Original Article

Peer-Rated Organizational Citizenship Behavior

Does Familiarity Improve Rating Quality?

Kevin Doyle[1], Richard Goffin[2], and David Woycheshin[3]

[1]Organizational Development and Coaching, GALE Partners, Toronto, ON, Canada
[2]Industrial/Organizational Psychology, Department of Psychology, Western University, London, ON, Canada
[3]General Military Personnel Research and Analysis, National Defence Headquarters, Ottawa, ON, Canada

Abstract: Organizational Citizenship Behavior (OCB) is valuable to organizations and has become an important focus of employee performance evaluation. Employees' peers may be particularly well-situated to rate their OCB. We investigated the proportion of variance in peer-rated OCB attributable to the ratee (true score) versus the rater (rater bias). Furthermore, we investigated whether these proportions were affected by the familiarity of the peer with the ratee. We found that high familiarity was associated with a greater proportion of ratee variance (.43 vs. .18), and a lower proportion of rater bias (.30 vs. .51), than was the case with low-to-moderate familiarity. Thus, when choosing peers as raters of OCB, there may be value in carefully considering the peers' familiarity with the ratees.

Keywords: performance management, contextual performance, organizational citizenship behavior, familiarity, rater bias

Organizational Citizenship Behavior (OCB) is an important correlate of organizations' success, including outcomes such as performance quantity and quality, financial efficiency, and customer service (Podsakoff, Mackenzie, Paine & Bachrach, 2000). Additionally, Podsakoff, Whiting, Podsakoff, and Blume's (2009) meta-analysis found that OCBs account for at least as much variance in managerial evaluations of job performance as task performance does, and a recent benchmarking study found that 64% of organizations rate their employees' levels of OCB (Gorman, Meriac, Roch, Ray, & Gamble, 2017). Thus, OCB has important consequences and has become a crucial part of employees' performance evaluations, underscoring the importance of validly assessing OCB.

However, recent studies have found that the majority of variance in job performance ratings is attributable to rater bias, not ratee behavior, calling into question the validity of job performance ratings (Hoffman, Lance, Bynum, & Gentry, 2010). This research has focused solely on task performance, leaving ratings of OCB unaccounted for (Lance, Baxter, & Mahan, 2006). In addition, studies investigating the proportion of variance attributable to rater bias versus ratee behavior have focused exclusively on supervisory ratings. Peers are likely exposed to many of their co-workers' OCBs because they are often the recipient or target of these behaviors (Carpenter, Berry, & Houston, 2014; Podsakoff, Whiting, Welsh, & Mai, 2013). As a result, peers may be well-positioned to evaluate co-workers' OCB, so we took advantage of this potentially valuable rating source. To our knowledge, this is the first investigation to assess the proportions of variance in peer-ratings of OCB that are attributable to rater bias versus ratee behavior, in relation to peer-raters' familiarity with the ratees. In particular, we assessed (a) the proportion of variance in peer-rated OCB that is attributable to the ratee (true score) versus the rater (rater bias), and (b) whether these proportions were affected by the level of familiarity of the peer with the ratee.

Organizational Citizenship Behavior, Conceptual Framework, and Potential Contributions

Organ (1988) drew attention to the "good soldier" prototype when he coined the term "OCB." Similarly, Borman and Motowidlo (1993, p. 73) referred to a category of work behaviors called Contextual Performance that "support the organizational, social, and psychological environment in which the technical core must function." It was later acknowledged that both terms encompass the same domain (Organ, 1997), thus we simply use the term "OCB" to refer to this domain throughout this article. Moreover, our focus was on the broad *overall* domain of

https://doi.org/10.1027/1866-5888/a000229

OCB ratings. An alternative OCB structure supported by Williams and Anderson (1991) made the distinction between OCB directed toward individuals (OCB-I), and OCB directed toward the organization (OCB-O). We acknowledge that OCB-I/OCB-O is a popular model in the literature. Despite our primary focus on analyzing overall OCB, for the sake of comprehensiveness we present the corresponding OCB-I and OCB-O analyses in the Electronic Supplementary Material (ESM 1).

The Realistic Accuracy Model (RAM; Funder, 1995) was developed with ratings of personality in mind, but we believe it also provides an apt conceptual framework for studying OCB ratings. The RAM suggests that validly rating another's personality requires the ratee to exhibit relevant behaviors, the behaviors must be made available to the rater (either visibly or audibly), the rater must detect the behaviors, and finally the rater must utilize the relevant, available, and detected behaviors. Funder's (1995) propositions have been supported by studies examining the correspondence between self-ratings and others' ratings of personality (Colvin & Funder, 1991; Connelly & Ones, 2010; Paunonen, 1989). Extending the RAM to OCB, OCBs are highly visible, and peers are frequent recipients of OCBs. Therefore, peers are likely to base their ratings on their observations of the ratee's OCB, and peer-ratings should contain high levels of ratee variance. In accordance with the RAM (Funder, 1995), familiarity with the ratee's performance is likely to be associated with increased opportunity for the detection of OCB behaviors, allowing for the utilization of those behaviors in providing ratings of OCB. Thus, greater familiarity with the ratee's performance, should be associated with more variance attributable to actual ratee OCB, as well as less variance associated with rater bias. Following this line of reasoning, our main focus was on (a) whether variance attributable to the rater versus the ratee's behavior dominates peer ratings of OCB, and (b) whether the proportions of variance accounted for by the rater and ratee in peer-ratings of OCB vary when the peer rater's familiarity with the ratee's performance is high versus low-to-moderate.

To the extent that the ratee as opposed to the rater accounts for more variance in peer-rated OCB, peer ratings of OCB have more potential value to researchers and practitioners alike. Accordingly, their use could complement the use of supervisory ratings of OCB and provide fresh perspectives from a different source. Furthermore, if peer raters' high familiarity with the ratee's performance is associated with a higher proportion of ratee variance compared to rater variance, this may be an indication that only highly familiar raters should be sought out as peer raters of OCB. More ratee variance and less rater variance associated with peer raters who report high familiarity would also provide preliminary support for generalizing the RAM into the domain of peer-rated OCB. Further application of this model could then provide additional insights into the mechanisms that may improve or attenuate the validity of OCB ratings. We next describe our partitioning of variance in peer ratings of OCB, and the development of our hypotheses.

Variance Components in Peer Ratings of OCB

Three main variance components were of particular interest in this work (see Putka, Le, McCloy, & Diaz, 2008). First, as mentioned above, ratee variance is the variance attributable to ratee differences in OCB. It is analogous to "true score" variance. Second, rater variance comprises raters' systematic deviations from the typical rating. Examples of this include leniency or severity. Finally, residual variance is the variance not accounted for by the ratee or rater. Variance components are similar to reliability estimates or multiple correlations. The values reported correspond to the amount of variability in the OCB ratings that can be explained by the different rating sources. We partitioned the variance in peer ratings using random coefficient modeling (RCM) within a multi-level modeling (MLM) framework (O'Neill, Goffin, & Gellatly, 2012). The model-testing sequence began with a focus on the estimation of each of the three variance components across all of the raters and concluded with the estimation of each component separately for raters who were high versus low-to-moderate in their level of familiarity with the ratee's performance. For a more statistically oriented treatment, see Appendices A and B.

Are Peer Ratings of OCB Dominated by Ratee Variance?

As explained earlier, peers are often the recipients of OCB, and, on the basis of the RAM's fundamental tenets (Funder, 1995), peer raters should provide OCB ratings with more ratee variance and less rater variance through increased detection and utilization of OCB cues. Therefore, we predicted the following:

Hypothesis 1: In peer ratings of OCB, the proportion of ratee variance will exceed the proportion of rater variance.

Is Rater Familiarity With Ratee Performance Associated With More Ratee Variance in Peer Ratings of OCB?

Rater selection is very important when collecting peer ratings, as there are often many peers from whom to choose potential raters. Self-reported familiarity with the

ratee's performance could be a simple and cost-effective criterion for selecting raters. From a theoretical perspective, and in accordance with Funder's propositions with regard to the RAM (Funder, 1995), those who are more familiar with the ratee are likely to have more opportunity to detect the relevant OCB of the ratee, allowing them to utilize the observed behaviors as the basis for their OCB ratings. Further, peer-raters likely observe many of the ratee's OCB behaviors because peers are often the target of OCB. These OCB observations, if detected, provide raters with salient cues to utilize when providing ratings of their peers. By this reasoning, peer raters who are highly familiar with the ratee's performance should engender a larger proportion of variance accounted for by the ratee and a smaller proportion of variance accounted for by the rater, than raters who report being less familiar with the performance of the ratee. Therefore, we predicted:

Hypothesis 2a: OCB ratings by peers who are highly familiar with the ratee's performance will have a greater proportion of ratee variance than will the ratings from peers who report low-to-moderate familiarity with the ratee's performance.

Hypothesis 2b: OCB ratings by peers who are highly familiar with the ratee's performance will have a lower proportion of rater variance than will the ratings from raters who report low-to-moderate familiarity with the ratee's performance.

Method

Participants

The participants were 240 military recruits undergoing Basic Recruit Training in the Canadian Forces. Recruits represented each major branch of the Canadian forces which are Air Operations, Combat Arms, Communications and Electronics, Electrical and Mechanical Engineering, Health Services, Logistics, Military Engineering, and Naval Operations. All participants had the rank of Private except those in Naval Operations, whose rank was Ordinary Seaman. The recruits were enrolled in one of six identical Basic Training courses, which contained 39, 39, 42, 41, 39, and 37 recruits, respectively. Cohorts of recruits start training at the same time, undergo the same training curriculum, and graduate at the same time. The recruits are given ample opportunity to get to know and observe each other. For the duration of the training, recruits live together in military accommodations. Training not only includes classroom instruction and field exercises, but maintenance of personal items and accommodations. In a sense, recruits are in training from the moment that they wake up to the moment they go to sleep.

As is often the case (Balzer, Greguras, & Raymark, 2004), each participant served as a peer rater and as a ratee. To allow computation of the proportions of rater and ratee variance, each rater had to provide at least two ratings and each ratee had to receive at least two ratings. Three of the 240 recruits did not meet the inclusion criteria, resulting in a final N of 237 and a total of 622 ratings (averaging 2.62 ratings per ratee). Of the 237, 92.1% were male and their ages ranged from 17 to 48 years (M = 22.98, SD = 5.12). The recruits did not receive feedback on their OCB ratings.

Measures

Organizational Citizenship Behavior

Participants' OCB was rated using the Relative Percentile Method (RPM), which asks the rater to rate all of his or her ratees relative to the superordinate reference group on a percentile scale ranging from 0 to 100 (see Goffin & Olson, 2011, for a complete description). The RPM is rooted in social comparison theory (SCT; e.g., Festinger, 1954; Kruglanksi & Mayseless, 1990) which suggests that people have a natural facility for rating themselves and others through comparative evaluations. The RPM has been shown to improve the validity of ratings even in situations where ratings tend to be elevated (Goffin & Olson, 2011; McCarthy & Goffin, 2001), as is often the case in military settings when performance-related constructs are being evaluated (e.g., Kozlowski, Chao, & Morrison, 1998).

The OCB items came from Podsakoff, MacKenzie, Moorman, and Fetter (1990) and Hogan, Rybicki, Motowidlo, and Borman (1998). Five-item scales were used to represent each of the nine subdimensions of OCB that were described in Goffin, Woycheshin, Hoffman, and George (2013). The subdimensions included altruism, conscientiousness, courtesy, sportsmanship, and civic virtue from Podsakoff et al.'s (1990) measure. It also included the subdimensions of persisting, volunteering, helping, following, and endorsing from Hogan et al.'s (1998) measure. Each subdimension score was computed as the sum of the five items for that subdimension, thus, subdimension scores could range from 0 to 500. Goffin et al. (2013) found that when these ten subdimensions were factor analyzed, a nine-factor structure that combined the altruism and the helping subdimensions fit the data best. Therefore, we combined the subdimensions of helping and altruism. The nine OCB subscales were highly correlated, and there is a large body of research suggesting that a single factor of OCB captures the majority of its predictiveness (Hoffman, Blair, Meriac, & Woehr, 2007; Lepine, Erez, & Johnson, 2002). With this in mind, a single OCB

score with a possible range of 0–500, equal to the average of all the subdimension scores, was computed for each participant.

Familiarity With the Ratee's Performance

Participants lived and trained together, as is typical during military basic training. This resulted in enhanced opportunity for peer raters to become familiar with the ratees' performance. Raters were instructed "It would be helpful to know how familiar you are with each person's performance on recruit training. Please use the following scale to indicate how familiar you are with the performance of each recruit on the rating list by writing the number from the scale in the space provided below. These ratings are for your familiarity with performance on recruit training only, not on your familiarity with them as a friend". The corresponding scale points were: 1 = *Not familiar at all*, 2 = *Fairly familiar*, 3 = *Quite familiar*, and 4 = *Extremely familiar*. It was clear from the ordering of these scale points that higher numbers referred to successively higher levels familiarity with the ratee's performance.

Procedure

In accordance with best-practice recommendations for peer ratings (Balzer et al., 2004), each recruit was asked to rate the OCB of three of their peers and was rated by three other recruits. A list of recruits was generated for each Basic Training class, and each recruit became the intended ratee of the three recruits after them on the list. For instance, Recruits 2, 3, and 4, were asked to rate Recruit 1, and so on. This approach ensured that each recruit was targeted to receive the same number of ratings and provide the same number of ratings. The actual rating instructions to the recruits were based directly on typical RPM instructions as described in detail in Goffin and Olson (2011). The OCB ratings were part of a battery of measures for a larger project. A researcher was present while the ratings were being made in order to thoroughly explain the purpose of the study, provide assurances of confidentiality, provide instructions on completing all of the measures, and answer any questions that arose.

Results

The mean OCB rating was 308.97 (*SD* = 64.67, range = 45.78–452.33). The internal consistency (α) reliability of the 45 item OCB scale was very high, .97. Alpha (α) corresponds to the estimated proportion of shared variance across the different raters.

Does the Ratee or the Rater Account for the Largest Proportion of Variance in Peer Ratings of OCB?

In order to determine whether the ratee would account for a greater proportion of variance than the rater in peer ratings of OCB, as predicted in Hypothesis 1, several models were tested and compared. Following standard RCM methodology (e.g., Bliese & Ployhart, 2002), Model 1 (Table 1) was a completely unpartitioned model where all the variance was contained in the residual. Thus, Model 1 provided an estimate of the overall variance of the OCB ratings. The fit of this model, as assessed via the −2 log likelihood ratio (−2LL), served as a baseline to which the fit of later models could be compared using the Likelihood Ratio Test (LRT; see Bliese & Ployhart, 2002; Han, 2005). Comparing the fit of Models 2, 3, and 4 to Model 1 allowed us to assess whether separately partitioning ratee variance, rater variance, and residual variance tended to improve model fit.

Model 2 (Table 1) partitioned the variance attributable to the ratee from the residual, estimating two variance components instead of one. Support for partitioning ratee variance from the overall variance would be provided if the fit of Model 2 were found to be superior to that of Model 1. The LRT results were: $\Delta\chi^2(1) = 30.44$ ($p < .01$), which suggested that partitioning ratee variance did in fact provide improved model fit.

Model 3 (Table 1) partitioned the variance attributable to the rater from the residual. Similar to Model 2, Model 3 estimated two variance components. Support for partitioning rater variance from the overall variance would be provided if the fit of Model 3 were found to be superior to that of Model 1. The LRT indicated superior fit of Model 3 over Model 1, $\Delta\chi^2(1) = 91.74$ ($p < .01$; Table 1, Model 3) which indicated that estimating the rater variance separately from the overall variance was advantageous.

Model 4 (Table 1) was estimated in order to determine whether partitioning both the ratee variance and the rater variance from the overall variance would provide greater fit over the preceding models. Thus, Model 4 was tested against each previous model to assess whether estimating all three of the variance components improved model fit (see Table 1). The LRT value comparing Model 4 and Model 1 was $\Delta\chi^2(2) = 163.76$ ($p < .01$); comparing Model 4 and 2 the respective value was $\Delta\chi^2(1) = 133.32$ ($p < .01$); and, finally, comparing Model 4 and 3 the LRT value was $\Delta\chi^2(1) = 72.02$ ($p < .01$). Because Model 4 evidenced the best fit of all the models, Model 4 was used to estimate the proportions of variance attributable to the rater and the ratee in order to test Hypothesis 1. Accordingly, the proportion of variance attributable to the ratee was calculated by dividing the variance attributable to the ratee (1,117.13)

Table 1. Random coefficient models of organizational citizenship behavior

Model	Familiarity level	Ratee main effects	Rater main effects	Residual variance	Proportion ratee main effects	Proportion rater main effects	Proportion residual	Model fit	Parameters	Model comparison
1	Combined			4,182.01			1.00	7,315.87	2	
2	Combined	963.16		3,216.13	0.23		0.77	7,285.43	3	30.44 (Model 1)**
3	Combined		1,658.22	2,516.77		0.40	0.60	7,224.13	3	91.74 (Model 1)**
4	Combined	1,117.13	1,865.28	1,323.04	0.26	0.43	0.31	7,152.11	4	163.76 (Model 1)**
										133.32 (Model 2)**
										72.02 (Model 3)**
5	LTM	713.540	2,045.50	1,278.79	0.18	0.50	0.32	6,824.75	7	491.12 (Model 1)**
										460.68 (Model 2)**
	High	2,147.51	1,461.58	1,035.88	0.46	0.31	0.22			399.38 (Model 3)**
										327.36 (Model 4)**

Notes. Combined = both levels of familiarity (LTM & High) were combined; LTM = low-to-moderate. Model fit assessed using −2 log likelihood ratio; model comparison assessed using −2 log likelihood difference test. $^{**}p < .01$.

by the total variance for Model 4 (1,117.13 + 1,865.28 + 1,323.04 = 4,305.45), which equaled .26. Similarly, the proportion of variance attributable to the rater was calculated by dividing the variance attributable to the rater (1,865.28) by the total variance for Model 4 (4,305.45), which equaled .43. Finally, the proportion of variance attributable to the residual was calculated by dividing the variance in the residual (1,323.04) by the total variance for Model 4 (4,305.45) and was found to be .31. Therefore, Hypothesis 1 was not supported as the proportion of variance representing the ratee (.26) was less than the proportion of variance representing the rater (.43).

Is Rater Familiarity With Ratee Performance Associated With Higher Quality Ratings?

Of the 622 ratings that were collected, 18 were from raters who described themselves as *not familiar at all* with the ratee's performance (1 on the familiarity scale), 152 were *fairly familiar* (a score of 2), 277 were *quite familiar* (a score of 3), and 175 were *extremely familiar* (a score of 4). The mean familiarity score was 2.99 (SD = 0.80). Because one of the goals of investigating familiarity with the ratee's performance is its potential use as a rater selection criterion, we divided the ratings into two levels of familiarity. High familiarity was operationalized as a score of 4 (N = 175 out of 622 ratings) and represented peer raters who would be deemed most appropriate as raters in a rater selection context. Low-to-moderate familiarity was operationalized as a score from 1 to 3 (N = 447 out of 622 ratings) and represented those with an average level of familiarity with the ratee's performance, and below. As stated above, the average reported familiarity was 2.99. This dichotomization provided adequate samples sizes to allow the required models to be estimated separately for each level of familiarity (for more discussion of dichotomization within RCM, see O'Neill et al., 2012).

We predicted that the ratings from raters who were highly familiar with the ratee's performance, as opposed to low-to-moderate in familiarity, would tend to have a greater proportion of variance associated with the ratee (Hypothesis 2a), as well as a smaller proportion of variance associated with the rater (Hypothesis 2b). In order to test these hypotheses, we first needed to test whether estimating familiarity improved model fit. This involved estimating a model which partitioned the variance attributable to the ratee, the rater, and the residual separately for high and low-to-moderate familiarity (Model 5) and comparing the fit to Model 4. Thus, Model 5 estimated three more parameters than Model 4. The LRT comparing Models 4 and 5 (Table 1; $\Delta\chi^2(1) = 327.35$, $p < .01$) indicated a significantly better fit for Model 5, thus, separately estimating the variance attributable to the ratee, the rater and the residual for high and low-to-moderate familiarity provided greater fit.

In support of Hypothesis 2a, the proportion of variance accounted for by the ratee in Model 5 was greater when raters reported high familiarity (.46) than when raters reported low-to-moderate familiarity (.18). Similarly, the proportion of variance accounted for by the rater was less when raters reported high familiarity (.31) than when they reported low-to-moderate familiarity (.50), supporting Hypothesis 2b.

Discussion

It is often taken for granted that ratings reflect the ratee's behavior rather than rater bias. However, within investigations of other performance constructs, Hoffman et al. (2010) and O'Neill et al. (2012) reported that the variance

accounted for by the rater typically outweighs that of the ratee. We hypothesized that peer raters should be good sources of OCB ratings as they are often the recipients of OCB. Contrary to our hypothesis, we found that the proportion of variance associated with the rater in peer-rated OCB (.43) was larger than that associated with ratee behavior (.26). Thus, peer-rated OCB conforms to the same general pattern as ratings of other work performance constructs (Hoffman et al., 2010; O'Neill et al., 2012). These findings are disheartening and indicate that research should investigate methods to reduce the proportion of rater variance and increase the amount of ratee variance in peer-rated OCB.

One avenue for potentially improving peer OCB ratings is to consider the rater's familiarity with the ratee's performance. It was found that peer raters who were highly familiar with the ratee's performance had a larger proportion of ratee variance and a lower proportion of rater variance compared to those who reported low-to-moderate familiarity. The results indicated that the proportion of variance accounted for by the ratee was .43 for those indicating high levels of familiarity compared to .18 for those who indicated low-to-moderate levels of familiarity, a difference of .25. Additionally, peer-ratings from raters indicating high levels of familiarity with the recruit had a lower proportion of variance accounted for by the rater (.30) than did those from peer-raters who reported low-to-moderate familiarity (.51), resulting in a difference of .21. Furthermore, the proportion of variance associated with the ratee for highly familiar raters exceeded the proportion of variance associated with the rater. This suggests that, with regard to peer-rated OCB, highly familiar raters do not conform to the general pattern of higher proportions of rater variance than ratee variance than other performance constructs do.

With regard to Funder's (1995) RAM, it is likely that those more familiar with the ratee's performance had greater opportunity to detect relevant OCB, which allowed them to provide ratings based more in ratee behavior. The results for familiarity with the ratee's performance provide support for the use of Funder's (1995) RAM for understanding OCB rating behavior and suggest that familiarity with the ratee's performance could be used to select peer raters of OCB in an effort to ensure that ratings are more representative of the ratee's OCB.

Limitations and Future Research

All the participants in the present study were military recruits who were receiving basic training and did not yet have extensive military experience. Similarly, all of the data came entirely from a single large military organization. As a result, replication of these findings utilizing a diversity of private-sector and other organizations that are not military in nature would be useful in order to enhance generalizability. Similarly, with regard to generalizing within the military domain, replication with more experienced military personnel would be advisable. Further, we mentioned that the military sample used in the present study had extensive exposure to each other as they not only trained together but lived together for the duration of their training. This level of familiarity is not common in many workplaces, suggesting that future research should extend the hypotheses of the present study to additional industry samples.

Another limitation is the use of a one-item familiarity scale. Although the item was specific in asking raters to indicate their familiarity with the ratee's performance and to exclude their friendship biases, this measure could be improved upon by incorporating additional familiarity items specifically asking about their interaction with the ratee and the context of the interaction. Future research should replicate the present study using such a measure.

Additional research should also investigate the hypotheses of the present study using a larger multisource data set that includes enough information to report the analyses for supervisors, peers, and subordinates. This would allow for the comparison of rater and ratee variance across rating sources. This data set would have to be large as it would require multiple supervisors to rate each ratee, and subordinates to provide ratings for multiple ratees. Such a project would involve the collection of a lot of very specific data and would be a large endeavor.

Finally, we evaluated only one persuasive reason why rater variance may tend to be high, and ratee variance may tend to be low in typical OCB peer ratings. This was the lack of rater familiarity with the ratee's performance which likely impacted the detection component of the RAM (Funder, 1995). Other rater variables are likely to impact the quality of OCB ratings. For instance, Frame of Reference Training (FOR; Bernardin, 1979) may affect the validity of peer raters' OCB ratings by addressing the utilization component of the RAM (Funder, 1995). The purpose of FOR training is to help consistently calibrate the raters in terms of their conceptualization and operationalization of the constructs being measured. As such, raters who have received FOR training are likely to interpret and utilize the OCB behaviors they have detected similarly, resulting in more consistent and valid ratings. Future research should address the utilization component by comparing the rater and ratee variance components for those who have received FOR training versus those who have not.

Conclusion and Implications

The present study investigated the quality of ratings provided by peer raters by examining the proportion of

variance in OCB ratings that was attributable to ratee behavior versus rater bias. The results indicated that peer-rated OCB tends to be dominated by rater variance, and ratee variance composed only a small portion of the variance. However, the results also indicated that greater familiarity with the ratee's performance was associated with increased ratee variance and decreased rater variance, supporting the notion that highly familiar peer-raters provide ratings that are based more on actual ratee behavior. To the extent that peer ratings of OCB are entrusted to raters who are highly familiar with the ratee's performance, they do not follow the pattern of low ratee variance and high rater variance that is characteristic of other performance constructs (e.g., O'Neill et al., 2012). Thus, in consonance with the RAM (Funder, 1995), it seems plausible that greater familiarity with the ratee's performance may be associated with more detection of relevant OCB behavior exhibited by the ratee. Accordingly, choosing peer raters who are highly familiar with the ratee's performance, or possibly increasing raters' familiarity with ratee's performance, may be practical and cost-effective solutions to ensuring that ratings of OCB are representative of ratees' behavior and not systematic error from the rater. There is already evidence from Oh and Berry (2009) that peer-ratings may increase the operational validity of OCB ratings if combined with supervisory ratings. If peers highly familiar with the ratee's performance were used, the increment may be larger still.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1866-5888/a000229

ESM 1. OCB-I and OCB-O analyses

References

Balzer, W. K., Greguras, G. J., & Raymark, P. H. (2004). Multisource feedback. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment, Vol. 4: Industrial and organizational assessment* (pp. 390–411). Hoboken, NJ: Wiley.

Bernardin, H. J. (1979). Rater training: A critique and reconceptualization. *Academy of Management Proceedings, 1979*, 131–135.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustration. *Organizational Research Methods, 5*, 362–387. https://doi.org/10.1177/109442802237116

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmidt & W. C. Borman (Eds.), *Personnel selection inorganizations* (pp. 77–98). San Francisco, CA: Jossey-Bass.

Carpenter, N. C., Berry, C. M., & Houston, L. (2014). A meta-analytic comparison of self-reported and other-reported organizational citizenship behavior. *Journal of Organizational Behavior, 35*, 547–574. https://doi.org/10.1002/job.1909

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology, 60*, 884–894. https://doi.org/10.1037/0022-3514.60.6.884

Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092–1122. https://doi.org/10.1037/a0021212

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117–140. https://doi.org/10.1177/001872675400700202

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6*, 48–60. https://doi.org/10.1177/1745691610393521

Goffin, R. D., Woycheshin, D. E., Hoffman, B. J., & George, K. (2013). The dimensionality of contextual and citizenship performance in military recruits: Support for nine dimensions using self-, peer, and supervisor ratings. *Military Psychology, 25*, 478–488. https://doi.org/10.1037/mil0000012

Gorman, C. A., Meriac, J. P., Roch, S. G., Ray, J. L., & Gamble, J. S. (2017). An exploratory study of current performance management practices: Human resource executives' perspectives. *International Journal of Selection and Assessment, 25*, 193–202. https://doi.org/10.1111/ijsa.12172

Han, J. (2005). Crossover linear modeling: Combining multilevel heterogeneities in crossover relationships. *Organizational Research Methods, 8*, 290–316. https://doi.org/10.1177/1094428105278177

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92*, 555–566. https://doi.org/10.1037/0021-9010.92.2.555

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. https://doi.org/10.1111/j.1744-6570.2009.01164.x

Hogan, J., Rybicki, S. L., Motowidlo, S. J., & Borman, W. C. (1998). Relations between contextual performance, personality, and occupational advancement. *Human Performance, 11*, 189–207. https://doi.org/10.1207/s15327043hup1102&3_5

Kozlowski, S. W. J., Chao, G. T., & Morrison, R. F. (1998). Games raters play: Politics, strategies, and impression management in performance appraisal. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 163–205). San Francisco, CA: Jossey-Bass.

Kruglanski, A. W., & Mayseless, O. (1990). Classic and current social comparison research: Expanding the perspective. *Psychological Bulletin, 108*, 195–208. https://doi.org/10.1037/0033-2909.108.2.195

Lance, C. E., Baxter, D., & Mahan, R. P. (2006). *Evaluation of alternative perspectives on source effects in multisource performance measures*. Mahwah, NJ: Erlbaum.

Lepine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87*, 52–65. https://doi.org/10.1037/0021-9010.87.1.52

McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three

standardized reference forms. *Military Psychology, 13*, 199–222. https://doi.org/10.1207/S15327876MP1304_2

Oh, I., & Berry, C. M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology, 94*, 1498–1513. https://doi.org/10.1037/a0017221

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods, 15*, 436–462. https://doi.org/10.1177/1094428112438699

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*, 85–97. https://doi.org/10.1207/s15327043hup1002_2

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

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56*, 823–833. https://doi.org/10.1037/0022-3514.56.5.823

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *The Leadership Quarterly, 1*, 107–142. https://doi.org/10.1016/1048-9843(90)90009-7

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26*, 513–563. https://doi.org/10.1177/014920630002600307

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual- and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 94*, 122–141. https://doi.org/10.1037/a0013079

Podsakoff, N. P., Whiting, S. W., Welsh, D. T., & Mai, K. M. (2013). Surveying for "artifacts": The susceptibility of the OCB-performance evaluation relationship to common rater, item, and measurement context effects. *Journal of Applied Psychology, 98*, 863–874. https://doi.org/10.1037/a0032588

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*, 601–617. https://doi.org/10.1177/014920639101700305

History
Received August 22, 2017
Revision received November 29, 2018
Accepted December 6, 2018
Published online August 1, 2019

Acknowledgments
A portion of the data used in this article was also used in Goffin et al. (2013), but that study addressed a distinctly different set of research questions.

Funding
This research was partially supported by funding from the Social Sciences and Humanities Research Council of Canada to Richard Goffin (Grant # 435-2014-1353).

Kevin Doyle
Organizational Development and Coaching
GALE Partners
171 E Liberty St., Unit 360
Toronto, ON M6K 3P6
Canada
kevin.doyle@galepartners.com

Appendix A

Analytic Procedure – Random Coefficient Modeling

The analytic procedure for the present study is based upon that of O'Neill et al. (2012). The goal of the present study was to determine how much variance is attributable to the ratee (true score analog) versus systematic and random error. This procedure requires that ratee main effects, rater main effects, and the residual variance be partitioned from random error. The first step is to model individual OCB ratings which is referred to as the baseline model and is referred to as Model 1 in Table 1. This first step is conducted using the following equation:

$$\text{OCB performance}_{ij} = \Upsilon_{00} + r_{ij}; \quad \text{Var}(r_{ij}) = \sigma_e^2, \quad (1)$$

where OCB performance$_{ij}$, is the OCB rating for ratee i provided by rater j, Υ_{00} is the grand mean of all OCB performance ratings, r_{ij} is the rating-specific residual, and the variance of r_{ij} is estimated by σ^2_e. The residual in the first step contains all sources of variance including the ratee main effects variance, rater main effects variance, and random error. The second step is to partition the ratee main effects variance from the rating-specific residual (Model 2).

$$\begin{aligned}\text{OCB performance}_{ij} &= \Upsilon_{00} + u_{0i} + r_{ij};\ \text{Var}(u_{ij}) \\ &= \tau_T, \ \text{and}\ \text{Var}(r_{ij}) = \sigma_e^2, \end{aligned} \quad (2)$$

where u_{0i} is the deviation from the grand mean for ratee i, and r_{ij} comprises the rater main effect, and the random error. The variance of ratee main effects is estimated by τ_T. The residual in the second step includes systematic and random error. This would include rater main effects variance, and random error. Model 3 is computed using a similar equation that isolates rater main effects from ratee main effects and random error.

The third step is to isolate rater main effects in addition to ratee main effects from the residual (Model 4). This can be done by computing the following:

$$\begin{aligned}\text{OCB performance}_{ij} &= \Upsilon_{00} + u_{0i} + v_{0j} \\ &\quad + r_{ij};\ \text{Var}(u_{ij}) \\ &= \tau_T,\ \text{Var}(v_{ij}) \\ &= \omega_R, \ \text{and}\ \text{Var}(r_{ij}) = \sigma_e^2, \end{aligned} \quad (3)$$

where v_{0j} is the deviation from the grand mean for rater j, r_{ij} is now the residual variance and the random error, and the variance of rater main effects is estimated by ω_R. Equation 3 permits estimation of each of the following: ratee main effects variance (τ_T), rater main effects variance (ω_R), and the residual variance (σ^2_e).

Appendix B

Analytic Procedure – Heterogeneous Variance Structures

The effect of high and low-to-moderate rater familiarity can be investigated by implementing heterogeneous variance structures (O'Neill et al., 2012). This procedure allows for the comparison of variance proportions across different conditions. In reference to the present study, ratee main effects, rater main effects, and the residual variance will be estimated for both high and low-to-moderate rater familiarity with the ratee (Models 5 and 6).

$$\text{Var}(u_{0i}) = \left(\tau_{T,\,\text{familiarity low-to-moderate}};\ \tau_{T\,\text{familiarity high}}\right), \quad (4)$$

$$\text{Var}(v_{0j}) = \left(\omega_{R,\text{familiarity low-to-moderate}};\ \omega_{R\,\text{familiarity high}}\right), \quad (5)$$

$$\text{Var}(r_{0j}) = \left(r_{ij\,\text{familiarity low-to-moderate}};\ r_{ij\,\text{familiarity high}}\right), \quad (6)$$

where Var(τ_T) comprises ratee main effects variances for each of high and low-to-moderate familiarity, Var(ω_R) comprises rater main effects variances for each of high and low-to-moderate familiarity, and Var(r_{ij}) comprises the residual variances for each of high and low-to-moderate familiarity.