The Relationship Between the Number of Raters and the Validity of Performance Ratings

Matt C. Howard

Industrial and Organizational Psychology / Volume 9 / Issue 02 / June 2016, pp 361 - 367
DOI: 10.1017/iop.2016.26, Published online: 04 July 2016

McKay, P. F., & McDaniel, M. A. (2006). A reexamination of Black–White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91*, 538–554.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable. *Journal of Applied Psychology, 75*, 175–184.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive ability: A European community meta-analysis. *Personnel Psychology, 56*, 537–605.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology, 52*, 445–464.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Taylor, P. J., Russ-Eft, D., & Taylor, H. (2009). Transfer of management training from alternative perspectives. *Journal of Applied Psychology, 94*, 104–112.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: Wiley.

The Relationship Between the Number of Raters and the Validity of Performance Ratings

Matt C. Howard
University of South Alabama and Pennsylvania State University

In the focal article "Getting Rid of Performance Ratings: Genius or Folly? A Debate," two groups of authors argued the merits of performance ratings (Adler et al., 2016). Despite varied views, both sides noted the importance of including multiple raters to obtain more accurate performance ratings. As the pro side noted, "if ratings can be pooled across many similarly situated raters, it should be possible to obtain quite reliable assessments" (Adler et al., p. 236). Even the con side noted, "In theory, it is possible to obtain ratings from multiple raters and pool them to eliminate some types of

Matt C. Howard, Mitchell College of Business, Department of Management, University of South Alabama, and Department of Psychology, Pennsylvania State University.

Thanks to Rick Jacobs and Alex McKay for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Matt C. Howard, Mitchell College of Business, Department of Management, University of South Alabama, 5811 USA Drive South, Room 346, Mobile, AL 36688-0002. E-mail: mhoward@southalabama.edu

interrater agreement" (Adler et al., p. 225), although this side was certainly less optimistic about the merits of multiple raters. In the broader industrial–organizational psychology literature, authors have repeatedly heralded the benefits of adding additional raters for performance ratings, some even treating it as a panacea for inaccurate ratings. Although these authors extol the virtues of multiple raters, an important question is often omitted from relevant discussions of performance ratings: To what extent do additional raters actually improve performance ratings? Does adding an additional rater double the validity of performance ratings? Does an additional rater increase the validity of performance ratings by a constant value? Or is the answer something else altogether?

It is possible, if not probable, that many researchers and practitioners do not exactly know the benefits of adding additional raters, and some authors may be blindly overemphasizing the importance of multiple raters. For this reason, in the following, I provide quantitative inferences about the actual impact of adding additional raters on the validity of performance ratings. In doing so, I also provide useful tables that future researchers and practitioners can use to determine whether adding additional raters to their performance rating systems would result in benefits that might outweigh the costs. To conclude, I discuss four primary inferences about the relationship of adding additional raters and the validity of performance ratings. From achieving these objectives, I provide a more accurate view of the benefits obtained from adding additional raters to a rating system, thereby allowing researchers and practitioners to more accurately determine whether getting rid of performance ratings is, in fact, genius or folly.

Determining the Impact of Adding Additional Raters

To determine the impact of adding additional raters to a rating system, a classic psychometric formula can be applied. Ghiselli (1964) created an equation to determine the validity of a test as the number of test items increases, but the same equation can compute the validity of a composite rater as the number of raters increases (Hogarth, 1978; Tsujimoto, Hamilton, & Berger, 1990). When applying the formula for the latter purpose, it is as follows:

$$\frac{\sqrt{m}E\left(r_{jx}\right)}{\sqrt{1+(m-1)\,E\left(r_{ij}\right)}} = E\left(r_{mjx}\right) \quad \textbf{(1)}$$

Whereas m is the number of raters, r_{jx} is the average correlation between each rater and true performance, r_{ij} is the average correlation between the raters, and r_{mjx} is the correlation between the average of the raters (composite rater) and true performance. In this article, the r_{mjx} is labeled the validity coefficient, and it represents the accuracy of a rating system. When assuming

Table 1. Correlation Between Average Observed Score and True Score (Validity Coefficient) Assuming No Systematic Error

Average corr. of each rater w/true performance	One rater	Two raters	Three raters	Four raters	Five raters	Six raters	Seven raters	Eight raters	Nine raters
.10	.10	.14	.17	.20	.22	.24	.26	.27	.28
.20	.20	.28	.33	.38	.42	.45	.48	.50	.52
.30	.30	.41	.48	.53	.58	.61	.64	.66	.69
.40	.40	.53	.60	.66	.70	.73	.76	.78	.79
.50	.50	.63	.71	.76	.79	.82	.84	.85	.87
.60	.60	.73	.79	.83	.86	.88	.89	.90	.91
.70	.70	.81	.86	.89	.91	.92	.93	.94	.95
.80	.80	.88	.92	.94	.95	.96	.96	.97	.97
.90	.90	.95	.96	.97	.98	.98	.98	.99	.99

	Average correlation change for			
Increasing the rater and true score correlation for each rater	.10 to .20	.19	.09	1 to 2
	.20 to .30	.15	.05	2 to 3
	.30 to .40	.12	.04	3 to 4
	.40 to .50	.09	.03	4 to 5
	.50 to .60	.07	.02	5 to 6
	.60 to .70	.06	.02	6 to 7
	.70 to .80	.05	.01	7 to 8
	.80 to .90	.04	.01	8 to 9

Increasing the number of raters

that all shared variance between raters is through true performance (i.e., no systematic error), the average correlation between raters is the indirect effect via true performance. For example, if the correlation between each of two raters and true performance is .20, and no other shared variance is assumed, then the correlation between the two raters is .04 (.20∗.20). Thus, when assuming that all shared variance between raters is via true performance, the formula can be rewritten as

$$\frac{\sqrt{m}E\left(r_{jx}\right)}{\sqrt{1+(m-1)\,E\left(r_{jx}{}^{2}\right)}} = E\left(r_{mjx}\right) \tag{2}$$

The only difference between Formulas 1 and 2 is that r_{ij} is replaced by $r_{jx}{}^{2}$, reflecting that the average correlation between raters is solely through the indirect effect of true performance. Using this formula, we can determine the validity coefficient when the rating accuracy and number of raters varies (see Table 1). Before analyzing such results and drawing

inferences about adding additional raters, however, another important factor should be considered.

The results in Table 1 assume that raters are independent and no systematic error exists, but many authors have demonstrated that this assumption is rarely held in practice (Murphy, Cleveland, & Mohler, 2001; Ones, Viswesvaran, & Schmidt, 2008; Viswesvaran, Ones, & Schmidt, 1996). Raters often demonstrate systematic variance that is independent of true performance, especially when ratings are provided from the same organizational level (i.e., peer, subordinate, supervisor, etc.), and this systematic variance decreases the validity coefficient. Ignoring this variance would depict an inaccurate view of performance ratings.

In a noteworthy study, Hoffman, Lance, Bynum, and Gentry (2010) demonstrated that rater source effects account for approximately 22% of the shared variance between raters. Taking this figure, we can assume that the correlation between raters that is solely due to rater source effects is .469 ($\sqrt{.22}$), and the total correlation between raters is an additive function of these source effects and the indirect effect of true performance. Given this, the prior formula can be modified to determine the validity coefficient when accounting for rater source effects. The modified formula is as follows:

$$\frac{\sqrt{m}E\left(r_{jx}\right)}{\sqrt{1+(m-1)\,E\left(.469+\ r_{jx}{}^{2}\right)}}=E\left(r_{mjx}\right) \tag{3}$$

The only difference between Formulas 2 and 3 is that the rater source effect (.469) is included in the calculation of the average correlation between raters. Using this formula, we can once again determine the validity coefficient when the rating accuracy and number of raters varies—this time accounting for rater source effects. Table 2 includes these validity coefficients. Four primary inferences should be taken from these results.

First, the impact of adding raters may be smaller than many would expect. As mentioned, many researchers and practitioners believe that adding raters is a panacea for inaccurate ratings. At best, however, adding an additional rater only increases the validity coefficient by .06. On average, adding an additional rater only improved the validity coefficient by .01. Although explaining any additional variance in performance is valuable, these results are almost assuredly smaller than many expectations, and additional raters are almost certainly not a panacea for inaccurate ratings.

Second, the benefits of adding raters decrease as the number of raters increases. For example, when increasing the number of raters from one to two when the average correlation between each of the raters and true performance is .50, the validity coefficient increases from .50 to .56; however, when increasing the number of raters from eight to nine, the validity coefficient

Table 2. Correlation Between Average Observed Score and True Score (Validity Coefficient) Assuming Rater Source Effects

Average corr. of each rater w/true performance	One rater	Two raters	Three raters	Four raters	Five raters	Six raters	Seven raters	Eight raters	Nine raters
.10	.10	.12	.12	.13	.13	.13	.13	.14	.14
.20	.20	.23	.25	.25	.26	.26	.27	.27	.27
.30	.30	.34	.36	.38	.38	.39	.39	.39	.40
.40	.40	.45	.48	.49	.50	.50	.51	.51	.51
.50	.50	.56	.58	.60	.61	.61	.62	.62	.62
.60	.60	.66	.68	.70	.70	.71	.71	.72	.72
.70	.70	.75	.77	.78	.79	.80	.80	.80	.80
.80	.80	.84	.86	.86	.87	.87	.87	.88	.88
.90	.90	.92	.93	.94	.94	.94	.94	.94	.94

	Average correlation change for			
	.10 to .20	.12	.04	1 to 2
	.20 to .30	.12	.02	2 to 3
	.30 to .40	.12	.01	3 to 4
Increasing the rater and true score correlation for each rater	.40 to .50	.11	.01	4 to 5
	.50 to .60	.10	.00	5 to 6
	.60 to .70	.09	.00	6 to 7
	.70 to .80	.08	.00	7 to 8
	.80 to .90	.07	.00	8 to 9

Increasing the number of raters

remains virtually constant at .62. When inspecting Table 2, it appears the benefits of adding raters begin to bottom out after three. Therefore, diminishing returns are received when adding raters, and authors should strongly consider whether the small benefits of including more than two or three raters outweigh the costs.

Third, the benefit of additional raters is decreased when rating accuracy is either high or low. For example, when the correlation between each of the raters and true performance is .10, increasing the number of raters from one to two only increases the validity coefficient from .10 to .12. When the correlation is .90, increasing the number of raters from one to two only increases the validity coefficient from .90 to .92. When the correlation is .50, however, increasing the number of raters from one to two increases the validity coefficient from .50 to .56. It appears that, when ratings are inaccurate, additional raters are unable to provide additional meaningful information, and when ratings are extremely accurate, additional raters do not provide additional novel information. Thus, researchers and practitioners should

heavily consider whether adding additional raters sufficiently improves performance ratings, such as through analyzing the accuracy of their current ratings, or whether they should allocate their resources toward other aspects of their rating systems.

Fourth, the impact of adding raters is smaller than improving measures. At best, improving the correlation of each rater with true performance by .10 results in a .16 increase in the validity coefficient. On average, improving the correlation of each rater with true performance by .10 results in a .10 increase to the validity coefficient. Although it is largely impossible to precisely increase each rater's correlation with true performance by .10, these results nevertheless show that improving rating accuracy is as effective as expectations. In almost any circumstance, researchers and practitioners receive their expected benefits from improving rating accuracy, which is not the case with increasing the number of raters. Once again, it may be more beneficial to allocate resources toward developing better measures and rating systems to improve performance ratings, rather than adding additional raters.

Together, these four inferences suggest that adding additional raters to a rating system may not actually provide noteworthy improvements to rating accuracy, contrary to common thought on the topic. These inferences do not explicitly show that adding getting rid of performance ratings is genius, but assuming that adding additional raters is a solution to this debate is a folly.

Conclusion

The goal of the current article was to reevaluate common thought about adding additional raters to performance rating systems. As the results of Ghiselli's (1964) classic formula demonstrated, adding additional raters may not provide as much of a benefit as commonly believed. Further, adding additional raters beyond two or three provides marginal benefits to performance ratings, and extremely inaccurate or accurate ratings systems likewise receive few benefits from adding additional raters. Nevertheless, improving the accuracy of ratings almost always provides the expected benefits. Together, whereas many authors laud the importance of multiple raters, the results of this commentary showed that adding raters might only provide marginal benefits to the validity of a rating system. Although these results may not argue that removing performance ratings is genius, they certainly demonstrate that even the most lauded strengths of performance ratings have serious concerns.

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.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York, NY: McGraw-Hill.

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119–151.
Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*, 40–46.
Murphy, K. R., Cleveland, J. N., & Mohler, C. (2001). Reliability, validity, and meaningfulness of multisource ratings. In D. Bracken, C. Timmreck, & A. Church (Eds.), *Handbook of multisource feedback* (pp. 130–148). San Francisco, CA: Jossey-Bass.
Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 174–179.
Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*(4), 432–437.
Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Getting Rid of Performance Ratings

Melvin Sorcher
Westport, Connecticut

And the beat goes on. The same questions about performance appraisals keep popping up despite significant changes in work environments, contexts, and expectations over the past 2 or 3 decades (Adler et al., 2016). Even after decades of research and debate about the benefits and construction of performance appraisal ratings, no closure is reached or "best practice" identified. The application of ratings differs widely among companies, and the criteria, scaling, and language are tweaked by virtually every human resources group. In my experience, each organization believes that its performance criteria are unique. This should not be surprising because supervisors who observe and rate human performance do not react like a school of fish. What most human resources managers miss is that each of the supervisors who apply ratings are also unique, and they do not perceive performance consistently—except, perhaps, for the most exceptional and the poorest performers. Methods of quantifying or behaviorally slotting employee performance along a variety of dimensions to arrive at some accurate scaled rating have not made employees happy and are a painful chore for most supervisors.

There are some work situations where the key tasks are exceptionally intense and require extraordinary focus, like landing an airplane on an aircraft

Melvin Sorcher, Westport, Connecticut.
Correspondence concerning this article should be addressed to Melvin Sorcher, 31 Dogwood Lane, Westport, CT 06880. E-mail: melsorcher@optonline.net