Industrial and Organizational Psychology (2024), **17**, 269–282
doi:10.1017/iop.2024.10

CAMBRIDGE UNIVERSITY PRESS

FOCAL ARTICLE

Selection tests work better than we think they do, and have for years

Jeff L. Foster[1], Piers Steel[2], Peter D. Harms[3], Thomas A. O'Neill[2], and Dustin Wood[3]

[1]Missouri State University, Springfield, MO, USA, [2]University of Calgary, Calgary, AB, Canada and [3]University of Alabama, Tuscaloosa, AL, USA
Corresponding author: Jeff L. Foster; Email: jfoster@missouristate.edu

(Received 10 November 2022; revised 11 July 2023; accepted 15 July 2023; first published online 20 August 2024)

Abstract
We can make better decisions when we have a better understanding of the different sources of variance that impact job performance ratings. A failure to do so cannot only lead to inaccurate conclusions when interpreting job performance ratings, but often misguided efforts aimed at improving our ability to explain and predict them. In this paper, we outline six recommendations relating to the interpretation of predictive validity coefficients and efforts aimed at predicting job performance ratings. The first three focus on the need to evaluate the effectiveness of selection instruments and systems based only on the variance they can possibly account for. When doing so, we find that it is not only possible to account for the majority of the variance in job performance ratings that most select systems can possibly predict, but that we've been able to account for this variance for years. Our last three recommendations focus on the need to incorporate components related to additional sources of variance in our predictive models. We conclude with a discussion of their implications for both research and practice.

Keywords: job performance ratings; job performance; employee selection; criterion-related validity

Selection tests work better than we think they do, and have for years

Job performance ratings are widely considered the primary criterion for evaluating the validity of selection practices. Furthermore, having a clear understanding of how well predictors work is an essential requirement for understanding the state of the field and its progress. However, although job performance ratings are widely used in research, the casual way in which they are often treated does not reflect their complicated nature. This, in turn, often does not give us an adequate benchmark against which to evaluate the effectiveness of selection procedures.

Confusion regarding the effectiveness of selection procedures may arise because of a failure to distinguish total variance in job performance ratings from the percentage of that variance that can be attributed to actual performance. For example, generalizability or G-theory research (see DeShon, 2003) has shown that performance ratings often do not contain much "performance valid variance" (cf. O'Neill et al., 2015; Putka & Hoffman, 2013). Indeed, Scullen et al. (2000) found that only 21%–25% of variance in ratings can be attributed to ratee main effects, which is the variance associated specifically with the person being rated. Similarly, Viswesvaran et al. (2005) concluded that a general factor accounts for 27.4% of uncorrected between-rater correlations. More recently, Jackson et al. (2020) determined that rater main effects account for as little as 20%–30% of the variance in performance ratings depending on the level of aggregation of scores across dimensions and raters. Although results relating to specific sources of variance have varied throughout the years (e.g., Greguras & Robie, 1998; Hoffman et al., 2010; Jackson et al., 2020;

© The Author(s), 2024. Published by Cambridge University Press on behalf of Society for Industrial and Organizational Psychology.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Kraiger & Teachout, 1990; O'Neill et al., 2012; O'Neill et al., 2015; Schmidt et al., 2021; Scullen et al., 2000; Woehr et al., 2012), likely due to differences in available data, research designs, and organizational environments and contexts, one thing that remains consistent is that variance specific to only the person being rated (i.e., ratee main effects) represents only a fraction of the total variance in job performance ratings.

Despite the consistency of these findings and their potential implications for faith in the construct validity of performance ratings, many organizational scholars and practitioners regularly treat job performance ratings only as measures of performance. For example, most organizations base at least some decisions on performance ratings provided by managers based on the assumption that employees with higher ratings are better performers than their counterparts. Similarly, researchers regularly evaluate the effectiveness of selection measures based on the variance they account for in overall job performance ratings. However, the intent is usually to identify how well they predict performance, which represents only a small percentage of this variance. As a result, even the most effective performance predictors look dismal.

Predicting performance versus performance ratings

It is important, therefore, to distinguish between predicting performance versus predicting job performance ratings. Although actual performance is one factor that influences job performance ratings, other sources of variance such as rater main effects (e.g., various rater biases and response patterns) also influence results, thereby contributing to total variance in job performance ratings. Therefore, when trying to predict performance, at least when using job performance ratings as criterion variables, we can only hope to predict variance attributed to ratee main effects with *most* selection instruments. This is because other known sources of variance require consideration of or interaction with raters who job applicants have not yet met, let alone worked with.

Therefore, when trying to predict performance, we should focus only on accounting for variance in performance ratings that we can attribute to the person being rated. Often, that may be limited to variance attributed to ratee main effects, but this can also include rater × ratee interactions, such as when an employee performs better for a particular supervisor or when a supervisor especially values a specific work style or approach.

Furthermore, we need to distinguish efforts to predict performance from those aimed at predicting performance ratings. For example, we can estimate response tendencies such as leniency bias by making comparisons among multiple raters. Potentially we could develop a Severity, Strictness or Leniency Index to estimate it directly (Cheng et al., 2017). When used with other predictors, such estimates of rater leniency can help us better predict performance ratings. In contrast, when used as a covariate, rater leniency can help us better predict actual performance because it removes variance not attributed to the person being rated.

Such a distinction highlights that we can often make better decisions when we have a better understanding of the different sources of variance that impact job performance ratings. A failure to do so not only leads to inaccurate conclusions when interpreting results, but often misguided efforts aimed at improving our ability to explain and predict performance ratings. For example, as we outline below, using a battery of well-built and valid selection tools, we can already account for a large percentage of variance attributed to ratee main effects in job performance ratings. In fact, common and well-established selection procedures can already account for most of the variance in performance ratings that they are able to predict, which means they work much better than we often give them credit for.

However, we are not nearly as good at predicting the full range of variance components of job performance ratings. To do so requires not only a better understanding of the different sources of variance in those ratings, but that we use predictive models that account for those sources of variance. The more common approach is treating and correcting for other sources of variance as

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

though they were error (e.g., correcting for inter-rater reliability), which may be useful for helping estimate the value of selection tools but don't help us actually predict performance or performance ratings more accurately. Outside of selection, uncorrected performance ratings are of critical importance in other HR decisions, such as promotions, demotions, resource allocation, project assignments, pay raises, and developmental opportunities. In other words, we shouldn't always treat these other sources of variance as error (Deshon, 2003). Correcting for artifacts has its role, but it can mask the challenges needed to predict or decompose actual job performance ratings.

Recommendations

Discussions about performance ratings are not new to the field of industrial-organizational psychology or this journal. However, what is often lacking is a thorough understanding and treatment of the different sources of variance in performance ratings. A more thorough recognition these sources of variance, along with a better understanding of the limitations of each, leads to several considerations and suggestions for using and interpreting individual predictors of job performance ratings and the ratings themselves. We outline six such recommendations relating to: (a) the interpretation of predictive validity coefficients and (b) efforts aimed at predicting job performance ratings. We conclude with a discussion of their implications for both practice and research.

Interpreting predictive validity coefficients

Research shows that ratee main effects, or those that are specific to the person being rated, accounts for a fraction (20-30%) of the variance in job performance ratings (Jackson et al., 2020). This relatively small percentage of variance reflects both overall performance and domain specific performance, indicating that the complexity we often attribute to variations in performance both within and across individuals is often dwarfed by other sources of variance in performance ratings. More importantly, it is the only source of variance that can be accounted for by most selection instruments, where job applicants answer questions or perform in a context that is independent of their future raters. Such selection assessments simply cannot produce results that can account for variance attributed to rater idiosyncratic effects (i.e., rater main effects and rater × ratee interaction effects).

A G-theory model (Cronbach et al., 1972) can help one identify what variance to consider as performance valid for generalizability purposes. For example, as DeShon (2003) noted, in top-down selection one way to approach the issue is to conceptualize the percent of performance valid variance as the components involving only ratee main effects divided by those effects plus rater × ratee interaction variance plus residual variance (as well as other ratee interaction effects if they exist, such as ratee × item variance) because that is the variance associated with rank order differences. Other approaches in other contexts may be more appropriate. The key is that failing to make an argument about the nature of variance components of job performance ratings, and to not model them in empirical validation research, can lead to widely inaccurate conclusions (see O'Neill et al., 2012; Putka et al., 2008).

In a cavalier and more typical way, it is common to refer to the squared correlation between a predictor score and job performance ratings as accounting for the percentage of variance in job performance. Therefore, if a study finds an observed correlation between a selection test score and future measure of job performance of .20, one might conclude that scores on this predictor account for only 4% of the variance in job performance, leaving the remaining 96% unaccounted for. This ignores the components of valid and invalid performance variance, thereby making the predictors look negligible when they might not be.[1]

[1]More generally, see Funder and Ozer (2019) for a discussion about how the widespread practice of squaring correlations to estimate variance explained is actually a highly misleading metric.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Given that even our most predictive assessments rarely result in observed correlations with job performance ratings larger than .20, such a conclusion paints a rather dim picture of our ability to predict job performance. Instead, although it is true that a correlation of .20 indicates that a selection measure only accounts for 4% of the variance in job performance ratings, one must keep in mind when interpreting this result that actual job performance can account for only a fraction of the variance in the ratings itself. To emphasize the importance of this distinction and suggest better practices, we offer the following three recommendations concerning the percentage of variance in job performance ratings that can be attributed to performance:

It is more appropriate to report the effectiveness of most selection assessments as a percentage of variance that can be attributed to ratee main effects than as a percentage of overall variance in job performance ratings

Although the variance associated with ratee main effects in job performance ratings might include more than just job performance, they set an upper limit on the percentage of variance that can be impacted by the individual characteristics of the person being rated. For example, how someone scores on a cognitive ability test or personality test simply cannot account for variance in job performance ratings attributed to rater idiosyncratic effects. To do so requires at least some information about or participation by the rater. But in most selection testing scenarios, this is not the case.

Therefore, rather than simply reflecting 4% of the total variance in job performance ratings, an observed correlation of .20 between an individual predictor measure and performance ratings accounts for closer to 13%–20% of the variance that can be attributed to ratee main effects (e.g., 4% out of a total of 20-30%), which paints a more realistic and positive outlook for any single predictor. This is a common step, for example, in meta-analysis, where a pseudo R^2 is reported, which ignores the non-predictable sampling error.

For example, consider cognitive ability, which often has one of the highest observed correlations with job performance ratings of any commonly used selection technique or measure. Further consider Hunter's, 1986 meta-analysis, which reported an observed correlation with job performance ratings of .32 for complex jobs, corresponding to an R^2 of 10.24%. When compared to the total variance in performance ratings, one might be tempted to conclude that the ability to process, manipulate, and make sense of complex information accounts for only 10% of the variance in job performance for jobs that require a large degree of information processing. And although we can correct this value for artifacts such as unreliability and range restriction (see more on that below), such corrections can introduce systematic bias and produce unrealistically high estimates (Deshon, 2003; LeBreton et al., 2014; Sackett, 2014; Sackett et al., 2022). Furthermore, they may result in lower total estimates for multivariate models because they generate high estimated intercorrelations between predictors, misleadingly suggesting high overlap between contributing variables. But perhaps most importantly, they don't actually increase our ability to predict job performance ratings.

However, rather than arguing for the validity and accuracy of statistical corrections that are potentially problematic themselves, we believe it is more reasonable and accurate to compare observed correlations and R^2 values to the percentage of variance that can be attributed to ratee main effects. A conservative estimate of performance relevant ratee main effects may be 25% of the variance in performance ratings, conservative in that ratee main effects can include shared bias among supervisors (e.g., if supervisors share a bias against a group, this increases the size of ratee main effects as well as supervisor inter-rater reliability). Comparing predictor contributions with this 25% predictable performance relevant variance invisible ceiling (which could be even lower), indicates that cognitive ability's R^2 of roughly 10% actually predicts 40% of the variance in ratee main effects in complex jobs. This seems a much more reasonable expectation concerning the likely influence of cognitive ability on performance for these jobs.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Or consider results from Park et al.'s (2020) meta-analysis comparing personality results to a variety of outcomes. Focusing on just task performance as an example, we find observed correlations of .14 for Emotional Stability, .08 for Agreeableness, .16 for Conscientiousness, .07 for Extraversion, and −.01 for Openness. Squaring the results and comparing them to a 25% ceiling for rater main effects indicates that across a range of jobs, Emotional Stability accounts for approximately 8% of the variance in that can be attributed to ratee main effects, Agreeableness around 2.5%, Conscientiousness a little over 10%, Extraversion 2%, and Openness almost nothing. Adding the values up indicates that roughly 20-25% of the variance in task performance ratings that can be attributed to ratee main effects across a range of jobs can be predicted with accurate and reliability personality instruments.

However, this sum may be inflated due to intercorrelations between the predictors. So instead, we can look at regression weights from Park et al. (2020), which give us an R^2 of .04, indicating that approximate 4% of the variance in performance ratings can be accounted for by personality predictor scales. Comparing that to a 25% ceiling indicates that an appropriate combination of personality scores can predict approximately 16% of the variance in ratee main effects. Results are two to three times greater (40-48%) when predicting OCBs and CWBs and greater still when examining predictive validity coefficients for specific jobs or job families.

Although our purpose here is not to reexamine every predictive validity coefficient that has been reported or advocate for the use of some selection instruments over others, we think it is important to examine all such results with a more appropriate reference concerning the amount of variance in job performance ratings that can possibly be predicted using most selection instruments. When doing so, we get a much more reasonable and realistic picture of how those instruments perform. Furthermore, a ceiling of 25% not only provides a number that is easy to work with, but one that is very similar to estimates of ratee main effects from previous research (e.g., Jackson et al., 2020; O'Neill et al., 2015; Scullen et al., 2000; Viswesvaran et al., 2005), We propose this number only as a rough estimate that is useful in that it allows us to better evaluate and understand the effectiveness of selection instruments for predicting performance.

It is also important to note that other factors might affect the percentage of variance in performance ratings attributed to ratee main effects, such as the content of the ratings, potential job characteristics, or study design issues such as aggregating ratings across multiple raters. For example, collecting data from and averaging ratings across multiple raters should reduce at least some of the overall variance in job performance ratings that can be attributed to rater main effects, thereby increasing the "ceiling" representing the percentage of variance that is left to ratee main effects. In other words, if you remove or reduce the effects of one source of variance, every other source represents a larger percentage of what is left. Therefore, if aggregating results across multiple raters, it may be more appropriate to compare observed correlations to a 30% ceiling based on Jackson et al. (2020)'s estimate that 29.77% of the variance in job performance ratings results from rater main effects after aggregating ratings across multiple raters within rater source. Rater training or other attempts to potentially mitigate the influence of rater main effects and/or rater × ratee interaction effects may also result in the need for a higher ceiling, which highlights the need for future research to continue to examine the potential influence of factors that impact the percentage of variance attributed to ratee main effects and other sources of variance.

We need to look at the overall effectiveness of selection systems containing multiple steps and selection measures in this same fashion

Similarly, we should evaluate most selection systems containing multiple diverse predictors based on the variance they can account for attributed to rater main effects. When doing so, we find that it is not only possible to account for most of the variance in job performance ratings that can be attributed to ratee main effects, but that we have been able to account for this variance for quite some time. In short, our assessments work well and have for years.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Table 1. Relationships Between Common Predictors and Job Performance Ratings[i]

Variable	1	2	3	4	5	6	7	8	9
1. Performance									
2. Interview	0.19[a]								
3. EI	0.24[c]	0.22[f,g]							
4. GMA	0.29[h]	0.14[e]	0.05[c]						
5. N	−0.09[h]	−0.04[e]	−0.47[c]	−0.07[d]					
6. E	0.04[h]	0.10[e]	0.42[c]	0.00[d]	−0.26[b]				
7. O	0.05[h]	0.04[e]	0.33[c]	0.19[d]	−0.07[b]	0.29[b]			
8. A	0.08[h]	0.06[e]	0.32[c]	0.00[d]	−0.16[b]	0.20[b]	0.19[b]		
9. C	0.17[h]	0.08[e]	0.32[c]	0.02[d]	−0.23[b]	0.19[b]	0.16[b]	0.29[b]	

[a]Schmidt & Rader, (1999).
[b]Park et al. (2020).
[c]O'Boyle et al. (2011).
[d]Stanek (2014).
[e]Salgado & Moscoso (2002).
[f]Christina & Latham (2004).
[g]Kluemper et al. (2015).
[h]Schmidt et al. (2008).
[i]Salgado et al. (2003).

To illustrate our point, we first identified results from published empirical research examining intercorrelations between common predictors and their relationships with job performance (Christina & Latham, 2004; Kluemper et al, 2015; O'Boyle et al., 2011; Park et al., 2020; Salgado, 2003; Salgado et al., 2003; Salgado & Moscoso, 2002; Schmidt & Rader, 1999; Schmidt et al., 2008; Stanek, 2014). Predictors included general mental ability (GMA), structured behavioral interviews, and emotional intelligence. We also included results for personality where each of the FFM model scales (Digman, 1990; Goldberg, 1992; John, 1990; McCrae & Costa, 1987) served as predictors, resulting in eight predictor scales derived from four assessments commonly used for selection.

Next, we created a validity coefficient matrix including relationships between predictors and relationships with each predictor and job performance ratings (see Table 1). To determine the percentage of variance in job performance ratings accounted for from a combination of these predictors, we conducted an OLS multiple regression analysis based on a harmonic mean of 2277, which produced an R^2 of .164.

When compared to our 25% ceiling, we see that a combination of four of our field's most commonly used job selection tools predicts about 66% of the variance in job performance ratings that can be attributed to ratee main effects. Given that we did not correct for reliability in our predictors, this is operational validity and could be improved with more reliable assessments. And given that we did not correct for range restriction in the performance criterion, as per Sackett et al. (2022), this is likely a small underestimate. In other words, four well-designed and valid selection procedures, which may take no more than a couple of hours to administer, can account for most of the variance we can possibly predict with traditional selection tests.

Furthermore, our ability to predict most of this variance should not only give us more confidence in relying on assessments for which there exists an abundance of validity evidence but emphasizes the need to focus on other sources of variance if we want to increase our ability to predict performance ratings. For example, consider meta-analysis and its validity generalization sub-discipline of synthetic validity (Johnson et al., 2010). Specifically, if we are confident that we can predict performance sub-dimensions such as Data, People, Things, and Organizational

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Citizenship and then combine then into an overall weighted composite score (Schmidt & Hunter, 2014), we could implement a three-stage or multi-hurdle system.

Using the same principles used in most screening systems, we move from broad to narrow (e.g., for water filtration it moves from gravel to sand to charcoal). In the first stage (e.g., gravel), we focus on Person-Organization (P-O) fit, which has its own set of criteria and standards (e.g., complementary versus supplementary fit; Barrick & Parks-Leduc, 2019). In this second stage (e.g., sand), we focus on Person-Job (P-J) fit and seek to account for variance associated with ratee main effects. Here, Meta-Analytic Structural Equational Modeling can be employed, with the validity correlations moderated by job characteristics and the weighted composite score influenced by organizational strategy (Jak & Cheung, 2020; Steel et al., 2021). In the third stage (e.g., charcoal), presumably after people are hired to minimize adverse treatment, we are now finding the best *position* for them within the *job* family by focusing on remaining variance and match to maximize Person-Supervisor (P-S) fit (i.e., address rater main effects and ratee × rater interaction effects).[2]

For example, one way we could operationalize this third stage is by reestablishing the importance (i.e., the weights) that individual supervisors give to the different performance sub-dimensions (e.g., Data, People, Things, and Organizational Citizenship), and then recalculate composite scores as to rank and match accordingly. Although such a process requires that we learn more about the nature of and how to account for these effects, it is superior to the tradition of relegating rater × ratee interactions as error to be statistically corrected. Far from being error, rater × ratee interactions are reflected in P-S fit, so they lead to higher leader-member exchange, supervisory trust, job satisfaction, affective commitment, and Organizational Citizenship Behaviors (Baik & Pak, 2020; Van Vianen, 2018). Consequently, individual supervisors or sole proprietors should want to know who they specifically would view as high performers, not what happens on average. Furthermore, individual employees would not consider working with supervisors who value *their* contributions as error, especially with subsequent performance assessments and interpersonal relationships influencing their job satisfaction, promotion and pay raises. Notably, such a methodology could provide insights to help reduce adverse impact and treatment while simultaneously improving productivity, two of the great challenges of our time.

Efforts aimed at creating new selection tests should still focus on being as predictive as possible but should highlight other benefits as well

As stated above, most selection tests are typically administered to job applicants outside of the context of the work environment or without any interaction with the person(s) providing performance ratings. As a result, a 25% ceiling is a reasonable estimate to examine the effectiveness of these instruments because it represents the maximum percentage of variance in performance ratings that they can capture. Therefore, if a combination of selection instruments can account for two-thirds and likely more of the variance associated with ratee main effects, there may be little room for improvement for new selection instruments in terms of predictive accuracy. Of note, even if we could predict 100% of this variance, this does not mean organizational main effects, such as reward systems or training programs, are ineffective. Affecting almost everyone within an organization, they are not attributable to ratee effects although they can increase the impact of individual differences. In particular, training programs not only are given to but also benefit most those who are conscientious and clever (Schmidt et al., 2008), augmenting the relationship these individual variables have with performance.

Also, this high variance percentage does not mean that such efforts cannot produce more efficient means of obtaining predictive information from applicants or ways that may reduce bias

[2]Aside from these formal assessments, we could even add an informal, preliminary stage at the start, that is Person-Occupation or Person-Vocation fit, where vocational counseling results in self-selection to a discipline or a field.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

or discrimination, or that are easier or less costly to administer, all of which are important goals. But if the primary goal of such efforts is to increase predictive validity over and above currently existing procedures, there may simply be too little additional variance left to predict, at least in terms of overall job performance. Although creators of new selection tests or procedures must still show that their instruments are as predictive as the instruments they intend to replace, they should focus more on highlighting other benefits of their approach. Predictive validity is certainly important, but it should not be the only thing we focus on when evaluating the effectiveness and utility of a selection instrument.

Furthermore, some selection practices can include one or more individuals who may later rate a person's performance. This is especially common for interviews involving managers who will supervise the new hire. In such cases, it is possible that any predictive validity associated with such interviews reflects at least some variance in future job performance ratings associated with rater main effects and ratee × rater interaction effects. If a manager has some understanding of the unique characteristics they associate with high performance in subordinates, they may focus on these characteristics during their interactions with job applicants. In fact, it might be possible to design interviews specifically with this goal in mind, which emphasizes the importance of understanding variance associated with rater main effects and ratee × rater interaction effects so we might be able to account for larger percentages of variance in job performance ratings above and beyond variance attributed to ratee main effects.

Dealing with rater main effects and rater × ratee interactions

Rather than toiling over validity coefficients that represent small effect sizes, are often plagued by sampling error with small sample sizes, and generally seem to indicate an inability to effectively predict job performance ratings, we should celebrate the decades of work that have gone into developing effective tools for capturing the wide variety of individual characteristics that influence job performance. In fact, the ability to predict the majority of the variance in job performance ratings that can be attributed to the person being rated using no more than a couple of hours of sound psychological assessment is quite impressive given the diversity and complexity of human characteristics and behaviors.

However, this doesn't mean we shouldn't focus on other methods for predicting and/or explaining variance in performance ratings beyond simply trying to create better selection instruments. Performance ratings usually have serious consequences for employees even though they represent only a modest estimate of actual performance. When it comes to many important organizational decisions, such as pay raises and promotions or project assignments and training opportunities, the ratings themselves often matter a great deal. They may also serve as a self-fulfilling prophesy by leading to opportunities that actually do impact future performance, and the only way to substantially improve our ability to predict performance ratings is to focus on and incorporate other sources of variance in our predictive models.

We should do a better job of measuring rater main effects and incorporating them into our predictive models

Instead of continuing to focus almost entirely on measures that can only predict ratee main effects, future efforts should focus more on including variables that account for variance attributed to other components (Kane, 2011). For example, we often attempt to minimize the impact of sources of rater main effects, such as leniency and halo biases, through rater training (i.e., a standardization approach). And although such training might be beneficial, it can be easy and sometimes just as practical to estimate leniency bias in raters who regularly rate the performance of multiple employees over multiple time periods. Essentially, such error is *random* between individuals but *systematic* within, a crucial difference. Measuring and incorporating leniency bias

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

into predictive models would not only potentially offer a long-term solution for accounting for their effects but should increase our ability to predict performance ratings. In other words, we might only need to only include a variable in our model that accounts for individual leniency effects by predicting higher ratings from traditionally lenient raters and lower ratings from traditionally less lenient raters (e.g., the Personal Equation approach; Cattell, 1893). So far, we have only made nascent steps towards creating performance specific self-report leniency indices, but it remains a promising venue (Cheng et al., 2017; Burchett & Ben-Porath, 2019; Rohling et al., 2011).

It would also be useful to rely on multiple raters to assess performance (Deshon, 2003). As per Generalizability Theory, averaging scores across multiple raters inevitably serves to reduce rater main effects, thereby increasing the percentage of remaining variance that we can attribute to ratee main effects (i.e., it raises our ceiling). This might not be possible for all jobs – such as jobs where evaluating performance requires extremely specialized knowledge and intimate exposure to the person's behavior. However, it should be possible for most. If the goal of data collection is to assess performance as accurately as possible, then researchers should always consider obtaining ratings from multiple raters familiar with an employee's behaviors and performance.

For example, when collecting performance data as part of a criterion-related validity study, relying on individual raters results in less reliable criterion measures, which reduces the overall magnitude of the correlations one can find (i.e., lowers the ceiling by increasing the size of rater main effects and ratee × rater interaction effects). This makes it less likely that one will even get results in the expected direction, let alone results that are statistically significant. Simply relying on statistical corrections after the fact is not a good substitute for collecting more reliable data to begin with. The default, therefore, should be to always try to collect data from multiple raters and only rely on ratings from individual sources for the rare occasions when multisource ratings are not feasible.

We should stop incorrectly categorizing known sources of variance as error and always treating them as such

Of course, using multiple raters will not eliminate the influence of rater main effects, especially when relying on a small number of raters or when ratings vary drastically across raters. As a result, it may still be reasonable to assess and correct for inter-rater reliability at times. And although an average reliability estimate of around .52 has been amazingly consistent across multiple studies (e.g. Salgado et al., 2016; Shen et al., 2014; but see also LeBreton et al., 2014), these studies also show that inter-rater reliability varies across samples, highlighting the value of sample specific estimates when possible. Therefore, before automatically employing a standard correction for all rater main effects, such as an estimate of .52, it is possible to calculate inter-rater reliability for specific samples to use for corrections while reducing the effects of low inter-rater reliability simply using multiple raters.

Even so, it is still possible to overcorrect. A combination of corrections for various forms of range restriction and reliability can produce correlations that seem unrealistically high (Schmitt, 2007). For example, Schmidt et al. (2008) estimated that the average correlation between GMA and job performance was as high as .88 for some jobs after applying multiple corrections to an observed correlation of .23, representing over a 1400% increase in the variance accounted for in the criterion measure (i.e., R^2 increasing from 5.5% to 77.4%). One issue with such estimates is that they suggest that only one or two predictors could account for nearly all the variance in ratee main effects, thereby ignoring predictive validity evidence for other selection instruments, as well as other factors known to influence performance such as person–organizational fit (Nye et al., 2017; van Vianen, 2018), network and mentoring (Wai & Rindermann, 2017), and employee training (Arthur et al., 2003). Another is that they are based on assumptions that essentially discard the vast majority of the variance in the criterion items they intend to predict.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

If our goal is to estimate the relationship between a specific predictor and generically perceived job performance, then correcting for inter-rater reliability may make sense, but it is important to remember what we are doing in this case. Specifically, we are estimating the expected relationship between the predictor and the mean estimate from an infinite number of raters, sampled in the same way as the observed raters (e.g., socialized, selected and trained similarly, so likely seasoned supervisors from the same occupation and organization), which is not that same thing as predicting actual job performance ratings. Instead, when predicting job performance ratings is our goal (i.e., how a specific supervisor would rate an individual), we must measure and incorporate variables relating to these additional known sources of variance into our predictive models rather than duping ourselves into thinking that statistical corrections have somehow made our estimates and conclusions more accurate.

We need to focus more on rater × ratee interaction effects

We know less about rater × ratee interaction effects than either of their associated main effects even though they account for substantial variance in performance ratings. One difficulty with pinpointing rater × ratee interactions is that we rarely have data where multiple raters have all rated the same set of ratees.[3] Instead, performance data are often nested within raters whereas individual raters, often managers, each rate separate employees. As such, Scullen et al. (2000) notes the presence of such interactions but only that they are combined in their analyses with other rater main effects such as leniency and halo.

However, by relying on data where each ratee was rated by multiple raters and raters were identified by source (e.g., manager, peer, or subordinate), Jackson et al. (2020) were able to estimate that interactions attributed to both rater source × ratee main effects, which accounted for between 16–28% of the variance in performance ratings and rater (nested within source) × ratee effect, which accounted for an additional 5–8% of the variance in performance ratings. These results are particularly intriguing because they indicate that the largest source of interaction effects may not be from interactions between individual raters and ratees but are interactions relating to rater source.

Much of the variance in multisource performance ratings, therefore, likely results from some individual ratee behaviors being viewed more favorably by raters representing specific perspectives (e.g., managers) than others (e.g., peers or direct reports). In other words, the characteristics and behaviors that managers associate with high performance might be different than the characteristics that peers or direct reports associate with high performance. Or in relation to variance associated with rater × ratee interaction effects within rater source, the characteristics and behaviors that one manager views as effective might be different than those viewed as effective by a different manager. It might also be useful to partition broad groups such as coworkers, peers, or direct reports to those working within particular teams, functions, or environments, or who have varying frequencies or forms of interaction with the ratee. Research examining the complexities of roles and relationships with ratees could help further predict, explain, and inform uses of performance ratings provided by a diverse range of other potential raters.

Unfortunately, we don't currently know much about ratee × rater interaction effects or how to incorporate them into our predictive models. We know that different raters think more favorably of some ratees than others, and likely vice versa, but we often don't know why. Some work has been done on person–supervisor (P–S) Fit, though it typically focuses on perceived match, congruence or similarity between the two (e.g., Peng et al., 2022). More relevant research draws on implicit leadership theory and implicit followership theory, where performance ratings fluctuate

[3]One caveat is that it may be unwise to move beyond rater × ratee interactions to consider more elaborate combinations because there is the potential to create a near-infinite number of potential interactions between different combinations of raters and/or traits, particularly when most studies are woefully underpowered to robustly test even a few moderators (Murphy & Russell, 2017).

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

massively to the degree the ratee matches the rater's prototypes (Junker & Van Dick, 2014). Furthermore, such variance might reflect performance, such as when one rater simply has more experience with a ratee than another rater, or might not reflect performance, such as when these effects are the result of actual biases. Regardless, we rarely account for these interactions in predictive models. However, we believe this is an area ripe for exploration and one that could substantially improve our ability to predict and explain job performance ratings. We envision a day where we can directly assess supervisors' ideal employees in terms of desired behaviors, which then can be predicted by individual differences and subsequently matched.

Discussion/implications

Exploring and incorporating multiple sources of variance in models aimed at predicting job performance could have several benefits. The first, which simply comes from recognizing the limitations of accounting for only one potential source of variance, is that it provides us with a means of better representing our ability to predict job performance. As a field, we are much better at identifying high performing employees than we often give ourselves credit. However, when we judge our measures against their ability to predict job performance ratings, we not only sell them short, but potentially make it more difficult to convince others to use otherwise effective procedures and instruments. We need to always be mindful to not interpret correlations with job performance ratings as indicators of how well selection measures predict actual performance.

Second, considering multiple sources of variance should help us design better studies. Short of using corrections for artifacts such as unreliability and range restriction, there may be little we can do to improve evidence that highlights the usefulness of our instruments when we only have data representing our predictors and limited performance ratings provided by one rater, usually a manager, at one point in time. In fact, when that is all that is available, it might be better to not conduct a criterion-related validity study at all, especially when dealing with small samples. Instead, researchers and practitioners should both carefully take into account such considerations and make every effort to incorporate measures relating to other sources of variance, such as rater source or biases, into their research when trying to predict either performance or performance ratings. Furthermore, nearly all studies would benefit from obtaining data from multiple raters when the goal is to examine actual job performance.

More focus on rater main effects could also be beneficial in terms of providing training and feedback to raters regarding the characteristics and utility of the ratings they provide. Although training aimed at reducing rater biases can be effective, these effects are often short lived and deteriorate over time (Roch et al., 2012). However, the data that are often necessary for examining rater biases, such as ratings for multiple people and/or over multiple points in time, should also be useful for communicating and tracking the impact of such biases for specific individuals. In other words, rather than just talking about potential biases and their impact, regularly collecting and examining data relating to biases could not only improve our predictive models but have the added benefit of assessing them (e.g., a Leniency Index) as well as facilitating training to have more profound and lasting positive effects. Such data might also be useful as a means for holding raters accountable for more accurate and useful ratings.

Furthermore, focusing more on rater × ratee interaction effects could help us design selection systems that more effectively predict additional sources of variance in job performance ratings. For example, identifying characteristics viewed by specific raters (e.g., a person's future manager) or groups of raters (e.g., subordinates or peers) and then focusing on these characteristics could improve prediction by capturing at least some variance attributed to sources of variance other than ratee main effects. Different managers inevitably hold different assumptions or for how they view performance for specific subordinates in specific roles. Although we might view these differences as biases, they still impact an employee's likelihood of success and, therefore, of

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

receiving higher performance ratings. Identifying the characteristics valued by specific managers and focusing on these characteristics in selection processes could help us better identify job applicants who are not only more likely to perform well in a specific job, but also those who are most likely to perform well in that job for a specific manager, which would increase the overall value and utility provided by the selection system. This might be particularly applicable to sole proprietors or within small organizations with a limited number of incumbents in each job.

Finally, regularly considering and evaluating multiple sources of variance in job performance ratings should also help us better understand and explain what the ratings given to individual employees actually mean and represent. A failure to accurately reflect performance is particularly problematic when organizations continue to use performance ratings as if they reflected a Platonic ideal instead of an imperfect operationalization. A better job of modeling the various components that contribute to variance in ratings will not only help us understand ratings better, but should help us communicate their meaning, limitations, and proper use within organizations.

References

Arthur, W., Bennett, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, **88**(2), 234–245. https://doi.org/10.1037/0021-9010.88.2.234.

Baik, S., & Park, H. (2020). A meta-analytic study on the relationships of person-supervisor fit with job attitudes and behaviors. *Korean Journal of Industrial and Organizational Psychology*, **33**(3), 233–265.

Barrick, M. R., & Parks-Leduc, L. (2019). Selection for fit. *Annual Review of Organizational Psychology and Organizational Behavior*, **6**, 171–193.

Burchett, D., & Ben-Porath, Y. S. (2019). Methodological considerations for developing and evaluating response bias indicators. *Psychological Assessment*, **31**(12), 1497–1511. https://doi.org/10.1037/pas0000680.

Cattell, J. M. (1893). On errors of observation. *American Journal of Psychology*, **5**(3), 285–293.

Cheng, K. H., Hui, C. H., & Cascio, W. F. (2017). Leniency bias in performance ratings: The big-five correlates. *Frontiers in psychology*, **8**, 521.

Christina, S. C., & Latham, G. P. (2004). The situational interview as a predictor of academic and team performance: A study of the mediating effects of cognitive ability and emotional intelligence. *International Journal of Selection and Assessment*, **12**(4), 312–320. https://doi.org/10.1111/j.0965-075X.2004.00286.x.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

DeShon, R. P. (2003). A generalizability theory perspective on measurement error corrections in validity generalization. In K. R. Murphy (Eds.), *Validity generalization: A critical review* (pp. 365–402). Lawrence Erlbaum Associates Publishers.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, **41**, 417–440.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, **2**, 156–168. https://doi.org/10.1177/2515245919847202

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, **4**, 26–42. https://doi.org/10.1037/1040-3590.4.1.26.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, **83**, 960–968.

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

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge and job performance. *Journal of Vocational Behavior*, **29**, 340–362.

Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, **105**, 312–329.

Jak, S., & Cheung, M. W. L. (2020). Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological methods*, **25**(4), 430–455.

John, O. P. (1990). The big-five factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Eds.), *Handbook of personality theory and research* (pp. 66–100). Guilford.

Johnson, J. W., Steel, P., Scherbaum, C. A., Hoffman, C. C., Jeanneret, P. R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. *Industrial and Organizational Psychology*, **3**(3), 305–328.

Junker, N. M., & Van Dick, R. (2014). Implicit theories in organizational settings: A systematic review and research agenda of implicit leadership and followership theories. *Leadership Quarterly*, **25**(6), 1154–1173.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, **48**(1), 12–30.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Kluemper, D. H., McLarty, B. D., Bishop, T. R., & Sen, A. (2015). Interviewee selection test and evaluator assessments of general mental ability, emotional intelligence and extraversion: Relationships with structured behavioral and situational interview performance. *Journal of Business and Psychology*, **30**(3), 543–563. https://doi.org/10.1007/s10869-014-9381-6.

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance*, **3**, 19–35.

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

McCrae, R. R., & Costa, P. T., Jr (1987). Validity of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, **52**, 81–90.

Murphy, K. R., & Russell, C. J. (2017). Mend it or end it: Redirecting the search for interactions in the organizational sciences. *Organizational Research Methods*, **20**(4), 549–573.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2017). Interest congruence and performance: Revisiting recent meta-analytic findings. *Journal of Vocational Behavior*, **98**, 138–151. https://doi.org/10.1016/j.jvb.2016.11.002.

O'Boyle, E. H.Jr, Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior*, **32**(5), 788–818. https://doi.org/10.1002/job.714.

O'Neill, T. A., Goffin, R. D., & Gallatly, 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.

O'Neill, T. A., McLarnon, M. J. W., & Carswell, J. J. (2015). Variance components of job performance ratings. *Human Performance*, **28**(1), 66–69. https://doi.org/10.1080/08959285.2014.974756.

Park, H.(H), Wiernik, B. M., Oh, I.-S., Gonzalez-Mulé, E., Ones, D. S., & Lee, Y. (2020). Meta-analytic five-factor model personality intercorrelations: Eeny, meeny, miney, moe, how, which, why, and where to go. *Journal of Applied Psychology*, **105**(12), 1490–1529. https://doi.org/10.1037/apl0000476.

Peng, Q., Zhong, X., Liu, S., Zhou, H., & Ke, N. (2022). Job autonomy and knowledge hiding: the moderating roles of leader reward omission and person-supervisor fit. *Personnel Review*, **51**(9), 2371–2387.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, **98**(1), 114–133. https://doi.org/10.1037/a0030887.

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**(5), 959–981.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, **85**(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x.

Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben-Porath, Y. S., Lees-Haley, P., Green, P., & Greve, K. W. (2011). *A misleading review of response bias*. Comment on McGrath, Mitchell, Kim, and Hough.

Sackett, P. R. (2014). When and why correcting validity coefficients for interrater reliability makes sense. *Industrial and Organizational Psychology*, **7**(04), 501–506.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, **107**(11), 2040–2068.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, **76**(3), 323–346. https://doi.org/10.1348/096317903769647201.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, **88**(6), 1068–1081. https://doi.org/10.1037/0021-9010.88.6.1068.

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology*, **11**(3), 299–324. https://doi.org/10.1080/13594320244000184.

Salgado, J. F., Moscoso, S., & Anderson, N. (2016). Corrections for criterion reliability in validity generalization: The consistency of Hermes, the utility of Midas. *Journal of Work and Organizational Psychology*, **32**(1), 17–23.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage publications.

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**(2), 445–464. https://doi.org/10.1111/j.1744-6570.1999.tb00169.x.

Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, **61**(4), 827–868. https://doi.org/10.1111/j.1744-6570.2008.00132.x.

Schmidt, J. A., O'Neill, T. A., & Dunlop, P. D. (2021). The effects of team context on peer ratings of task and citizenship performance. *Journal of Business and Psychology*, **36**(4), 573–588. https://doi.org/10.1007/s10869-020-09701-8

Schmitt, N. (2007). The value of personnel selection: Reflections on some remarkable claims. *Academy of Management Perspectives*, **21**(3), 19–23.

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, **85**, 956–970.

Shen, W., Cucina, J., Walmsley, P., & Seltzer, B. (2014). When Correcting for Unreliability of Job Performance Ratings, the Best Estimate Is Still .52. *Industrial and Organizational Psychology*, **7**(4), 519–524.

Stanek, K. C. (2014). *Meta-analyses of personality and cognitive ability. Unpublished doctoral dissertation*. University of Minnesota.

Steel, P., Beugelsdijk, S., & Aguinis, H. (2021). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews. *Journal of International Business Studies*, **52**(1), 23–44.

van Vianen, A. E. (2018). Person-environment fit: A review of its basic tenets. *Annual Review of Organizational Psychology and Organizational Behavior*, **5**, 75–101. https://doi.org/10.1146/annurev-orgpsych-032117-104702.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, **90**(1), 108.

Wai, J., & Rindermann, H. (2017). What goes into high educational and occupational achievement? Education, brains, hard work, networks, and other factors. *High Ability Studies*, **28**(1), 127–145. https://doi.org/10.1080/13598139.2017.1302874.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, **15**(1), 134–161.

Cite this article: Foster, J. L., Steel, P., Harms, P. D., O'Neill, T. A., & Wood, D. (2024). Selection tests work better than we think they do, and have for years. *Industrial and Organizational Psychology* **17**, 269–282. https://doi.org/10.1017/iop.2024.10

https://doi.org/10.1017/iop.2024.10 Published online by Cambridge University Press