Organizational Psychology, 2022, Vol. 12, No. 4, P. 123–143.
DOI: 10.17323/2312-5942-2022-12-4-123-143

ORGANIZATIONAL PSYCHOLOGY

www.orgpsyjournal.hse.ru

Revisiting performance appraisal measurement: Construction and validation of Rater Bias Scale

Oladimeji Jamiu ODETUNDE
University of Lagos, Lagos, Nigeria

Abstract. Given the importance of the performance appraisal system and despite continued discussions and efforts to ensure the quality of supervisor rating in the appraisal system, rater effects are still pervasive and continue to undermine the validity and credibility of the appraisal system. *Purpose*. This study aimed to develop and validate Rater Bias Scale for the detection and control of rater effects in performance appraisal system. *Method*. Extensive review of literature was conducted. Separate data were collected for exploratory and confirmatory factor analyses. Samples consisted of diverse academics and professional managers who have been involved in performance ratings in their career. Findings. The exploratory and confirmatory factor analyses conducted on the data revealed eight major dimensions of rater bias with 24 items, three items for each dimension. The scale demonstrates high content reliability and convergent and discriminant validities. *Value of the results*. Using configural and metric invariance, as well as chi-square and one-way ANOVA, the scale displays conceptual equivalence across diverse group. *Implications for research and practice*. The study has important implications for detecting and controlling supervisor errors in performance appraisal for research and practice. The study contributes to the literature and practice of human resource management in organisation. Limitations of the study are highlighted and potential problem in its administration are discussed.

Keyword: performance appraisal system, rater bias scale development, exploratory and confirmatory factor analysis.

Introduction

Job performance measurement is an integral component in test validation research, assessing training needs and evaluating training programmes, promotion and succession planning, salary administration, recruitment, selection and placement, and developmental feedback toward performance maintenance and improvement (Smither, London, Reilly, 2005). The most frequently used measure of employee performance measurement is performance ratings. Ratings of job performance are intended to reflect the proficiency with which employees perform behaviours that contribute toward organizational goals (Campbell, 2012). Though several systematic methods and approaches are available to assess employee performance, the most prominent source remains the incumbents' immediate supervisors.

Supervisors' ratings of subordinates' job performance are however as pervasive as they are notoriously biased or subjective, making job performance rating a major problem. Biased appraisals

Address: University Road Lagos Mainland Akoka, Yaba, Lagos, Nigeria. **E-mail**: oodetunde@unilag.edu.ng

hurt both organizations and employees. Because supervisors differ in their ability to manage subordinates, it reflects in their ratings of employees and affects how their subordinates perform on the job (Lazear, Shaw, Stanton, 2015), and consequently undermine the performance management system. If employee performance is misidentified, flawed information means adjustments cannot be made and the organization's own performance is undermined. Inaccurate performance rating resulting from supervisor rating errors also impedes attempts to appropriately reward performance, undermine employees' sense of fairness and trust in their organizations (Mayer, Davis, 1999). They may also lead to compensating and promoting employees who are less qualified, as well as affect task assignment, employee retention (Frederiksen, Lange, Kriechel, 2018) and career advancement. Indeed, rater bias is considered a substantial source of error within psychological research (Hoyt, 2000) as research findings based on biased supervisor ratings of subordinates may be seriously confounded. Yet, despite many years of research targeting the best ways to address this problem, it has defied all solutions, prompting calls to jettison the use of supervisor rating or the entire performance appraisal system as they are incapable of being used to objectively determine and improve employee performance (Elder et al., 2005).

As ratter's errors continue to constitute major threat to internal validity of performance rating instrument as well as decisions based on such ratings, it will be a huge error to continue to rely on such performance appraisal methods that have been adjudged inadequate and sometimes invalid. Some research and practical attempts have been made to mitigate the negative consequences of rater error in performance appraisal with only marginal success prompting calls for the scrapping of the performance appraisal system, especially the supervisor rating scale. Despite these calls, performance rating of subordinates by supervisors continues to be widely adopted. As a result, a new method of detecting and controlling supervisor ratings error is warranted in order to establish some confidence in the subordinate performance appraisal system. The major aim of this study therefore is to develop and validate a supervisor rating error scale to detect and control performance rating errors. This study therefore contributes to both research and practice efforts to find lasting solution to the enduring rating errors in performance appraisal.

Performance rating errors

Performance measurements require the supervisors to objectively reach a conclusion about the performance of their subordinates. The use of ratings assumes that the supervisor is substantially objective and accurate. However, in reality, supervisors' memories are quite fallible, and supervisors subscribe to their own sets of attributes and expectations by assigning ratings that reflect their own patterns of rating behaviour rather than the actual performance of the target employee or are influenced by extraneous factors that bear no relevance to their subordinates' performance (cf. Kumar, 2005). These biases are commonly viewed as a source of systematic variance in observed ratings that is associated with the raters and not the ratees (Myford, Wolfe, 2003). In other words, rater effects are irrelevant to the construct being assessed through ratings and, thus, threaten the validity of the assessment procedure (Weir, 2005). They produce rating errors, or deviations between the "true" rating an employee deserves and the actual rating assigned. Therefore, rating errors reduce the reliability, validity, and utility of performance appraisal systems.

The general dissatisfaction with supervisor appraisal on subordinate performance stem from many reasons. L. R. Gomez-Meijia with colleagues pointed out that most surveys of raters, ratees, and even human resource professors find that people generally believe the performance appraisal process is unsuccessful (Gomez-Meijia, Balkin, Cardy, 2004). Appraisals are conducted typically based on supervisors' subjective judgments rather than on objective indicators of performance, leading many people to believe that such appraisals are full of errors. However, despite extensive evidence of their inaccuracy, there is a continued reliance on supervisors' ratings in performance assessment. Supervisory ratings of performance persist because the alternatives are worse and according to J. L. Pearce and Q. L. Xu, there are only a few jobs with comprehensive objective measures of performance (Pearce, Xu, 2012).

Early approaches to the problem of supervisory rating inaccuracy treated it as a problem of either supervisor's flawed perceptions of reality, or of their poor assessments of their subordinates' individual differences in performance. Both approaches drew on substantial research in psychology on person perception and psychometrics but proved unable to do much to improve supervisors' ratings accuracy in practice. More systematic measurement systems and better supervisory training in assessing individual differences based on these theoretical explanations have had only limited effects on ratings accuracy. Others have tried to improve the accuracy of supervisor's ratings by triangulating them with others' ratings (Pearce, Xu, 2007). However, peer ratings are as less accurate than supervisors' ratings (Motowidlo, 1982).

For most jobs, an employees' performance needs to be assessed via a retrospective assessment of job effort, judgment, and performance after the fact (Pearce, 1987). The persistent use of supervisory ratings of performance, though flawed, strongly suggests that there is no viable alternative to them yet. Their inaccuracy remains an important practical problem that can be addressed by better approach. There is strong evidence that supervisor ratings inaccuracy is not random, but the result of systematic errors.

Types of performance rating error

Research on performance rating accuracy and the development of accuracy criteria focusing on the psychometric quality of the performance ratings have shown that performance appraisal is fraught with biases and errors. The many psychometric biases, called rater error (bias) is defined as appraisal errors and biases in judgment observations which influences rating score variance (Hoffman et al., 2010; Javidmehr, Ebrahimpour, 2015). Most common and mostly investigated of these biases are leniency (severity), central tendency, halo (horn), primacy (recency), and similarity with the assumption that these implied a lack of accuracy (Appelbaum, Roy, Gilliland, 2011). While other rater errors exist such as contrast error, logical error, proximity error, order effects, they are less commonly studied (Myford, Wolfe, 2003) and substantially share elements of the commonly studied ones.

Primacy and recency errors

Primacy and recency errors occur when a rater overemphasises an employee's previous or most recent behaviour. First impression bias also known as primacy effect occurs when a rater is overwhelmed by the outcome of first assignment of a ratee that this impression is carried over in the memory for long. In most cases, this impression beclouds the rater and makes retrieval from memory of more recent performance information difficult. The opposite direction occurs with recency bias. Most supervisors do not have the time or resources to closely monitor an employee's performance over a year or make detailed notes. Before the appraisal, the rater is forced to consult memory, which is clearer and more dependable in the months leading up to review, as opposed to the earlier part of the rating period. As such many managers tend to rate employee's job performance based on a "what has he done lately" mindset instead of basing assessment on performance from earlier in the year. Consequently, achievements and events that happened lately tend to bear more influence on the employee's performance rating than achievements and events from earlier in the evaluation period. Ratings that unduly reflect recent events can present a false picture of an employee's job performance during the entire rating period. For instance, the employee may have received a poor rating because he or she performed poorly during the most recent month, despite an excellent performance during the preceding eleven months. This may result since past events cannot be remembered objectively. Thus, recent events tend to overshadow the overall performance as a result of "short memories". Therefore, a person who has worked very hard and excelled throughout the year, but for some inadvertent reasons had faced performance issues in the last weeks or month may at times get a poor appraisal from the supervisor, showing a recency bias.

Leniency and severity errors

Leniency and severity describe a situation where ratings are rather too high (leniency) or too low (severity) than warranted. Raters are considered lenient or severe when they systematically assign higher-than-expected or lower-than-expected ratings, respectively, than is warranted by the quality of performances (Wind, Engelhard, 2017). Leniency is conceptually defined as the rater ratings that are well above the midpoint in the evaluation scales used as indicated by average ratings over all ratees. The tendency in lenient rating is to evaluate everyone favourably similarly. However, if everyone gets a similar above-average score on a generic performance evaluation, the rating process has limited value (Eckes, 2005). The downside of this error is that even poor performers may get good ratings, and this could create resentment among good performers. Such a bias is undesirable since it results in subordinates appearing to be more competent than in fact they are (Wildman, Bedwell, Salas, Smith-Jentsch, 2010).

On the other extreme side of leniency is severe rating where a rater constantly gives out low ratings. Severe raters can also cause problems in making valid inferences from assessment results. Because both lenient and severe raters introduce a similar kind of problems to performance rating, researchers generally use the term 'leniency error' to apply to a general, constant tendency of a rater to rate too high or too low for whatever reasons (Myford, Wolfe, 2003).

Central tendency error

Central tendency is the raters' unwillingness to give ratings in either favourable or unfavourable direction or avoiding using high or low ratings even when they are warranted. A rater assesses a disproportionately large number of ratees as performing in the central part of a distribution of rated performance, in contrast to their true level of performance (Muchinsky, 2006). The rater may believe that all the employees are equal, and do not want to rock the boat. The result is a failure to reflect the true range of differences among the employees. For example, when rating subordinates on a scale that ranges from one to five, an appraiser would avoid giving any 1s or 5s. When this error occurs, all employees end up being rated as average or near average, and the employer is thus unable to discern who its best and worst performers are. Common reason deduce for this type of rating is to 'play safe' and avoid the necessity to justify scoring across the two extremes as some systems expect managers to specify additional comments as they give too high or too low ratings to employees.

Halo and horn effects

Halo and horn effects are types of cognitive bias whereby our perception of someone is positively or negatively influenced by our opinions of that person's other related traits (Perera, 2021). Halo error is the tendency to focus on the global impression of each ratee rather than to carefully differentiate among specific levels of different performance dimensions (Martin, Bartol, Kehoe, 2010). This rating error conceptually implies that the rater depends on a general view of the ratee. Halo error occurs when the rater perceives one factor as having paramount importance and gives a good rating to an employee on all other dimensions based on this one factor. One patent attribute of a certain person leads an observer to draw a generalising conclusion about that person (Ellis, 2018). The rater fails to discriminate between the employee's strong points and weak points; and the halo is carried over from one dimension to the others. For example, the overall impression of an employee is based on a particular characteristic, such as intelligence or appearance.

The definition of horn effect is already subsumed in the definition of halo effect as the opposite of halo effect. It occurs when an employee has one hindering weakness, and the manager allows this to seep into other rating categories or the overall outcome of the employee's performance appraisal. For example, if an employee is especially weak in the category of "customer satisfaction," it is not necessarily true that the employee may need to make improvement in the categories of "job knowledge" or "problem-solving." Similarly, the horn effect may cause us to stereotype that someone who is physically overweight is also lazy, although there is no evidence to indicate that morality is tied to appearance (Perera, 2021). Thus, while the halo effect works in the positive direction, horn works in the negative direction. A single positive quality of a person may induce a positive predisposition toward every aspect of the person while one negative attribute of that person may induce an overall negative impression of the person (Perera, 2021).

Similarity effect

Similarity effect reflects a tendency of supervisors to judge their subordinates more favourably who they perceive as similar to themselves. The more similar subordinates are to their supervisors in attitude, background, demographic, and other characteristics, etc. the higher the subordinates' performance ratings (Baltes, Bauer, French, 2007). Related to this, A. Zahed and F. S. Ardabili opine that majority of managers prefer to employ assistants and subordinates with whom they have many similarities in terms of religion, education, ideas and traits and in such cases, the differences in ideas and attitudes are exacerbated in the organisation and those who are similar to the manager are appointed to management workgroup (Zahed, Ardabili, 2017). This will reflect in the performance appraisal of such employees by the manager. In the views of H. Alves with colleagues with the similarity effect, managers attract and favour subordinates with whom he shares some similarities (Alves et al., 2016). Research finds that evaluators of all kinds tend to rate those more similar to themselves more positively than those who are dissimilar (e.g., Cascio, Aguinis, 2004). Researchers also argued that high perceptions of (ethnic / racial) similarity tend to elicit favourable responses such as interpersonal attraction, perceptions of procedural fairness, and increased job satisfaction (Avery, 2003).

A brief overview of previous efforts to address performance rating error

Over the decades, researchers have made several efforts to address errors in performance appraisal. Four of the most widely used methods are briefly reviewed here. They include rater training, rating scale methods, behaviourally anchored scale, 360-degree feedback methods, and management-by-objective. Rater training programmes generally aim to influence ratings by educating raters about key cognitive and observational demands of the rating process. They focus on reducing rating errors, increasing accuracy and providing raters with a common frame of reference to attain rater reliability (Elder et al., 2005; Roch, Woehr, Mishra, Kieszczynska, 2012). However, results of the effect of rater training on rater errors have been mixed. For example, some studies provide support for the positive effects of rater training in lessening rater errors (e.g., Kang, Rubin, Kermad, 2019; Bijani, 2018; Davis, 2016). Other studies however show that it is impossible to fully eliminate rater error even after training (e.g., Elder et al., 2005) as raters who want some employees to get a raise still, irrespective of rater error training, gave higher rating to them to achieve it.

Graphical rating scales (GRS) consist of a checklist of job performance dimensions, job-relevant human traits (e.g., cooperation, flexibility, initiative, sociability), or both, accompanied by an evaluative continuum (e.g., below average to outstanding, very high to very low) upon which supervisors are asked to indicate their judgments about target employees. The quantitative nature of the rating scheme makes it more standardised and objective than the other rating schemes which are more susceptible to rater subjectivity. However, despite their level of objectivity, they are still susceptible to rater bias like halo (horn), leniency (strict) and central tendency error. This is because the behavioural traits measured by the scale are not standardised, are vague or ambiguous giving room for different interpretations, thereby leading to subjective judgements by different users (cf. Klieger et al., 2018).

Most Behaviourally Anchored Rating Scales (BARS) consist of a set of five to ten vertical scales, with each scale representing a major performance dimension of the job which is anchored by five or more critical incidents that reflect highly effective to highly ineffective observable job behaviours relevant to the job dimension under consideration. Scale values are assigned to the critical incidents, which correspond to the approximate degree of effectiveness with the highly effective behaviour being assigned the highest value on the scale. The critical incidents are defined by observable job-related behaviours and reflect various levels of desirable performance. However, despite the findings of studies that BARS has high reliability and validity in the evaluation of job performance (e.g., Debnath, Lee, Tandon 2015; Matosas-López, Leguey-Galán, Doncel-Pedrera, 2019), leniency rating bias and lack of discriminant validity between performance dimensions are still major drawbacks (Kingstrom, Bass, 1981).

The 360-degree appraisal system involves the use of many assessors of some specific sets of employee work behaviour. These may include self-assessment, immediate supervisor assessment, subordinate assessment, and peer assessment (Grund, Przemeck, 2012) and everyone else who is directly work-related to the ratee. The underlying premise that the use of 360-degree performance appraisal is that with multi-sources of assessment, significant amount of information of performance of employees can be obtained (Sahoo, Mishra, 2012). 360-degree method ensures that subjectivity and bias inherent in a single supervisor performance assessment of employee is controlled and thus ensuring a more valid rating outcome; 360-degree appraisal is more accurate and more reflective of employee performance (Espinilla et al., 2013; Sahoo, Mishra, 2012). However, despite the effectiveness of the appraisal system, only modest ratings improvements over time and not among all those appraised have been found (e.g., Smither, London, Reilly, 2005; Atwater, Brett, Charles, 2007). The 360-degree ratings are often subject to biases: subordinates typically providing overly lenient ratings of their supervisors for fear retaliation (Smith, Fortunato, 2008), employees providing more lenient and less reliable self-ratings (e.g., Atwater et al., 2007), and supervisors engaging in centrality bias and leniency bias (Bol, 2011).

Management by objectives

Management by objectives (MBO) entails the subordinates participating in setting short-term performance which are discussed with the superior and performance evaluated against these goals (Stein, 2010). The extent to which the employee is able to implement the action plan and achieve the objectives and set goals are then appraised by key stakeholders like the subordinates, supervisors and the employees themselves. This allows for appraisal of performance in an objective manner. P. Drucker underscores that the approach when correctly implemented helps in establishing a performance appraisal system that is based on efficiency and fairness and promotes objectivity of the appraisal system (Drucker, 2013). A critical review of literature however suggests that MBO in performance appraisal has its drawbacks. It is not applicable to all jobs and interpretation of goals may vary from manager to manager which may lead to subjectivity in ratings of subordinates' performance (Aggarwal, Thakur, 2013; Dagar, 2014).

Thus, while studies have provided some evidence of improvements in the qualities of the performance appraisal methods with the various previous attempts to control rater bias, other studies found only negligible or partial support, and yet others found no support at all, prompting continued lack of confidence in the performance appraisal system. Major reason for this is the vulnerability of the measures to both intentional and inadvertent rater errors, thus warranting more efforts to improve the performance appraisal system. The present is another effort directed to developing and constructing scale to detect and control rater bias.

Methods

Participants

A total of 308 supervisors/managers across 15 service and manufacturing organisations in Lagos and Ibadan in the South-West Nigeria were accidentally selected and used in the validation exercise. From the service sector, three organisations each were from education and telecommunication, two each from finance and broadcasting, while one each was from communication and advertising. From manufacturing, two organisations were from and publishing and one each from power and diary. The supervisors/managers are distributed as follows: Education — 94 (30.5%), Telecommunication — 42 (13.6%), Finance — 48 (15.6%), Broadcasting — 54 (17.5%), Publishing — 18 (5.8%), Communication (Advertising) — 4 (1.3%), Power — 34 (11%), and Diary — 14 (4.5%). There were 242 (78.6%) male and 66 (21.4%) female supervisors (managers). Their education ranged from National Diploma or National Certificate in Education — 22 (5.1%), Higher National Diploma or B.SC — 156 (50.6%), Postgraduate Degree (Diploma) — 134 (43.5%) and Professional Certificate — 2 (.06%). The supervisors/managers occupied lower management — 70 (22.7%), middle management – 180 (58.4%) and top management — 58 (18.8%) levels. Fifty (16.2%) of the managers have appraised and rated subordinates for promotion (advancement) not less than twice and 72 (23.4%) for few times and 186 (60.4%) many times.

Measures

The proposed performance rating bias scale consisted of the 32 items with four items each on the eight commonly identified dimensions, namely, primacy and recency, leniency and severity, central tendency, halo and horn, and similarity effects. A five-point Likert response scale ranging from "1" (strongly disagree) to "5" (strongly agree) was adopted. Higher values indicate greater possession of the bias measured.

Procedure

Managers (supervisors) were purposively selected and administered with the survey instrument in each of the 15 service and manufacturing companies during their office hour after obtaining the permission from the Human Resources Departments to collect the data. One to three visits were made to each of the companies. The first visit was to distribute and collect administered survey instrument from the respondents. Second and third visits were made to retrieve the survey instruments from those who requested to take them home owing to being unable to complete them in the office. A total of 336 survey instruments were collected in eight weeks with only 308 of them usable.

Results. Development and construction of Performance Rater Bias Scale

Domain specification and item generation

Literature reviewed above, comprising theories and empirical studies on performance appraisal provided the basis for the identification of the dimensions of rating bias. Several dimensions were identified, and subsequent analysis revealed that many of these dimensions are conceptually related or overlapping and eventually eight most common biases were identified as primacy and recency effects, leniency and severity effects, central tendency effect, halo and horn effects, and similarity effect. Initial pool of 80 items were generated, comprising 10 items for each of the eight dimensions. Editing and eliminating redundant statements reduced the item-pool to 48 items with six items on each dimension, all of which have the desirable property of being worded in the common rating bias parlance rather than formal academic language.

Scale purification

A twelve-person judgmental panel was used to provide indication of face and content validities of the scale (Nunnally, Bernstein, 1994). They comprised senior faculty members of the Department of Psychology of a Nigerian university who are industrial psychologists (with position ranging from senior lectureship to full professorship), hold PhD degrees, and possess expert knowledge and professional experience on the subject matter. It is a common practice to use expert faculty members who are knowledgeable about concept domain to rate items on content validity at this initial stage of scale development. Subject expert judges are preferrable over target population judges for content validation of scales, but ideal to combine them. With resource constraint, however, it is recommended, at least, the use of expert judges (Boateng et al., 2018; DeVellis, 2012; Morgado et al., 2018).

They rated each of the 48 items on the extent to which they agree that each item on the eight dimensions is adequate to measure each dimension on a five-point Likert scale ranging from “1” (Strongly disagree) to “5” (Strongly agree). To adequately rate the items, the panel members were provided with the conceptual definitions of each dimension of rater bias. Ratings were summed for each item. An a priori decision rule specified retaining only items with average rating score of four and above as relevant and adequate to measure the construct. Thirty-two items satisfied this rule. Sixteen items were eliminated subsequently because they were rated below four and as such considered inappropriate and inadequate to measure the construct. The remaining 32 items with four items on each of the eight dimensions were subjected to further validation tests to ascertain their psychometric adequacy and construct validity.

Descriptive statistics

The descriptive statistics of the performance rater bias scale are shown in Table 1 below. The means and the standard deviations of the items ranged from 1.94 to 3.88 and 0.86 to 1.30 respectively. Given the five-point Likert rating scale used and the recorded average mean score of 2.95 (approx. 3.00) and compared with the lower standard deviations recorded, the biases were observed among the supervisors/managers some of the times.

Table 1. Descriptive Statistics of the Performance Rater Bias (Error) Scale

Dimensions and Items	Total Score	*M*	*SD*
Primacy Effect			
1. I believe that outcome of first assignment should guide the rating of anyone	377	2.45	1.30
2. Outcome of first assignment influences my general impression of anyone	513	3.33	1.11
3. Anyone who fails on first task may never succeed on other tasks	598	3.88	1.11
4. I never doubt the ability of anyone who succeeds on first assignment	522	3.39	1.18

Recency Effect			
5. I am more interested in current performance in rating anyone	299	1.94	.86
6. Irrespective of past performances, I will rate anyone low if current performance is unimpressive	499	3.24	1.12
7. Notwithstanding past performances, I will rate anyone high if current performance is impressive	376	2.44	1.24
8. Even if past punctuality record is poor, I will rate anyone high if current punctuality record is satisfactory	376	2.44	1.13
Leniency Effect			
9. I do not believe it is right to fail anyone whose performance is below expectation	475	3.09	1.10
10. I believe anyone whose performance fall short of expectation deserves commendation	343	2.23	1.04
11. Everyone in a great work team deserves to be rated equally high	343	2.23	1.11
12. Anyone who fails to meet expectations should be encouraged with good rating	353	2.29	1.08
Severity Effect			
13. I will reward only people who give their best performances at all times	408	2.65	1.14
14. Only exceptional performances interest me	414	2.69	1.15
15. I will take nothing short of excellent performances from anyone	435	2.82	1.21
16. Performances not outstanding are unacceptable to me	597	3.88	1.16
Central Tendency Effect			
17. I believe that an average score is fair for everyone	491	3.19	1.21
18. I avoid rating performance extremely high or extremely low	503	3.27	1.23
19. I believe it is safer to score people in the middle position	339	2.20	1.12
20. I do not believe anyone can be exceptionally good or exceptionally bad	439	2.85	1.25
Halo Effect			
21. Anyone who succeeds in a critical role will most certainly succeed in other roles	483	3.14	1.29
22. I consider anyone who succeeds on a difficult task as generally competent	438	2.84	1.16
23. I believe that meticulous people tend to do well at work	424	2.75	1.13
24. People with pleasing personality tend to be successful people	505	3.28	1.10
Horn Effect			
25. Anyone who fails in a critical role will most certainly fail in other roles	489	3.12	1.26
26. I consider anyone who fails on a difficult task as generally incompetent	467	3.42	1.18
27. I believe late comers never do well on work assignment	354	2.63	1.07
28. People without pleasant personality are not usually successful people	524	3.86	1.09
Similarity Effect			
29. I like to work with people who relate well with me	405	2.63	1.18
30. I prefer people who share my beliefs and values	569	3.69	1.14
31. I like people who share similar characteristics with me	500	3.25	1.07
32. I do not like anyone whose attitudes irritate me	338	2.19	.94

Exploratory factor analysis

Prior to conducting exploratory and confirmatory factor analyses to validate underlying factor structure of the scale, Kaiser — Meyer — Olkin Measure of Sampling Adequacy (KMO-MSA) and the Barlett's Test of Sphericity (BTS) were first examined, as suggested by Tabachnik and Fidell (2013), to test the appropriateness of conducting factor analysis on the correlation matrix of the scale. Significant Bartlett's Test of Sphericity (p < 0.001) and a KMO-MSA which ranged from 0.662 to 0.704, which were above the recommended 0.60 (Tabachnick, Fidell, 2007), confirmed that the correlation matrix and the sample size were adequate and suitable for factor analysis.

Exploratory factor analysis (EFA), using principal component analysis (PCA) with Varimax rotation, was conducted to extract the factor structure. Exploratory factor analysis is most helpful for identifying items that load on their respective factors as hypothesised. Varimax rotation loaded items on 14 factors as shown in Table 2. Four items loaded on Factor 1, three each on Factors 2, 3, 4, 6 and 7, two each on Factors 5, 8, 9, 10 and 12, and one each on Factor 11, 13, and 14. Three items each loaded highest on four factors which were the original factors for which they were drawn. They are retained under these factors, viz.: Factor 1 (Similarity Effect), Factor 3 (Primacy Effect), Factor 4 (Leniency Effect), and Factor 6 (Recency Effect). Two items each loaded highest on three factors which were the original factors on which they were and are also retained under these factors, viz: Factor 2 (Halo Effect), Factor 8 (Central Tendency), and Factor 11 (Horn Effect).

Table 2. Rotated component matrix and item-total correlation for the scale

Factor	1	2	3
Factor 1			
1. I prefer people who share my beliefs and values	.77	.68	Similarity
2. I like people who share similar characteristics with me	.74	.69	Similarity
3. *Performances not outstanding are unacceptable to me	.51	.68	Severity
4. I do not like anyone whose attitudes irritate me	.73	.69	Similarity
Factor 2			
5. Anyone who succeeds in a critical role will most certainly succeed in other roles	.59	.68	Halo
6. I consider anyone who succeeds on a difficult task as generally competent	.71	.68	Halo
7. *Even if past punctuality record is poor, I will rate anyone high if current punctuality record is satisfactory	.75	.69	Recency
Factor 3			
8. I believe that outcome of first assignment should guide the rating of anyone	.46	.69	Primacy
9. Outcome of first assignment influences my general impression of anyone	.73	.69	Primacy
10. I never doubt the ability of anyone who succeeds on first assignment	.54	.68	Primacy
Factor 4			
11. Anyone who fails to meet expectations should be encouraged with good rating	.76	.71	Leniency
12. I believe anyone whose performance fall short of expectation deserves commendation	.70	.70	Leniency
13. Everyone in a great work team deserves to be rated equally high	.45	.69	Leniency
Factor 5			
14. *I believe it is safer to score people in the middle position	.66	.68	Centr.Tend.
15. **I will take nothing short of excellent performances from anyone	.78	.67	Severity
Factor 6			
16. *I am more interested in current performance in rating anyone	.49	.69	Recency
17. Irrespective of past performances, I will rate anyone low if current performance is unimpressive	.83	.69	Recency
18. Notwithstanding past performances, I will rate anyone high if current performance is impressive	.60	.69	Recency
Factor 7			
19. **I will reward only people who give their best performances at all times	.72	.69	Severity
20. *I do not believe it is right to fail anyone whose performance is below expectation	.70	.68	Leniency
21. *I believe that meticulous people tend to do well at work	.43	.69	Halo
Factor 8			
22. I avoid rating performance extremely high or extremely low	.75	.69	Centr. Tend.
23. I do not believe anyone can be exceptionally good or exceptionally bad	.67	.68	Centr. Tend.
Factor 9			
24. **I believe that an average score is fair for everyone	.78	.70	Centr. Tend.
25. **People with pleasing personality tend to be successful people	.67	.68	Halo
Factor 10			
26. *I like to work with people who relate well with me	.72	.68	Similarity
27. *Anyone who fails on first task may never succeed on other tasks	-.46	.69	Primacy
Factor 11			
28. Anyone who fails in a critical role will most certainly fail in other roles	.71	.70	Horn
29. I consider anyone who fails on a difficult task as generally incompetent	.69	.68	Horn
Factor 12			
30. **Only exceptional performances interest me	.74	.71	Severity
Factor 13			
31. People without pleasant personality are not usually successful people	.63	.54	Horn
Factor 14			
32. I believe late comers never do well on work assignment	.61	.52	Horn

***Note*:** 1. Varimax Rotated Component Matrix; 2. Item-Total Correlation; 3. Original Factor for which item was generated; * — Item deleted from scale for cross-loading; ** — Item moved to original factor.

Further analysis reveals that the stand-alone single items that loaded on the other factors, loaded next highest on their original factors. Items 15 and 28 which loaded separately on Factors 5 and 11 and item 3 which loaded under Similarity Effect loaded next highest under Severity Effect (0.41, 0.40, and 0.48, respectively) and were subsequently taken together under Severity Effect. Item 24 which loaded on Factor 9 also loaded next highest on its original factor (Central Tendency Effect) and therefore retained on the factor. Thus, a total of 24 items with three items each recording the highest reliability and validity were identified for the eight factors for the performance rater bias scale in the validation exercise. The eight factors are similarity, halo, primacy, leniency, recency, central tendency, horn, and severity effects each having three items, with least eigen value of 3.56 for all eight factors. B. G. Tabachnick and L. S. Fidell suggested retaining at least three items per factor (Tabachnick, Fidell, 2007). Results of item-total correlation and factor analysis are shown in Table 2. Items and subscales extracted with Varimax rotation are shown in Table 3. Five-point Likert response format ranging from strongly disagree ("1") to strongly agree ("5") is adopted for the scale.

Table 3. Subscales and Items Identified with Varimax Rotation

Subscales and Items	**1**	**2**
Primacy Effect		
1. Outcome of first assignment influences my general impression of anyone	.73	.68
2. I never doubt the ability of anyone who succeeds on first assignment	.54	.68
3. I believe that outcome of first assignment should guide the rating of anyone	.46	.69
Recency Effec		
4. Irrespective of past performances, I will rate anyone low if current performance is unimpressive	.83	.69
5. Notwithstanding past performances, I will rate anyone high if current performance is impressive	.60	.69
6. I am more interested in current performance in rating anyone	.49	.69
Leniency Effect		
7. Anyone who fails to meet expectations should be encouraged with good rating	.76	.71
8. I believe anyone whose performance fall short of expectation deserves commendation	.70	.70
9. Everyone in a great work team deserves to be rated equally high	.45	.69
Severity Effect		
10. I will take nothing short of excellent performances from anyone	.78	.67
11. Only exceptional performances interest me	.74	.71
12. I will reward only people who give their best performances at all times	.72	.69
Central Tendency Effect		
13. I avoid rating performance extremely high or extremely low	.75	.69
14. I do not believe anyone can be exceptionally good or exceptionally bad	.67	.68
15. I believe that an average score is fair for everyone	.42	.70
Halo Effect		
16. I consider anyone who succeeds on a difficult task as generally competent	.71	.68
17. Anyone who succeeds in a critical role will most certainly succeed in other roles	.59	.68
18. People with pleasing personality tend to be successful people	.40	.68
Horn Effect		
19.Anyone who fails in a critical role will most certainly fail in other roles	.71	.70
20. I consider anyone who fails on a difficult task as generally incompetent	.69	.68
21. People without pleasant personality are not usually successful people	.63	.54
Similarity Effect		
22. I prefer people who share my beliefs and values	.77	.68
23. I like people who share similar characteristics with me	.74	.69
24. I do not like anyone whose attitudes irritate me	.73	.69

Note: 1. Factor Loading on Varimax Rotation; 2. Item — Total Correlation.

Confirmatory Factor Analysis

To examine and quantify the goodness-of-fit of the 8-factor structure identified on the EFA, a new set of data was collected, and confirmatory factor analysis (CFA) performed (Morgado et al.,

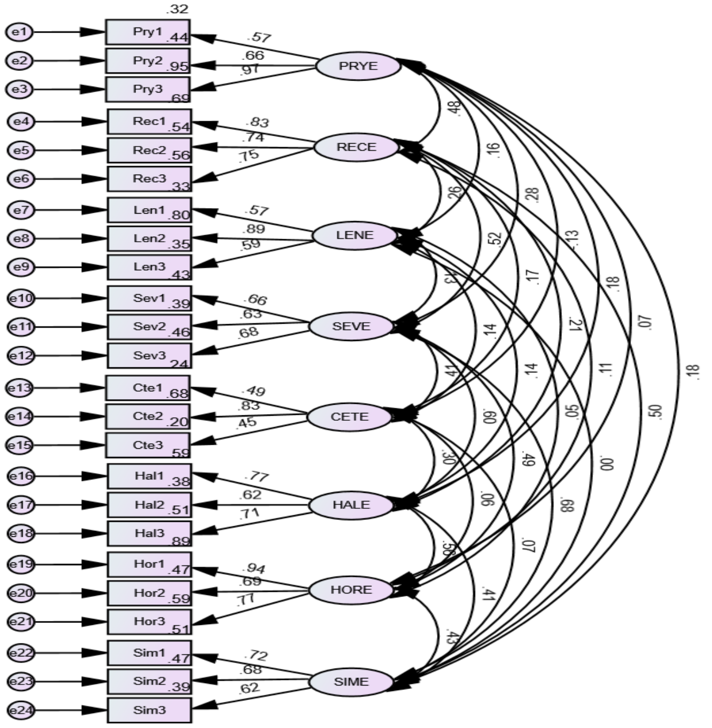
2018) using AMOS 26. The sample consisted of 351 respondents of which academics not lower than senior lecturers constituted 57%, and 43% management professionals not lower than lower-level managers. Male respondents are 69% and 31% are females. They are aged between 32 years and 56 years, have at least a first university degree, and have rated subordinates for promotion, training and development programmes, awards, raises, transfer, etc. at least a couple of times in their career.

Kaiser — Meyer — Olkin Measure of Sampling Adequacy (KMO-MSA) and Barlett's Test of Sphericity (BTS) re-examined for the new dataset, produced significant Bartlett's Test of Sphericity ($p < 0.001$) and a KMO-MSA ranging from 0.75 to 0.94. These were above the recommended 0.60 by Tabachnick and Fidell (2007) and confirmed that the correlation matrix and the sample size were adequate and suitable for factor analysis.

Table 5. CFA Goodness of fit test for performance rater bias scale (N = 351)

Model Fit Value	**Model 1 Observed Values**	**Model 2 Observed Values**	**Recommended Values Between**	**Source**
CMIN / *df*	6.61	2.81	≤ 5.00	Hair, Black, Babin, and Anderson, (2010).
CFI	.65	.94	≥ .90 and 1.00	
TLI	.56	.88	≥ .90 and 1.00	
GFI	.76	.93	≥ .90 and 1.00	
AGFI	.67	.86	≥ .90 and 1.00	
NFI	.61	.92	≥ .90 and 1.00	
SRMR	.12	.07	≤ .08	
RMSEA	.13	.07	.05 - .08	

Note. CFA = Confirmatory Factor Analysis; CMIN/df = Residual Degrees of Freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; NFI – Normed Fit Index; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation.



Fiure 1. Eihgt-factor structure of Performance Rater Bias Scale

As recommended (Byrne, 2010), a one-factor model (Model 1) comprising all the latent variables together on the scale was first tested before the eight-factor model structure obtained in the EFA (Model 2). Retaining the item loading criterion of > 0.40 (Howitt, Cramer, 1997), the least item loading observed was 0.45 (item 3 on central tendency effect). Other item loadings range from 0.49 (item 1 on central tendency effect) to 0.91 (items 3 on primacy effect) which signifies reasonable loading on their respective factors. Various criteria used to examine the fit of Model 1 in Table 5 show that none of the values to determine the model fit attained the recommended value and as such, the one-factor model shows a poor fit. On the other hand, the 8-factor model (Model 2) provides better values and a good fit with the model attaining all the recommended values.

Though the observedvalues of TLI = 0.88 and AGFI = 0.86 fell slightly short of Hair with colleagues' recommended values, some researchers held that values higher than 0.85 on the two criteria are acceptable (Schermelleh-Engel, Moosbrugger, Müller, 2003; Hu, Bentler, 1999).Parsimony-Adjusted Measures J. B. Schreiber with colleagues also affirmed that the measurement model that fulfils the majority of the model fit indexes value is a good and acceptable model fit (Schreiber et al., 2006). Therefore, it can be accepted that the eight-factor structure of Model 2 provides a better fit and thus confirms the hypothesis as derived from the literature and validates the exploratory factor analysis. The validated eight-factor structure performance rater bias scale is represented in Figure 1.

Reliability tests

The eight-factor structure derived from the CFA was subjected to reliability and validity tests. Test-retest reliability, item-total correlation, Cronbach alpha, and composite reliability tests were used to assess the internal consistency of the subscales. Test-retest reliability was 0.76, with item-total correlation coefficients ranging from 0.45 (CETE item 3) to 0.94. (HORE item 1). As shown in Table 6, all the scale's dimensions had Cronbach alphas greater than 0.70, indicating satisfactory reliability (Gadermann, Guhn, Zumbo, 2012), with the exception of CETE (0.62) and SEVE (0.67). Howit and Cramer (1997), however, indicated that reliabilities between –0.26 and +0.26 are significant at .01 level with a sample size as low as 100 and that such scales can be dependent upon to provide consistent and reliable scores. Cronbach alpha for the entire scale is 0.72. Composite reliability values obtained are equal to or higher than 0.70 (Hair, Black, Babin, Anderson, 2010), with the exception of CETE (0.63) which was also accepted as reliable given Loewenthal's (2004) suggestion that composite reliability lower than 0.70, but higher than 0.60 is acceptable to meet the composite reliability condition. Therefore, the scale can be assumed to be reliable.

Validity checks

Convergent and discriminant validities were used to assess the construct validity of the scale. Convergent validity was established with Eigenvalue > 1.00 and item loading of at least 0.40 on posited constructs, as suggested by some researchers (Hinton, Brownlow, McMurray, Cozens, 2004; Straub, Boudreau, Gefen, 2004). A further test of convergent validity was conducted by examining the Average Variance Extracted (AVE). As also evident in Table 6, the AVE values of five dimensions of the scale, ranging from 0.38 (CETE) to 0.49 (LENE and HALE), fell short of the threshold of equal or greater than 0.50 as recommended by some authors (Boateng et al., 2018; Hair et al., 2014).

Discriminant validity was assessed by examining item loadings on each dimension of the construct, which was observed to be higher than 0.40 with no cross-loading of items equal to or higher than the value to confirm discriminant validity as suggested by E. J. Pedhazur (Pedhazur, 1982). Other scholars also suggested an item loading of 0.30 as sufficient to accept a factor structure (Howitt, Cramer, 1997). As can also be observed in Figure 1 and Table 6, the relatively general low correlations (ranging from 0.00 to 0.68) among the dimensions of the construct also attest to the discriminant validity of the scale. To further corroborate this evidence, the square roots of AVE for each dimension of the construct are observed to be higher than the correlations between the dimension and the other dimensions, as suggested by researchers (Hair, Hult, Ringle, Sarstedt, 2014; Henseler, Ringle, Sarstedt, 2015). As asserted by A. O'Cass and L. Ngo, intercorrelations among dimensions of the construct and their respective composite reliabilities (CRs) are observed in Table 6 to be higher than their CRs to further substantiate the discriminant validity of the scale (O'Cass, Ngo, 2007).

Table 6. Descriptive statistics, intercorrelations among scale dimensions, composite reliability (CR) index, average variance extracted (AVE), and square root of the AVE

SN	Dimension	M	SD	1	2	3	4	5	6	7	8	α	CR	AVE
1.	PRYE	1.94	.50	(.76)								.73	.79	.57
2.	RECE	3.40	1.13	.48	(.78)							.77	.82	.60
3.	LENE	2.21	.74	.16	.26	(.70)						.70	.73	.49
4.	SEVE	2.99	.86	.28	.52	.13	(.66)					.67	.70	.43
5.	CETE	2.60	.90	.13	.17	.14	.41	(.62)				.62	.63	.38
6.	HALE	3.40	.85	.18	.21	.14	.60	.30	(.70)			.70	.74	.49
7.	HORE	2.00	.64	.07	.11	.05	.49	.06	.56	(.81)		.80	.84	.65
8.	SIME	3.34	.79	.18	.50	.00	.68	.07	.41	.43	(.68)	.67	.71	.46

Note: PRYE (Primacy Effect), RECE (Recency Effect), LENE (Leniency Effect), SEVE (Severity Effect), CETE (Central Tendency Effect), HALE (Halo Effect), HORE (Horn Effect), and SIME (Similarity Effect), *α* — Cronbach alpha, CR (Composite reliability), AVE (Average Variance Extracted).

Group comparisons on the construct

To explore if the eight dimensions of the construct have conceptually equivalent meaning across groups, measurement invariance using configural and metric invariance was conducted. Configural invariance was used to determine if item loadings on the latent factors differed across the groups by freely estimating all item loadings (parameters) in the different groups and holding the underlying measurement structure constant across the groups. Metric invariance was used to determine if the factor loadings are equivalent across the groups by constraining factor loadings of the items on the constructs to be equivalent across the groups and freely estimating the intercepts. Chi-square and One-way ANOVA were subsequently conducted to determine if there are group mean differences in the construct (Bialosiewicz, Murphy, Berry, 2013; Putnick, Bornstein, 2016).

For the different groups, both configural invariance and metric invariance were conducted. For this purpose, age was categorised into younger and older employees, gender into males and females, and work experience into short and long work experience. The three organisational management positions of lower, middle, and senior levels were collapsed into two, namely lower and senior level positions. This is to align it with the two academic positions involved in performance evaluation in higher education, namely senior lecturers and associate professors (readers, professors). Occupations were categorised into academics and professional managers.

Configural invariance and metric invariance were found for age (CFI = .92, TLI = .85, RMSEA = .08; CFI = .93, TLI = .86, RMSEA = .07), gender (CFI = .94, TLI = .87, RMSEA = .06; CFI = .95, TLI = .86, RMSEA = .06), work experience (CFI = .92, TLI = .85, RMSEA = .08; CFI = .93, TLI = .86, RMSEA = .06), organisational position (CFI = .92, TLI = .85, RMSEA = .08; CFI = .92, TLI = .85, RMSEA = .08), and occupation (CFI = .92, TLI = .85, RMSEA = .08; CFI = .93, TLI = .86, RMSEA = .06). These results demonstrate the measurement invariance of the construct across the different groups: age (younger vs older samples), gender (males vs females), work experience (short vs long), position (lower-level vs higher-level), and occupation (academics vs managers). Constraining the main and cross factor loadings (ΔCFI = .93; ΔTLI = .87; ΔRMSEA = .06), the intercepts (ΔCFI = .93; ΔTLI = .86; ΔRMSEA

= .08), and the variances and covariances (ΔCFI = .91; ΔTLI = .86; ΔRMSEA = .07) did not result in increases outside the expected range (Cheung, Rensvold, 2002).

Model comparison with Chi-square for gender ($\Delta\chi^2 = 35.84$, $\Delta df = 7$, $p < .05$), work experience ($\Delta\chi^2 = 51.69$, $\Delta df = 7$, $p < .05$), and occupation ($\Delta\chi^2 = 25.36$, $\Delta df = 7$, $p < .05$) further indicates that the groups are comparable on the construct. However, model comparison for age ($\Delta\chi^2 = 5.36$, $\Delta df = 7$, $p > .05$) and organisational position ($\Delta\chi^2 = 10.34$, Δdf = 7, $p > .05$) revealed that younger and older samples, as well as lower-level and higher-level samples, are not equivalent on the construct. In contrast, one-way ANOVA revealed conducted no significant differences between the groups: age ($F_{(7;15)} = 3.08$, $p = .069$), gender ($F_{(7;15)} = .25$, $p = .96$), work experience ($F_{(7;15)} = .23$, $p = .97$), organisational position ($F_{(7;15)} = .73$, $p = .66$), occupation ($F_{(7;15)} = .22$, $p = .97$). The between-group equality of indicator intercepts also showed that the means of the indicators were equivalent for all the groups. These results demonstrate the measurement invariance of the instrument across the different groups. Overall, the various results give credence to the comparability of the different groups on the performance rater bias construct and, as such, can be used across groups.

Discussion

Previous efforts at ensuring the validity of performance ratings (e.g., rater training, rating scale methods, behaviourally anchored scale, 360-degree feedback methods, and management-by-objective) have mostly centred on improving the psychometric properties of the rating scales used in performance evaluations. However, only marginal success has been recorded in reducing rating biases and improving the rating outcomes. The present research developed and validated a performance rater bias scale aimed at helping to detect and control rater errors and improving the quality of the employee performance appraisal system. Both exploratory (EFA) and confirmatory factor analysis (CFA) identified eight-factors structure on the performance rater bias scale, with each dimension measured by three items. The 24-item scale (shown in Table 3) has eight dimensions, namely, leniency, severity, central tendency, halo, horn, primacy, recency, and similarity effects.

The reliability of the scale was established through test-retests, item-total correlations, Cronbach alpha, and composite reliabilities (CRs). Construct validity was established with adequate eigenvalues, item loadings on posited dimensions of the scale, and the average variance extracted (AVE) for convergent validity and adequate item loadings on the dimensions with no cross-item loading, square roots of AVE for each dimension, moderately low intercorrelations among the dimensions, and their respective CRs for discriminant validity. Group comparison, a further measure of scale validity, was established with measurement invariance using configural and metric invariances and group mean differences with chi-square and one-way ANOVA. The results show that the groups are equivalent on all the dimensions of the scale, thus validating the generalisability of the scale across groups in the sample.

The construction and validation of this scale have several strengths. It engaged in an extensive review of the literature to provide the basis for very thorough scale development. Hence, the resulting scale has been subjected to a rigorous development and validation procedures. The items on the scale are carefully worded and the samples used were selected across demographics and organisations, thus increasing the utility and generalisability of the scale across demographics and organisational and performance appraisal contexts.

Implications for research and practice

The performance rater bias scale has implications for detecting and controlling supervisor errors in performance appraisals used for practice and research purposes. As part of the annual performance evaluation, many organisations evaluate supervisors on their ability to conduct performance evaluations and complete performance appraisal reports. Supervisors are assessed on the quality of evaluations regarding (1) their fairness and impartiality of ratings and (2) their ability to carry out their role in the performance evaluation system. This performance rater bias scale can complement the performance evaluation process. It can be used to detect supervisors who are susceptible to rating bias before or after the rating and improve the rating outcomes.

For supervisors' ratings of subordinates for research purposes and management practices, researchers and practitioners may correlate this performance rater bias scale with the supervisor ratings of the subordinates. Significant positive relationships between supervisor ratings of subordinates and this performance rater bias scale indicate that the supervisor rating is confounded and compromised by supervisor errors. The insignificant and negative relationship, however, indicates the absence of supervisor rater errors. Statistical control with partial correlation will help remove the rater effects and reveal the true supervisor ratings.

The results of measurement invariance confirmed the validity and generalisability of the scale in the Nigerian context. There is a need to replicate the study in other countries and cultures to validate the cross-country and cross-cultural validity of the scale.

Limitations

This study presents some limitations that need clarification. First, with respect to the further test of convergent validity using the average variance extract (AVE), five dimensions of the scale recorded AVE values slightly lower than 0.50. This, however, does not affect the convergent and discriminant validities of the scale. Literature (e.g., Ping, 2009) provides support that such slightly low AVE values are still acceptable, particularly in an "interesting" and "first-time" study and if they do not pose major discriminant validity problems as in this study. However, it is suggested that this should be considered when interpreting the result. It is further suggested that the results should be treated as provisional pending more replicative studies.

Second, results of the model comparison showed model invariance for age and position, indicating that younger and older and lower- and highest-level subgroups respectively are not equivalent and interpret the construct differently. However, other measurement invariance results indicate otherwise and overwhelmingly support equivalence of the two groups. It is therefore difficult to assume that the construct is noninvariant among these samples. As this is a first-time study in this area, as stated earlier, these findings should be treated as provisional pending further validation tests.

Third, an equivalent study with which the results of this study can be compared cannot be found. Most of the available studies only focus on the methods of appraisal and psychometric properties of the rating scales used, and none focus directly on the raters. One of the main strengths of the present study, therefore, is the use of a different method with a focus on the real sources of performance rating errors using diverse and experienced raters of performance appraisal in academia and industry.

Fourth, a potential problem must be borne in mind when using the scale. The scale may be vulnerable to test-sensitization effects. Performance appraisers who are administered the scale repeatedly over time may become sensitised to the instrument and form expectancies or hypotheses about its purpose. They may therefore learn to respond as desired in subsequent administrations, which compromises the validity and integrity of the instrument (Nunnally, 1978). It is recommended that the performance rater bias scale be administered to appraisers only once and that the test score be correlated with their ratings of subordinates over a period of five years. The five-year gap is sufficient to reduce test sensitisation effects in appraisers. Future research may explore the probability of the proneness of the scale to a sensitisation effect.

References

Aggarwal, A., Thakur, G. S. M. (2013). Techniques of performance appraisal: A review. International *Journal of Engineering and Advanced Technology (IJEAT),* 2249–8958.

Alves, H., Koch, A., Unkelbach, C. (2016). My friends are all alike: The relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology, 62*, 103–117.

Appelbaum, S. H., Roy, M., Gilliland, T. (2011). Globalization of performance appraisals: Theory and applications. *Management Decision, 49*(4), 570–585.

Atwater, L. E., Brett, J. F., Charles, A. C. (2007). Multisource feedback: Lessons learned and implications for practice. *Human Resource Management, 46*(2), 285–307.

Avery, D. R. (2003). Reactions to diversity in recruitment advertising: Are differences black and white? *Journal of Applied Psychology, 88*, 672–679.

Baltes, B. B., Bauer, C. C., Frensch, P. A. (2007). Does a structured free recall intervention reduce bias in performance ratings and by what mechanism? *Journal of Applied Psychology, 92*, 151–164.

Bialosiewicz, S., Murphy, K., Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants.* Retrieved from http://comm.eval.org/

Bijani, H. (2018). Investigating the validity of oral assessment training program: A mixed-methods study of ratrer' perceptions and attitudes before and after training. *Cogent Education, 5*(1), 1–20.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. F*ront Public Health, 6,* 149. doi: 10.3389/fpubh.2018.00149

Bol, J. (2011). The determinants and performance effects of managers' performance evaluation biases. *Accounting Review, 86*(5), 1549–1575.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS. Structural Equation Modeling (Vol. 22).* http://doi.org/10.4324/9781410600219

Campbell, J. P. (2012). Behaviour, performance and effectiveness in the twenty-first century. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology, Vol. 1* (159–194). New York, NY: Oxford University Press.

Cascio, W. F., Aguinis, H. (2004). *Applied psychology in human resource management (6th ed.).* Upper Saddle River, NJ: Prentice Hall. DOI: 10.1108/ijm.2002.23.6.580.3

Cheung, G. W., Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. doi: 10.1207/S15328007SEM0902_5

Dagar, A. (2014). Review of performance appraisal techniques. *International Research Journal of Commerce Arts and Science, 5*(10), 16–23.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. Language Testing, 33(1), 117–135.

Debnath, S. C., Lee, B. B., Tandon, S. (2015). Fifty years and going strong: What makes behaviorally anchored rating scales so perennial as an appraisal method? *International Journal of Business and Social Science, 6*(2), 16–25.

DeVellis, R. F (2012). *Scale Development: Theory and Application*. Los Angeles, CA: Sage Publications.

Downing, S. M., Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327–333.

Drucker, P. (2013). *People and Performance*. London: Routledge.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Elder, C, Knoch, U., Barhuizen, G., Von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Ellis, G. (2018). *Cognitive Biases in Visualizations (Ed.)*. New York, NY, USA: Springer.

Espinilla, M., Andres, R., Martinez, J., Martinez, L. (2013) A 360-degree performance appraisal model dealing with heterogeneous information and dependent criteria. *Information Sciences, 222*(4), 459–471.

Frederiksen, A, Lange, F., Kriechel, B. (2018). Performance evaluations and careers: Similarities and differences across firms. *Journal of Economic Behavior and Organization, 134*, 408–429.

Frederiksen, A., Khan, L. B., Lange, F. (2020). Supervisors and performance management systems. *Journal of Political Economy, 128*(6), 2123–2187. DOI: 10.1086/705715

Gadermann, A. M., Guhn, M., Zumbo, D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3), 1–13.

Gomez-Meijia, L. R., Balkin, D. B., Cardy, R. L. (2004). *Management human resources* (4th ed.). Prentice Hall.

Graen, G., Novak, M. A., Sommerkamp, P. (1982). The effects of leader-member exchange and job design on productivity and satisfaction: Testing a dual attachment model. *Organizational Behaviour and Human Performance, 30*, 109–131.

Grund, C., Przemeck, J. (2012) Subjective performance appraisal and inequality aversion. *Applied Economics, 44*(2), 2149–2155.

Hair, J., Black, W. C., Babin, B. J., Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.

Hair, J., Hult, G. T. M., Ringle, C., Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM).* Los Angeles: SAGE Publications, Incorporated.

Hair, J. F., Ringle, C. M., Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice, 19*, 139–152.

Henseler, J., Ringle, C. M., Sarstedt, M. J. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43,* 115–135. doi: 10.1007/s11747-014-0403-8

Heslin, P. A., Latham, G. P., VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*, 842–856.

Hinton, P. R., Brownlow, C., McMurray, I., Cozens, B. (2004). SPSS explained. England: Routledge Inc.

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

Howitt, D., Cramer, D. (1997). *An introduction to statistics in psychology*. London Prentice Hall; Harvester Wheatsheaf.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5,* 64–86.

Hu, L., Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. DOI: 10.1080/10705519909540118

Javidmehr, M., Ebrahimpour, M. (2015). Performance appraisal bias and errors: The influences and consequences. *International Journal of Organizational Leadership, 4*(3), 286–302.

Kang, O., Rubin, D., Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481–504.

Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., Shore, J. R. (2018). Development of the behaviorally anchored rating scales for the skills demonstration and progression guide (Research Report No. RR-18-24). Princeton, NJ: Educational Testing Service. DOI: 10.1002/ets2.12210

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). New York: The Guilford Press.

Kingstrom, P. O., Bass, A.R. (1981). A critical analysis of studies comparing behaviourally anchored rating scales (Bars) and other rating formats. *Personnel Psychology, 34*(2), 263–289. doi*:* 10.1111/j.1744-6570.1981.tb00942.x

Kumar, D. (2005). Performance appraisal: The importance of rater training. Journal of the Kuala Lumpur Royal Malaysia Police College, 4.

Landy, F. J., Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72–107.

Lazear, E., Shaw, K., Stanton, C (2015). The value of bosses. *Journal of Labour Economics, 33*(4), 823–861.

Levy, P. E., Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*(6), 881–905.

Loewenthal, K. (2004). An Introduction to Psychological Tests and Scales. 2nd Ed. Hove, UK: Psychology Press.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Matosas-López, L., Leguey-Galán, S., Doncel-Pedrera, L. M. (2019). Converting Likert scales into behavioral anchored rating scales (BARS) for the evaluation of teaching effectiveness for formative purposes. *Journal of University Teaching & Learning Practice, 16*(3), 1–24.

Mayer, R. C., Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84,* 123–136.

Martin, D. C., Bartol, K. M., Kehoe, P. E. (2010). The legal ramifications of performance appraisal: The growing significance. Public Personnel Management, 29(3), 379.

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., Ferreira, M. E. C. (2018). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica / Psychology: Research and Review, 34*(30), 3. doi: 10.1186/s41155-016-0057-1

Motowidlo, S. J. (1982). Relationship between self-rated performance and pay satisfaction among sales representatives. *Journal of Applied Psychology, 67*, 209–213.

Muchinsky, P. M. (2006). Psychology Applied to Work: An Introduction to Industrial and Organizational Psychology (8th Ed). California: Wadsworth.

Myford, C. M., Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Nunnally, J. C. (1978). *Psychometric Theory.* McGraw-Hill, New York.

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

O'Cass, A., Ngo, L. (2007). Market orientation versus innovative culture: Two routes to superior brand performance. *European Journal of Marketing, 41*, 868–887.

Pearce, J. L. (1987). Merit pay doesn't work: Implication from organization theory. In D. B. Balkin and L. R. Gomez-Meija (Eds.), *New Perspectives on Compensation* (169–178). Englewood Cliffs, NJ: Prentice Hall.

Pearce, J. L., Xu, Q. J. (2012). Rating performance or contesting status: Evidence against the homophily explanation for supervisor demographic skew in performance ratings. *Organization Science, 23*(2), 373–385. doi: 10.1287/orsc.1100.0585

Pedhazur, E. J. (1982). *Multiple regression in behavioural research* (2nd ed.). New York: Holt, Rinehart and Winston.

Perera, A. (2021). *Why the halo effect affects how we perceive others*. Simply Psychology. Retrieved from https://www.simplypsychology.org/halo-effect.html, 13 June 2021.

Ping, R. A. (2009). *Is there any way to improve Average Variance Extracted (AVE) in a Latent Variable (LV) X?* Retrieved from http://home.att.net/~rpingjr/ImprovAVE1.doc, (2009).

Putnick, D. L., Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Development Review, 41*, 71–90. doi: 10.1016/j.dr.2016.06.004

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*, 370–395.

Sahoo, C., Mishra, S. (2012). Performance management benefits organizations and their employees. *Human Resource Management International Digest, 20*(6), 3 – 5.

Schermelleh-Engel, K., Moosbrugger, H., Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research, 8*(2), 23–74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., King, J. (2006). Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Education Research, 99*(6), 323–338.

Smith, A. F. R., Fortunato, V. J. (2008). Factors influencing employee intentions to provide honest upward feedback ratings. *Journal of Business and Psychology, 22*(3), 191–207.

Smither, J. W., London, M., Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*(1), 33–66.

Stein, G. (2010). Managing People and Organizations. Amsterdam: Emerald Group Publishing.

Straub, D., Boudreau, M-C., Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*, 380–427.

Tabachnick, B. G., Fidell, L. S. (2007). *Using multivariate statistics*. (5th ed.). Boston, MA: Pearson Education.

Tabachnick, B. G., Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson Education.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Houndmills, England: Palgrave Macmillan.

Wildman, J. L., Bedwell, W. L., Salas, E., Smith-Jentsch, K. A. (2010). Performance measurement: Individual, team, and organizational strategies. In S. Zedeck (ed.), *APA Handbook of Industrial and Organizational Psychology, Vol. 1: Building and developing the organization* (303–341). Washington, DC: American Psychological Association.

Wind, S. A., Engelhard, G. (2017). Exploring rater errors and systematic biases using adjacent-categories Mokken models. *Psychological Test and Assessment Modeling, 59*(4), 493–515.

Woehr, D. J., Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189–205.

Zahed, A., Ardabili, F.S. (2017). Effect of similar-to-me effect on job satisfaction and organizational trust. *Problems and Perspectives in Management, 15*(4), 254–262. doi:10.21511/ppm.15(4-1).2017.09

Received 07.07.2022

Оценивание в ходе аттестации персонала: разработка и проверка шкалы предвзятости оценщика

ОДЕТУНДЕ Оладимейи Джамиу
Университет Лагоса, Лагос, Нигерия

Аннотация. Учитывая важность системы аттестации персонала и несмотря на непрекращающиеся дискуссии и прилагаемые усилия по обеспечению качества оценки руководителей в системе аттестации, эффекты оценивания по-прежнему широко распространены и продолжают подрывать достоверность и надёжность системы аттестации. *Цель*. Это исследование было направлено на разработку и проверку «Шкалы предвзятости оценщика» для обнаружения и контроля эффектов оценивания в системе служебной аттестации. *Метод*. Был проведён обширный обзор литературы. Отдельные данные были собраны для эксплораторного и конфирматорного факторного анализа. Выборки состояли из различных представителей профессорско-преподавательского состава вузов и профессиональных менеджеров, которые в своей карьере участвовали в оценивании эффективности. *Результаты*. Эксплораторный и конфирматорный факторный анализ собранных данных выявил восемь основных параметров систематической ошибки оценщика, вошедших в шкалу из 24 пунктов, по три пункта на каждый параметр. Шкала демонстрирует высокую надёжность содержания и конвергентную и дискриминантную валидность. *Ценность результатов*. Использование конфигурационной и метрической инвариантности, а также критерия Хи-квадрат и однофакторного дисперсионного анализа позволило установить, что шкала отображает концептуальную эквивалентность в различных группах. *Значение для исследований и практики*. Исследование имеет важное значение для решения проблемы обнаружения и контроля ошибок руководителя в ходе оценивания эффективности подчинённых, как в научном, так и в практическом плане. Исследование вносит свой вклад в науку и практику управления человеческими ресурсами в организации. Подчёркиваются ограничения исследования и обсуждаются потенциальные проблемы в его проведении.

Ключевые слова: аттестация персонала; шкалы предвзятости оценщика; эксплораторный факторный анализ; конфирматорный факторный анализ.