, J. P., Ray, J. L., & Roddy, T. W. (2015). Current trends in rater training: A survey of rater training programs in American organizations. In B. J. O'Leary, B. L. Weathington, C. J. L. Cunningham, & M. D. Biderman (Eds.), *Trends in training* (pp. 1–23). Newcastle upon Tyne, UK: Cambridge Scholars.

Hoffman, B. J., Gorman, C. A., Blair, C. A., Meriac, J. P., Overstreet, B. L., & Atchley, E. K. (2012). Evidence for the effectiveness of an alternative multisource performance rating methodology. *Personnel Psychology, 65*, 531–563.

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*(1), 119–151.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*(4), 905–926.

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*(4), 223–232.

Lance, C. E., & Woehr, D. J. (1989). The validity of performance judgments: Normative accuracy model versus ecological perspectives. In D. F. Ray (Ed.), *Southern Management Association Proceedings* (pp. 115–117). Oxford, MS: Southern Management Association.

Landy, F. J. (2010). Performance ratings: Then and now. In J. L. Outz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 227–248). New York, NY: Routledge.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80–128.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 478–500.

Levy, P. E. (2010). *Industrial/organizational psychology* (3rd ed.). New York, NY: Worth.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255–264.

London, M., Smither, J. W., & Adsit, D. J. (1997). Accountability: The Achilles' heel of multisource feedback. *Group & Organization Management, 22*(2), 162–184.

Meriac, J. P., Gorman, C. A., & Macan, T. (2015). Seeing the forest but missing the trees: The role of judgments in performance management. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 102–108.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Research and Practice, 1*, 148–160.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619–624.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., & DeShon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*(4), 913–924.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218–225.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology, 75*, 290–296.

Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*, 3–26.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology, 1*(2), 174–179.

Pulakos, E. D., Hanson, R. M., Arad, S., & Moye, N. (2015). Performance management can be fixed: An on-the-job experiential learning approach for complex behavior change. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 51–76.

Pulakos, E. D., & O'Leary, R. S. (2011). Why is performance management broken? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 146–164.

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs. relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*, 302–316.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2011). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*, 370–395.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901–912.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*, 22–40.

Steelman, L. A., Levy, P. E., & Snell, A. F. (2004). The feedback environment scale: Construct definition, measurement, and validation. *Educational and Psychological Measurement*, *64*(1), 165–184.

Tziner, A. (1984). A fairer examination of rating scales when used for performance appraisal in a real organizational setting. *Journal of Organizational Behavior, 5*, 103–112.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574.

Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes, 70*, 95–103.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.
Woehr, D. J., Sheehan, M. K., & Bennett, W., Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait–multirater approach. *Journal of Applied Psychology*, *90*(3), 592–600.

Beyond Performance Ratings: The Long Road to Effective Performance Management

Robert L. Cardy
University of Texas at San Antonio

Deeksha Munjal
University of Texas at San Antonio

Performance evaluation has long been a source of dissatisfaction for practitioners and a focus of research for scholars. The current call for the elimination of performance ratings is not new. This commentary considers the quality perspective as a historical context in which performance ratings were, at best, considered a misguided management tool. Although the current debate doesn't seem to be philosophically based, it may be useful to recognize that serious questions regarding performance ratings have come up before. Potential measurement problems with performance ratings are considered. It is concluded that performance ratings are not the major problem for performance management. Possible sources of problems with performance management are considered. Directions for improvement are discussed.

Giving and receiving performance ratings are probably seldom viewed as fun and relaxing activities by either party. Even top performers can be anxious about how they will be assessed, and the best managers can't be certain how their evaluations will be received. To put a number on it with a performance rating seems to amplify anxiety and concerns over equity. As reflected in the focal article (Adler et al., 2016), there are both pros and cons to performance ratings. Recently, attention has been given to organizations that are choosing to eliminate performance ratings, for example Adobe and General Electric (Garr, 2013; Pulakos, Mueller-Hanson, Arad, & Moye, 2015). Arguments about the downsides of performance ratings are

Robert L. Cardy, Department of Management, University of Texas at San Antonio; Deeksha Munjal, Department of Management, University of Texas at San Antonio.

Correspondence concerning this article should be addressed to Robert L. Cardy, Department of Management, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0634. E-mail: robert.cardy@utsa.edu