Revista Interamericana de Psicología/Interamerican Journal of Psychology
2021, Vol., 55, No. 1, e1229

Influence of measurement type and the moment of occurrence of low performance behaviour's on task and citizenship performance appraisal

Christian Rosales [1], María Dolores Díaz-Cabrera, & Estefenía Hernández-Fernaud [2]

Universidad de La Laguna, La Laguna, España.

ABSTRACT

This research studies whether the moment of occurrence of a task or contextual behaviour with a low performance produces a primacy or recency effect and whether it causes changes in performance appraisal. We also analyzed whether the nature of assessment questionnaire items affects raters' assessments and how the sequence of questionnaire presentation and completion may do so. Participants were 146 undergraduate students. We used a design with two inter-subject variables (questionnaire presentation and performance sequence) and one within-subject variable (global versus specific questionnaires). Findings show that if a low performance is presented at the beginning of the assessment period, the performance assessment will be more negative. Also, results indicate that task performance appraisals and contextual behaviour assessments are higher and less accurate when performed with a questionnaire that includes global items.

Keywords
performance appraisal; primacy effect; recency effect; global and specific items

RESUMEN

Esta investigación estudia si el momento de ocurrencia de una acción de desempeño de tarea o contextual bajo produce un efecto de primacía o recencia en la evaluación del desempeño. Asimismo, se analiza en qué medida la naturaleza de los ítems del cuestionario de evaluación y la secuencia de presentación y cumplimentación del cuestionario pueden influir en la valoración de los evaluadores sobre el desempeño laboral. La muestra está formada por 146 estudiantes de grado. Se emplea un diseño con dos variables inter-sujetos (presentación del cuestionario y secuencia de desempeño) y una variable intra-sujeto (cuestionarios globales versus cuestionarios específicos). Los resultados muestran que, si se presenta un bajo rendimiento al comienzo del período de evaluación, la evaluación del desempeño será más negativa. Además, entre los hallazgos, destaca que las evaluaciones del desempeño de la tarea y contextual son más altas y menos precisas cuando se realizan con un cuestionario que incluye ítems globales.

Palabras clave
desempeño laboral, efecto de primacía; efecto de recencia, cuestionarios; ítems globales y específicos

[1] Correspondence about this article should be addressed to **Christian Rosales Sánchez** crosales@ull.es

[2] ***Conflicts of Interest*:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Influencia del Tipo de Medición y del Momento de Aparición de las Conductas de Bajo Rendimiento en la Evaluación del Desempeño de Tarea y Cívico

Originally, Industrial/Organizational Psychology was considered a branch that would facilitate and relate socioeconomic development and work performance. Today, in the competitive and changing labor context, it is essential to know and manage the necessary skills of workers to carry out a good job performance (Vélez, Rosario, Méndez & Vargas, 2016). Likewise, there are multiple relevant variables when analyzing the evolution of the various indicators desirable for the organization, among which the performance evaluation stands out (Carmona-Halty & Villegas-Robertson, 2018). Thus, given the importance of this formal appraisal system, in this study was conducted to verify whether raters' assessment of an employee's task and contextual performance are influenced by a) the moment (at the beginning, in the middle or at the end of the review period), in which a task and/or contextual behaviour with a low or inadequate performance is presented, b) the nature, whether global or specific, of the items of the assessment method and, c) the order of presentation and completion of the assessment questionnaires influence raters' assessment of an employee's task and contextual performance.

Campbell (1999) points out that work performance is determined by the behaviours and actions of employees who are important for the objectives of the organization and can be measured and appraised in accordance with their contribution to these objectives. These behaviours are divided into three types: task, contextual and counterproductive. This study focuses particularly on the appraisal of task and contextual performance.

Task performance is the quality and frequency with which employees carry out the activities formally assigned to their work posts (Borman & Motowidlo, 1997; Motowidlo & Schmit, 1999). Moreover, contextual performance involves activities that are not part of the formal role but contribute to the efficiency of the organization, helping create and maintain an appropriate working atmosphere on a social and psychological level (Díaz-Vilela, Díaz-Cabrera, Isla-Díaz, Hernández-Fernaud, & Rosales-Sánchez, 2012). Thus, although this type of performance does not vary between the different jobs, it and the characteristics of each organization affect its operation, such as the organizational climate (Silvestre, 2017).

The existence of these two sets of activities makes performance appraisal a complex process by which an organization can determine the extent to which employees

are performing their work effectively (Griffin & Ebert, 2002). However, this appraisal can reveal the contribution of a specific employee, not only in the tasks assigned to the post and position, but also in relation to the contextual tasks that the employee undertakes as part of daily interaction with other organizational agents.

Following are some of the benefits contributed by performance management and assessment (Aguinis, Joo & Gottfredson, 2011; Schraeder, Becton, & Portis, 2007): a) performance improvement through feedback, b) suitable assignment of workers to work posts in accordance with their skills, capabilities and experience, c) training needs identification, d) acquisition of relevant information for recognition of work, management of salary incentives and promotions. However, if the organization is to benefit from these advantages, it is vital that the system of appraisal is not distorted, in which, even today, a distance is perceived between the theory and practice of the appraisal system due to the influence of the human element, distorting the objectivity and accuracy of the system and generating employee dissatisfaction with the appraisal scheme (Dauda, 2018). Along these lines, the aim of this paper is to determine the role in performance appraisal of variables such as the effect of primacy and recency, benevolent bias or the appraisal method used (Aguinis, 2013; Dewberry, Davies-Muir & Newell, 2013; Kane, Bernardin, Villanova & Peyrefitte, 1995; Steiner & Rain, 1989; Wagner & Goffin, 1997). In order to achieve this, it is important to analyze how performance assessment varies in terms of objectivity and accuracy. Traditionally, they have been studied against the four accuracy deviation measures outlined by Cronbach (1995) and by comparing them with the rating issued by a group of experts (McIntyre, Smith & Hassett, 1984; Smith, 1986; Woehr & Huffcutt, 1994).

First of all, we are interested in examining how performance is affected by the order in which employee information is processed. In the field of organizational psychology, research into performance assessment reveals the existence of certain biases and errors that can have a significant impact on raters' assessments (Kane et al., 1995). These biases include the effect of primacy and recency. In performance assessment, primacy occurs when raters base their judgement on the initial information obtained about the employee, while recency is a result of raters focusing on the latest known data about employee performance (Aguinis, 2013). The literature contains few studies on how the effect of primacy and recency occurs in performance assessment, and whether one is stronger than the other. Highhouse and Gallo (1997) found that by presenting a low performance sample at the end, performance assessment was lower. In other words, the

effect of recency modifies participants' assessment in a specific skill and in performance in general. By contrast, Steiner and Rain (1989) found that when raters observed several fragments of an employee's performance in a single session, performance assessment was influenced in experimental conditions in which an adequate performance level was placed last; this was not the case for conditions of low performance. Conversely, when raters are shown fragments of an employee's performance on several consecutive days, the resulting performance assessments are affected when conditions of low performance are placed at the end. Neither Highhouse and Gallo (1997) nor Steiner and Rain (1999) obtained an effect of primacy in any of the performance situations, irrespective of adequacy or inadequacy. Moreover, Steiner and Rain (1989) found no effect of recency in the performance assessments when they were carried out in a single session. In this paper, we aim to explore how the moment in which a task and/or contextual behaviour with a low performance is presented (at the beginning, in the middle or at the end of the review period) affects raters' performance assessment.

Hypothesis 1: Performance assessment will be more negative when the task and/or contextual behaviour with a low performance appears at the end (effect of recency) rather than at the beginning (effect of primacy) or in the middle (control) of the review period.

Secondly, another common error in performance assessment is the tendency of some raters to be systematically indulgent (Dewberry et al., 2013). This error is known as benevolent bias and occurs when raters tend to award high scores to most or all employees (Aguinis, 2013). This bias directly affects the performance assessment process, since it reduces the possibility of identifying and rewarding employee performance, and performance validity is therefore diminished (Bretz, Milkovich, & Read, 1992). The study of benevolent bias has been closely linked with training programmes created for raters and with the development and improvement of performance assessment techniques (Díaz-Cabrera et al., 2014; Rosales, Díaz-Cabrera & Hernández-Fernaud, 2019; Woehr & Huffcutt, 1994). Traditionally, a distinction has been made between two assessment methods: 1) absolutes, which involve the assessment of assessees against a universal standard or specific behaviour, such as behaviourally anchored rating scales (BARS, Smith & Kendall, 1963) and behavioural observation scales (BOS, Latham & Wexley, 1977) and 2) comparatives, which require the rater to assess the assessee against other assessees, including, for example, pairing comparisons. Absolute and comparative performance assessment methods use specific and global

items, respectively (Wagner & Goffin, 1997). Our research has therefore compared the accuracy of raters' assessments according to whether they used global or specific items. The results are not conclusive, since Fay and Latham (1982) found that raters using scales of specific items committed fewer errors than those using global items. Heneman (1988), however, found the opposite; that is, raters who used global items issued more accurate assessments than those who used specific items. Therefore, this paper aims to analyze how the nature of the assessment method used, whether with global or specific items, affects performance assessment.

Hypothesis 2a: By evaluating performance with an assessment method using global items, the assessment will be higher than the one issued with a questionnaire containing specific items.

Hypothesis 2b: By evaluating performance with an assessment method using global items, the assessment will be less accurate than the one issued with a questionnaire containing specific items.

We also explore whether the order in which the different types of items are completed affects assessment. Given that global items are expected to generate a higher and less accurate assessment, it is to be expected that when the assessment is done first, it will contaminate a subsequent assessment that uses a system of specific items.

Hypothesis 3a: Completion of a performance assessment questionnaire containing global items placed first will result in the subsequent assessment being higher, when a tool containing specific elements is used.

Hypothesis 3b: Completion of a performance assessment questionnaire containing global items placed first will result in the subsequent assessment being less accurate, when a tool containing specific elements is used.

In short, these hypotheses propose the way in which the moment information about employee performance and type of instrument used in performance assessment influences the accuracy of the assessment.

Method

Participants

The sample was made up of 146 university students, of whom 52.1% and 47.9% were second-year labour relations students and psychology students, respectively. Of the total, 28.1% were men and 71.9% women. The mean age of the patients was 22.7 years.

Moreover, 41.8% of the participants had prior work experience. The involvement of the participants in this research was voluntary although in exchange for their collaboration they were offered an incentive linked to the final grade of the subject.

Design

Three independent variables were used: two between subject and one within subject. The independent between subject variables were as follows: 1) the moment when a low performance task and/or contextual behaviour occurs on three levels, at the beginning (days 1-2), in the middle (days 2-3 and 3-4) or at the end of the assessment period (days 4-5), and 2) the order of presentation and completion of the performance assessment scales with two levels, Global-Specific and Specific-Global. The nature of the items of the performance assessment method on two levels, global or specific, was used as an independent within-subject variable. The global and specific task and contextual performance measurements were also used as dependent variables.

Materials and tools

This research used materials designed as stimuli and scales for performance assessment. Each one is outlined below:

- *Work performance description of a fictitious employee:* five samples of an employee's performance over five working days were created. These performance samples included a series of performance task activities habitually carried out by administrative and office staff, as well as the contextual behaviours of the employee in work interactions. These administrative tasks were identified by analyzing a previous post (Díaz-Vilela et al., 2015). In relation to contextual behaviours, according to an assessment by a group of experts and administrative professionals, the ten most representative and realistic descriptions were selected from the 20 created, in line with Borman and Motowidlo's (1993) contextual performance dimensions. On three of the five working days described, the employee carried out task and contextual performance adequately, while on the two remaining days, performance in both areas was low or inadequate. For example, on the subject of contextual performance, descriptions of situations of an adequate level of performance were given when the employee helped a colleague with computer program issues, as well as situations with a low level of contextual performance, such as when the employee refused a colleague's offer of a training course, on the basis

that such courses were useless. An example of adequate task performance was when the employee correctly filed the documentation of a user he/she had recently attended. Examples of low task performance included situations in which the employee made some procedural mistakes.

- *Spanish adaptation of Coleman and Borman's (2000) scale of citizenship performance behaviours* (Díaz-Vilela et al., 2012): 27 specific items, with a response scale of 1 to 7, with two anchors: "Not characteristic at all" and "That is very characteristic". This scale has an overall reliability of $\alpha = 0.96$. Moreover, the response option "Not applicable" was included for items that participants considered could not be answered with the information provided.

- *Questionnaire to assess global contextual performance:* this ad hoc tool comprised six global items. Five items evaluated employee *contextual* performance according to the five dimensions proposed by Borman and Motowidlo (1993), while the sixth item assessed general *contextual* performance. The response scale ranged from 1 to 7, where 1 was "Not characteristic at all" and 7 was "That is really characteristic". As with the previous scale, this questionnaire included the optional response: "Not applicable". The score of global *contextual* performance was found by calculating the arithmetic mean of the six items. This questionnaire has an overall reliability of $\alpha = 0.74$. *"Manifest enthusiasm and dedication to his/her work and strive to perform his/her tasks well"* is an example of items based on Borman and Motowidlo's five dimensions (1993). The overall item used to assess contextual performance was: *"He/she is an achiever, shows interest in his/her job, collaborates with peers and bosses, and is friendly when dealing with people (customers and users)"*.

- *Questionnaire to assess task performance:* it contained 14 specific items, based on a previous task inventory of the office position (Díaz-Vilela et al., 2015), with a response scale of 1 to 7, where 1 is "Improvable" and 7 is "Exceptional". The response option "Not applicable" was also available. For instance, these are among the items used: *"Communicate with the right people at all times, both in writing and verbally, providing information clearly and accurately"* and *"Send the necessary documents requested by other departments on time and in due form"*. This scale has an overall reliability of $\alpha = 0.85$

- *Assessment of global task performance:* a question to elicit an overall appraisal of performance was also added to the questionnaire to assess task performance. The responses for this question followed the same scale as for specific items. The question

used to elicit an overall appraisal in the assessment of global performance was: *"In your view, indicate the overall quality with which the employee performs the tasks assigned to his/her job"*.

At the end of the instruments, the participants were required to provide demographic data such as age, gender, on-going degree and previous work experience

Procedure

The participant's task consisted in slowly reading the description of five working days and then assessing the contextual and task performance of the employee.

The materials and tools were presented in a booklet, available in eight differentiated versions: 1) the place, at the beginning (days 1-2), in the middle (days 2-3 and 3-4) or at the end (days 4-5), in which a task and/or contextual behaviour appeared as low performance, and 2) the order of completion of the performance assessment scales (specific versus global).

Data collection was undertaken in group sessions in a classroom, for around one hour. Care was taken to ensure that the different versions of the booklets were represented equally and that participants in adjoining seats had different versions.

A group of three experts in performance assessment also evaluated the Work performance description of a fictitious employee. The aim of this step was to reach a consensus on the task and contextual performance of the fictitious employees, using the Delphi method (Linstone & Turoff, 1975). For this, the three experts received by email the work performance description and different questionnaires. Experts had one week in which to send their ratings. Then, based on the answers given, the response scale of the questionnaires was reduced and only the most frequently chosen alternative was maintained. The new questionnaires were sent back to the experts who again had one week in which to assess the performance of the fictitious employee. This time they were also required to explain their answers. Based on this second assessment, questionnaires were adapted to show only the most voted options of the response scale, as well as a summary of the most important comments. Finally, we gathered the three experts and gave them the questionnaires reduced to the most frequently chosen alternative. They were asked to assess each alternative, and discuss and reach an agreement on the task and civic performance (global and specific) shown by the fictitious employee. These assessments were used as a criterion to evaluate the goodness of fit of participants'

assessments, in the understanding that assessments are more accurate when they are more similar to expert appraisal (expert task performance assesment, X=4.93; expert contextual performance assesment, X=3.5).

Results

All analyses were carried out with the SPSS software program 21.0. Firstly, we checked for the presence or not of univariate and multivariate outliers. No univariate and multivariate outliers were found. Secondly, using independent samples *t*-test, we analyzed whether the fact that the participants (university students) had previous work experience or not would cause discrepancies in the assessment made, obtaining no statistically significant differences for this variable.

Thirdly, using Pearson's correlation, we checked whether there is a connection between performance assessment undertaken with global or specific elements, for both contextual (r=0.643; p< .001) and task (r=0.618; p< .001) performance.

Fourthly, we undertook a repeated measures ANOVA (3×2×2) with two intergroup variables: 1) the order of presentation of low performance, either at the beginning, in the middle or at the end of the employee's performance sample, and 2) the order of presentation of the items, global versus specific, and an independent within-group variable, the type of assessment measurement, global versus specific. Table 1 shows the means and standard deviations for each dependent variable.

Table 1.
Descriptive statistics of dependent variables (experimental conditions).

		Task performance				Contextual performance			
		Global		Specific		Global		Specific	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Primacy	Global-Specific (n=23)	4.30	1.52	4.51	.74	4.65	.89	4.06	.77
(n=47)	Specific-Global (n=24)	5.26	.92	4.75	.77	4.83	.78	4.21	.66
Recency	Global-Specific (n=24)	5.71	1.08	5.24	.85	5.44	.72	4.82	.88
(n=48)	Specific-Global (n=24)	5.54	.88	5.03	.81	5.08	.99	4.41	.76
Control	Global-Specific (n=27)	5.81	1.1	5.32	.87	5.43	.88	4.42	.96
(n=48)	Specific-Global (n=21)	5.38	1.07	4.83	1.2	4.80	.91	3.95	1.1

A significant main effect of one intergroup variable, the order of presentation of low performance, was obtained ($F(2, 139) = 8.55$; $p < .001$; $\eta^2 = .11$) in task performance assessment. Figure 1 shows the relationship between the total averages of the types of performance (contextual and task) in the different moments of presentation of low performance (primacy, middle or control, and recency) versus the assessment issued by the group of experts. Specifically, posteriori tests with Bonferroni adjustment indicate that participants who are presented with low performance at the beginning (primacy

effect, X=4.7), make a lower task performance assessment than participants in the recency (X=5.4) and control group (X=5.4) ($p< .05$).

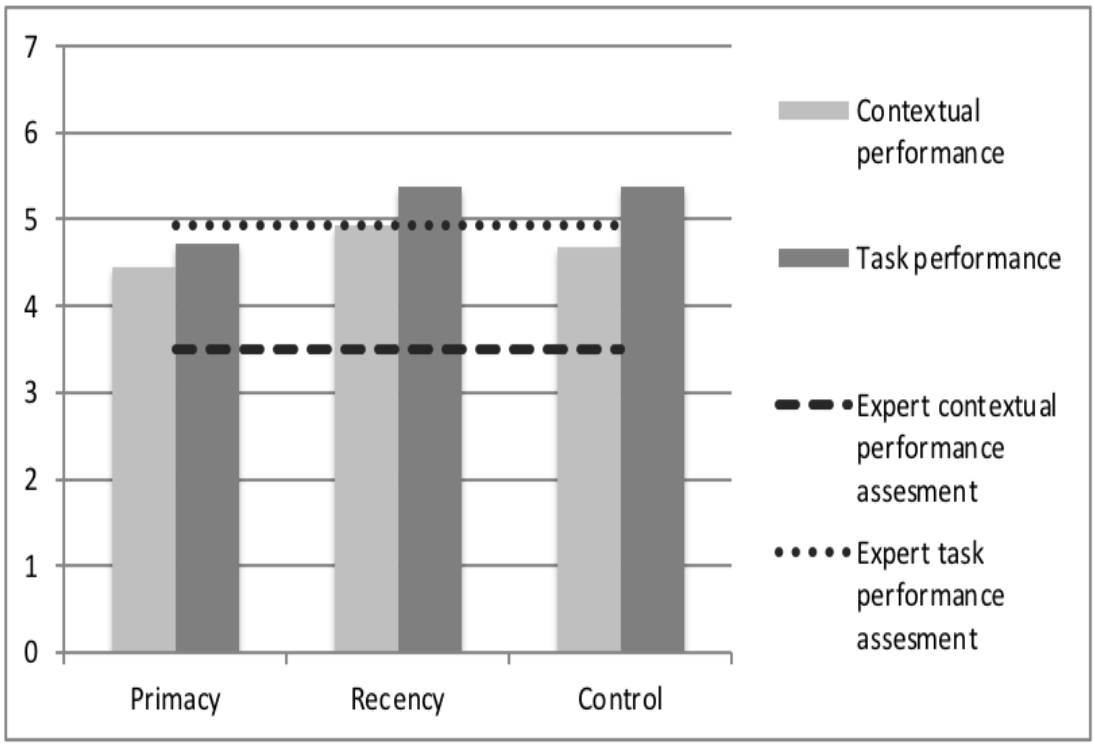

Figure 1. Relationship between the total averages of the types of performance (contextual and task) in the different moments (primacy, recency and control) versus the assessment of the experts.

Moreover, a significant main effect of the nature of the assessment measurement was obtained ($F(1, 139) = 28.16$; $p < .001$; $\eta 2= .17$) in task performance assessment, the highest assessments resulting from a questionnaire with global performance items. Figure 2 shows the relationship between the overall average performance types (contextual and task) with the nature of the items of the performance assessment method (global or specific) against the assessment issued by the group of experts. However, neither the main effect of the order of presentation of the items, nor the double interactions effects were significant. And no significant triple interaction effect was found.

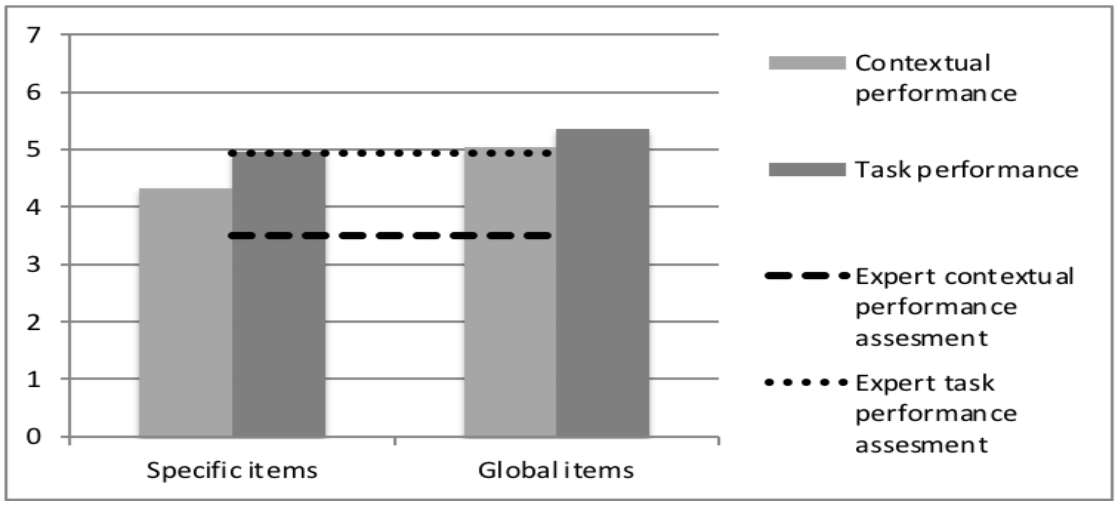

Figure 2. Relationship between the total means of performance types (contextual and task) with the nature of the items of the performance assessment method (global or specific) versus the assessment of the experts

In regard to contextual performance, the results are similar to those for task performance. The main effect of the order of presentation of low performance was significant ($F(2, 137) = 4.93$; $p < .001$; $\eta 2 = .07$) (Figure 1). Posteriori tests with Bonferroni adjustment show that participants who are presented with low performance at the beginning (primacy effect, X=4.4), make a lower contextual performance assessment than participants in the recency group (X=4.9; $p < .05$). Likewise, the independent within-group variable, the nature of the items, was significant ($F(1, 137) = 130.76$; $p < .001$; $\eta 2 = .49$) (Figure 2). No main effect of the order of presentation of the items, or double or triple interaction effects was found.

Fifthly, we calculated the distance between the direct score of each individual and the score given by the experts for each dependent variable. Table 2 shows the descriptive statistics of the distance variables.

Table 2.
Descriptive statistics of the distance between the direct scores of each participant and expert assessment for the dependent variables (experimental conditions)

		Distance-Task performance				Distance-Contextual performance			
		Global		Specific		Global		Specific	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Primacy	Global-Specific (n=24)	1.21	1.06	.70	.46	1.15	.83	.68	.63
(n=48)	Specific-Global (n=24)	.82	.50	.60	.48	1.33	.78	.79	.55
Recency	Global-Specific (n=24)	1.09	.75	.76	.48	1.94	.72	1.33	.86
(n=48)	Specific-Global (n=24)	.84	.65	.67	.43	1.67	.80	.98	.66
Control	Global-Specific (n=27)	1.17	.75	.82	.46	1.93	.88	.97	.91
(n=50)	Specific-Global (n=23)	1.00	.55	1.02	.47	1.32	.74	.86	.73

Finally, we repeated the repeated measures ANOVA (3×2×2) with the distance scores. The results for task performance show that measurements are not influenced by the moment when a low performance task and/or contextual behaviour is presented or by the order of presentation and completion of the questionnaires. However, statistically significant differences were found, depending on the nature of the items of the tool (global/specific) in task performance assessment ($F(1, 140)= 20.17$; $p< .001$; $\eta 2= .13$), so that evaluations undertaken with a specific tool were closer to those made by experts than those done with a global tool.

In relation to contextual performance, it has been found a significant main effect of one intergroup variable, the order of presentation of low performance ($F(2, 140) = 6.53$; $p< .001$; $\eta^2 = .09$). Specifically, posteriori tests with Bonferroni adjustment indicate that participants who are presented with low performance at the beginning (primacy effect, X=0.99), make a lower task performance assessment than participants in the recency (X=1.5; $p< .05$). In addition, it has been detected significant differences in assessment due to the nature of the items ($F(1, 140) = 101,53$; $p< .000$; $\eta^2= .42$), the global tools giving less accurate evaluations.

These results go against the first hypothesis, since the influence of the primacy effect on performance assessment has been found. In addition, the findings do not support hypotheses 3a and 3b, since the order of presentation of the global and specific tools produced no significant differences in performance assessment. However, results do support hypotheses 2a and 2b, the assessments undertaken using tools with global items being higher and less accurate than those conducted with specific items.

Discussion

DeNisi and Murphy (2017), in their review of research on performance appraisal and performance management, point out the existence of eight relevant research fields: (1) scale formats, (2) criteria for evaluating ratings, (3)) training, (4) reactions to appraisal, (5) purpose of rating, (6) rating sources, (7) demographic differences in ratings, and (8) cognitive processes. Based on this classification, this paper, through the objectives pursued, addresses three of these research fields (scale formats, criteria for evaluating ratings and cognitive processes), trying to provide new knowledge about the influence of certain variables on the accuracy of task and contextual performance appraisal issued by raters. In particular, we have explored the extent to which the moment a low or inadequate task and/or contextual behaviour performance, the nature of the items of the assessment

tools, global or specific, and the order of presentation and completion can affect the appraisal issued by raters.

The first hypothesis in this study (performance assessment will be more negative when the performance is low and contextual behavior at the end (effect of recency) and not the beginning (effect of primacy) of the review period) was considered because, although the results of some studies support this idea, they are not consistent (Gürbüz & Dikmenli, 2007; Highhouse & Gallo, 1997; Steiner & Rain, 1989). Performance appraisal should range all assessment period although, on many occasions, recent events or behaviors are more remarkable. So, some raters only use the latest behavior without paying attention to the performance shown by the employee over time (Gürbüz & Dikmenli, 2007). Our results indicate that performance assessment is dependent of the moment in which a low performance task and/or contextual behavior appears. In other words, when a low performance task and/or contextual behavior appears at the beginning (primacy effect) of the assessment period, task performance appraisals will be more negative than the others groups (recency and control). Nevertheless, there is no difference in terms of accuracy in the three groups (primacy, recency and control), since none is closer than another to the assessment of the group of experts. As for contextual behavioral assessments, these are also negative in front of the recency group when a low performance task and / or contextual behavior appears at the beginning (primacy effect) of the assessment period. Even though, in this case, the assessment of primacy group is more accurate than the recency group. No difference in accuracy was observed between the group primacy and control. This finding, on the one hand, contrasts with that obtained by Highhouse and Gallo (1997), who pointed out that the recency effect influences assessment when low performance is presented at the end. On the other hand, our results neither concur with those found by Steiner and Rain (1989), who obtained no recency effect of low performance situations when raters assessed employee performance in a single session. However, this phenomenon indicates that if the raters had information on the employee's performance throughout the assessment period, the appraisal would not be influenced by the action of the recency effect.

The second hypotheses, 2a and 2b (a performance assessment method using global items will be higher and less accurate than one using specific items) address the effect of benevolent bias on assessment, depending on the nature of the items that make up the scale. The results confirm these two hypotheses. In particular, an assessment undertaken using a tool of global items is higher and less accurate than that undertaken using a

questionnaire with specific items. This result is obtained in assessments of both types of performance: task and contextual. One possible explanation for this result is that when global assessment is undertaken, the rater makes a general estimation of employee performance as a whole, making it easier for mistakes to be made. However, when assessment is more specific, the rater is required to make more efforts of attention, concentration and recognition because different specific tasks and behaviours must be considered, among which there may be variations in employee performance level. This explanation is based on the conclusion found by Gaugler and Thornton (1989), who maintain that the complexity of the task increases the likelihood of cognitive bias, thereby reducing judgement accuracy. In particular, these authors point out that making judgements about a whole job is more complex than for specific tasks, since it calls for a greater use of memory and information integration. Depending on the type of measure used, these differences are in line with Fay and Latham's (1982) results, which indicate that raters tend to issue higher and more benevolent assessments when they appraise global questionnaires. The tendency to overestimate employee performance affects organizational decision-making (e.g., promotions, salaries, training needs identification), since it reduces the validity of assessments and decision-making (Bretz et al., 1992).

In relation to hypotheses 3a and 3b (an assessment using a tool with specific elements will be higher and less accurate when a performance assessment questionnaire containing global items has been completed beforehand), Dror and Fraser-Mackenzie (2008) alert to the considerable influence of first impressions on end assessments. Individuals tend to hold these prior beliefs, even though the new information differs or contradicts the former. As the results show, the appraisal carried out with an assessment system of global elements is more benevolent. If this global appraisal is undertaken first of all, it is to be expected that the global assessment can create an impression that could influence a second assessment undertaken with specific items. Contrary to the prediction of the third hypothesis, the results indicate that the assessment is not influenced by the order in which both types of tools are presented and completed. By forcing raters to concentrate on more detailed tasks and contextual behaviours, the benevolent bias that occurs in global measures tends to disappear. Thus, concentrating on specific items in the questionnaire and remembering relevant information for the response means that participants do not pay attention to the global assessment issued beforehand.

The results of this study must be interpreted bearing in mind that the sample is made up of students. Although performance assessment is a process traditionally

associated with employees in a given organization, numerous studies are based on samples of students who have never worked in the position that they are required to assess (Morgeson & Campion, 1997). In fact, the task required of participants consisted in issuing an evaluative judgement, based on their opinions; in other words, they were to assess the performance of a fictitious employee. Generally speaking, it can be said that, in our everyday and professional lives, all individuals make assessments in some way or other. Similarly, whether students had prior work experience or not was found to make no difference to the assessment. Furthermore, although students may not have sufficient knowledge of the work position or associated tasks and behaviours, the type of position and the administrative and office staff can be familiar to anyone who has enrolled at university, for example. However, that the performance appraisal made by the participants had no repercussion either on them or on the appraisee—possibly affecting their degree of involvement in the evaluation—would need to be taken into account.

Also, the novel contribution to the area of work performance assessment and the practical implications of this study are notable. The results of this paper bring new information to the few studies on how task and contextual performance assessment may be distorted by the action of different variables. In our opinion, this study brings new knowledge about questionnaires used in performance assessment. It finds that tools with specific items lead to a more objective and accurate assessment than questionnaires that use global elements, therefore greatly minimizing the influence of benevolent bias in both task and contextual performance assessment. Organizations should weigh the advantages and inconveniences of the assessment method they use to identify their employees' level of performance. Developing a tool with specific items is more costly in terms of time, staff and economic factors than creating an assessment system using global elements. However, the use of specific items increases the objectivity and accuracy of the assessment. When an organization opts to use tools with global items, data correction could be established, since benevolent bias is likely to occur (Díaz-Vilela et al., 2012). Likewise, this tendency can be reduced by instructing raters in one or more training programmes: Rater Error Training (RET), Performance Dimension Training (PDimT), Frame-of-Reference (FOR) and Behavioural Observation Training (BOT) (Rosales et al., 2019; Woehr & Huffcutt, 1994). This paper provides new information about how biases can influence the performance assessment process, depending on when low performance occurs. Thus, these results underscore the need to record, as far as possible, the performance of a worker throughout the assessment period. For this, Aguinis (2013)

proposes the use of notes or diaries as behavior registration strategies, with the purpose of reducing the influence of these biases in the performance assessment.

The findings of this research provide an interesting starting point for future studies, as long as researchers and professionals use these findings sensibly and responsibly, avoiding their misuse or abuse (Porras, 2016). Thus, for instance, it would be interesting to analyze whether previous experience as an appraiser affects the accuracy of the appraisal, regardless of the assessment method used or the moment in time when the information was obtained by the assessor (e.g., based on first impressions compared with recent information about the assessee). It would also be useful to explore whether the same pattern of results is obtained when performance appraisal has potential consequences for the appraisee. In short, notwithstanding the criticisms of performance appraisal, this is a system that works well and provides benefits to both workers and the organization (Dauda, 2018). Therefore, a further study of all the variables that could influence the accuracy and objectivity of the performance appraisal is necessary.

References

Aguinis, H. (2013). *Performance management*. Pearson.

Aguinis, H., Joo, H., & Gottfredson, R. K. (2011). Why we hate performance management - and why we should love it. *Business Horizons*, 54, 503-507. http://dx.doi.org/10.1016/j.bushor.2011.06.001

Borman, W. C., & Motowidlo, S. (1993). Expanding the criterion domain to include elements of contextual performance. En N. Schmitt, & W. C. Borman, *Personnel selection in organization*s (pp 71–98). Jossey - Bass.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, 10(2), 99-109. http://dx.doi.org/10.1207/s15327043hup1002_3

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, 18, 321-352. http://dx.doi.org/10.1177/014920639201800206

Campbell, J. P. (1999). The definition and measurement of performance in the new age. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 399-430). Jossey-Bass.

Carmona-Halty, M., & Villegas-Robertson, J. M. (2019). El Capital Psicológico Predice el Bienestar y Desempeño en Estudiantes Secundarios Chilenos. *Revista Interamericana de Psicología*, 52(3).

Dauda, Y. (2018). A review of performance appraisal systems in different countries: The UK, India, South Africa and Ghana. *International Journal of Applied Environmental Sciences*, 13(2), 203-221.

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, 102(3), 421. https://doi.org/10.1037/apl0000085

Dewberry, C., Davies-Muir, A., & Newell, S. (2013). Impact and causes of Rater Severity/Leniency in Appraisals without Postevaluation communication between raters and rates. *International Journal of Selection and Assessment*, *21*(3), 286-293. http://dx.doi.org/10.1111/ijsa.12038

Díaz-Cabrera, D., Hernández-Fernaud, E., Isla-Díaz, D., Delgado, N., Díaz-Vilela, L. & Rosales-Sánchez, C. (2014). Factores relevantes para aumentar la precisión, la viabilidad y el éxito de los sistemas de evaluación del desempeño laboral. *Papeles del Psicólogo*, *35*(2), 115-121.

Díaz-Vilela, L., Delgado, N., Isla-Díaz, R., Díaz-Cabrera, D. Hernández-Fernaud, E & Rosales-Sánchez, C. (2015). Relationships between contextual and task performance and interrater agreement: Are there any? Manuscript presented for publication. PLoS ONE 10(10): e0139898. https://doi.org/10.1371/journal.pone.0139898

Díaz-Vilela, L., Díaz- Cabrera, D., Isla-Díaz, R., Hernández-Fernaud, E., & Rosales-Sánchez, C. (2012). Spanish adaptation of the citizenship performance questionnaire by Coleman and Borman (2000) and an analysis of the empiric structure of the construct. *Revista de Psicología del Trabajo y las Organizaciones*, *28*(3), 135-149. http://dx.doi.org/10.5093/tr2012a12

Dror, I. E., & Fraser-Mackenzie, P. (2008). Cognitive biases in human perception, judgment, and decision making: Bridging theory and the real world. In K.

Rossmo (Ed.) *Criminal Investigative Failures* (pp. 53-67). Taylor & Francis Publishing.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology*, 35, 105–116. http://dx.doi.org/10.1111/j.1744-6570.1982.tb02188.x

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, 74, 611-618.doi:http://dx.doi.org/10.1037//0021-9010.74.4.611

Griffin, R., & Ebert, R. (2002). *Business*. Prentice Hall.

Gürbüz, S. & Dikmenli, O. (2007). Performance Appraisal Biases in A Public Organization: An Empirical Study. *Journal of the Kocaeli University of the Institute of Social Sciences*, *13*(1), 108-138.

Heneman, R. L. (1988). Traits, behaviors, and rater training: Some unexpected results. *Human Performance*, 1, 85–98. http://dx.doi.org/10.1207/s15327043hup0102_1

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance, 10*, 31-46.

Kane, J.S., Bernardin, J.J., Villanova, P. & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, 38, 1036-1051. http://dx.doi.org/10.2307/256619

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology*, 30, 225-268. http://dx.doi.org/10.1111/j.1744-6570.1977.tb02092.x

Linstone, H. A. & Turoff, M. (1975). *Delphi Method: Techniques and Applications*. Addison-Wesley Publishing.

Mclntyre, R., Smith, D., & Hassett, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, 69,147-156. http://dx.doi.org/10.1037/0021-9010.69.1.147

Morgeson, F.P. & Campion, M.A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, 82(5), 627-655. http://dx.doi.org/10.1037//0021-9010.82.5.627

Motowidlo, S. J., & Schmit, M. J. (1999). Performance assessment in unique jobs. In D. R. Ilgen, & E. D. Pulakos, *The changing nature of performance: Implications for staffing, motivation, and development*, (pp. 56-87). Jossey-Bass.

Porras, N. R. (2016). Aproximación histórica a la psicología del trabajo y de las organizaciones en Colombia. *Interamerican Journal of Psychology, 50*(3), 317-329. Disponible en: https://www.redalyc.org/articulo.oa?id=284/28450492002

Rosales, C., Díaz-Cabrera, D., & Hernández-Fernaud, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PLoS ONE* 14(9): e0222694. https://doi.org/10.1371/journal.pone.0222694

Schraeder, M., Becton, J.B., & Portis, R.(2007). A Critical Examination of Performance Appraisal: An Organization's Friend or Foe? *The Journal for Quality and Participation*, 30, 20-25.

Silvestre, E. (2017). Construcción y validación empírica de una escala de clima organizacional universitario. *Interamerican Journal of Psychology, 51*(1), 44-59. Disponible en: https://www.redalyc.org/articulo.oa?id=284/28452860005

Smith, D.E. (1986). Training Programs for Performance Appraisal: A Review. *The Academy of Management Review*, Vol. 11, 1, 22-40. http://dx.doi.org/10.2307/258329

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155. http://dx.doi.org/10.1037/h0047060

Steiner, D., & Rain, J. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136-142. http://dx.doi.org/10.1037/0021-9010.74.1.136

Vélez, S., Rosario, I., Méndez, V., & Vargas, L. (2016). Familia, capital humano, y Psicología Industrial/Organizacional. *Interamerican Journal of Psychology*, 50(3), 433-440. Disponible en: https://www.redalyc.org/articulo.oa?id=284/28450492011

Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes, 70*(2), 95-103. http://dx.doi.org/10.1006/obhd.1997.2698

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189-205. http://dx.doi.org/10.1111/j.2044-8325.1994.tb00562.x

Received: 2019-11-25
Accepted: 2021-01-26