1 2 9 0

## UNIVERSIDADE Ð COIMBRA

Mariana Lobo Moreno

# EMPIRICAL RESEARCH ON MULTISOURCE FEEDBACK IN PERFORMANCE APPRAISAL PROCESSES: A SYSTEMATIC LITERATURE REVIEW

## VOLUME 1

Fevereiro de 2020

**Agradecimentos**

Gosto de pensar que este momento se trata de enaltecer as pessoas que connosco percorreram um caminho nem sempre fácil.

Um muito obrigada a ti, mãe. Mais do que financeiramente, deste-me em dobro todo o amor, apoio, força, perseverança e determinação que por vezes, sozinha, não conseguia ter. Obrigada por suportares todas as ausências, todas as ansiedades, todos os medos. Obrigada por me dares coragem para continuar, sempre. O canudo é meu, mas a vitória é nossa!

A vocês, Rita, Helena, Inês, Sara, Elsa, Bruno e Pedro, que fizeram desta experiência e cidade, algo mais caloroso, mais amável, mais amigo, um muitíssimo obrigada. Por todas as sessões de estudo, por todos os momentos de descontração, por toda a força que por vezes ia buscar a vós. Quem sou hoje e a minha memória de Coimbra, será para sempre um reflexo de cada um de vós. Fizeram de mim melhor pessoa, melhor amiga, melhor psicóloga.

À RAJA! Ao 1º Esquerdo! Ás minhas colegas de quarto! Ás minhas madeirenses! À cave! Obrigada! Tornaram esta cidade a minha segunda, se não, a primeira. Ensinaram-me tanto, deram-me tanto. Levo cada um(a) de vós no meu coração! Parte de mim cresceu devido a vós e às pessoas que por já cá passaram. Não conheci a palavra solidão por vossa causa. Coimbra é Coimbra devido a vocês!

Á minha madrinha, Katia, que foi o meu suporte constante. A ti, sei que te levo para a vida. Apesar da distância, da vida atribulada, foste e és a primeira pessoa a quem recorro. Foste o melhor que Coimbra me deu, e és o melhor que levarei.

Aos meus professores, Doutora Teresa Rebelo, Doutor Paulo Renato, Doutora Leonor Pais, e em especial aos meus orientadores Doutor Nuno Rebelo dos Santos e Doutora Leonor Pais, um muitíssimo obrigada. Fizeram-me apaixonar e (re)apaixonar todos os dias por um trabalho, por uma área tão querida. Espero um dia poder transmitir ao mundo todo o carinho, toda a paciência e toda a paixão que sentem por psicologia!

E finalmente, a ti Coimbra! Obrigada, por tudo e todos!

"Uma vez Coimbra, para sempre saudade"

**Empirical research on multisource feedback in performance appraisal processes: A systematic literature review**

Abstract

The current organizational, strategic and political context and situation increased the need for an objective and fair performance appraisal process. The inclusion, by organizations and managers, of multisource feedback in performance appraisal systems responds to these needs. This systematic literature review aims to describe and further characterize empirical research of multisource feedback included in performance appraisal processes or when used as a measurement tool. Using the research platforms B-ON, EBSCO host and ProQuest and the keywords: 'multisource or multi-source', 'performance', 'appraisal or evaluation or assessment', 162 articles were retrieved. After applying the inclusion and exclusion criteria, 37 articles were retained for analysis. The results obtained demonstrate that the implementation of multisource feedback in performance appraisal processes is increasing rapidly. It has an intrinsically political component, since it allows organizations to reach the concept of decent work and the goals set for the year 2030, and a strategic component, once that by providing a full range of feedback information to employees, they can use it to enhance their performance within organizations. To deepen the knowledge about this organizational phenomenon, more studies about the psychometric properties of multisource feedback in performance appraisal, its cultural sensitivity, link with organizational strategy, policy and philosophy as well, its long-term impact, should be done.

*Keywords*: multisource feedback; performance appraisal; literature review

**Investigação empírica sobre o feedback de múltiplas fontes nos processos de avaliação de desempenho: Uma revisão sistemática de literatura**

Resumo

O contexto organizacional, estratégico e político atual incrementou a necessidade de um processo de avaliação de desempenho objetivo e justo. A inclusão, pelas organizações e pelos gestores, do feedback de múltiplas fontes na avaliação de desempenho vem responder a estas necessidades. O objetivo desta revisão sistemática de literatura é descrever e caracterizar a investigação empírica sobre o feedback de múltiplas fontes incluído nos processos de avaliação de desempenho ou quando utilizado como ferramenta de medição. Recorrendo às plataformas de pesquisa B-ON, EBSCO host e ProQuest e às palavras-chave: '*multisource* ou *multi-source*' , '*performance*', '*appraisal* ou *evaluation* ou *assessmen*t', 162 artigos foram obtidos. Após a aplicação dos critérios de inclusão e exclusão, 37 artigos foram mantidos para análise. Os resultados obtidos demonstram que a implementação do feedback de múltiplas fontes na avaliação de desempenho está a aumentar rapidamente. Este tem uma componente política, uma vez que permite que as organizações se aproximam do conceito de trabalho digno e dos objectivos definidos para o ano de 2030, e uma componente estratégica, uma vez que ao fornecer feedback de múltiplas fontes, inserido na avaliação de desempenho, aos colaboradores permite que incrementem o seu desempenho na organização. De modo a aprofundar o conhecimento sobre este fenómeno organizacional, devem-se realizar mais estudos sobre as propriedades psicométricas do feedback de múltiplas fontes inserido na avaliação de desempenho, a sua sensibilidade cultural, a sua ligação com a estratégia organizacional, política e filosofia, bem como o seu impacto a longo prazo.

*Palavras-chave:* feedback de fontes múltiplas; avaliação de desempenho; revisão de literatura

**Index**

**Tables and Figures Index**

## I. Introduction

Workers are vital elements since they undertake actions that lead to organizational goals achievement. Motivated and well-trained workers are essential to secure development and organizational success (Görün et al., 2018). To provide excellent training, it is crucial to determine the worker's weaknesses and strengths. By doing that, emphasis can be put into the structured actions of intentional development (dos Santos & Pais, 2015) such as training, which meets the worker's needs. Thus, performance and its measurement is a core tool to determine the current performance and identify those needs. Performance appraisals, included in human resource management, have a crucial role in overall organizational development (Baruch & Harel, 1993; Garg, 2018).

Defining performance has been a difficult task (dos Santos, 2011; Kline & Sulsky, 2009). However, "the total expected value to the organization of the discrete behavioral episodes that an individual carries out over a standard period of time" (Motowidlo, 2003, p. 39), it is so far a general, clear and useful definition. This concept covers the productive work behavior, the counterproductive work behavior (dos Santos, 2011), and both the qualitative and quantitative worker's contributions, of motivational or non-motivational nature (Görün et al., 2018; Rynes et al., 2005). Assessment is related to a variety of concepts, programs and procedures integrating measurement, evaluation, correlation and prediction of one or more variables and criteria as well as organizational performance standards (Baruch & Harel, 1993).

Fletcher (2001) defines performance appraisal, which can be informal or formal, as the "general heading for a variety of activities through which organizations seek to assess employees and develop their competence, enhance performance and distribute rewards" (p. 474). In the present study, only the formal performance evaluation system will be considered, and it consists of the evaluation of employee performance against the pre-determined organizational criteria, and of the feedback of the appraisal results (Görün et al., 2018). A formal performance appraisal system is a "valuable contribution to the purposeful development process of organizations" (dos Santos, 2011, p. 589) since it can provide objectivity, transparency, and fairness in the overall picture, and interconnection among the several human resources practices (dos Santos, 2011). Performance appraisal is vital for the organization, for the appraiser and for the appraisee. For the organization,

because it satisfies the needs of information about employees' abilities, qualifications, training needs and potential, and it improves communication, motivation and performance. It allows for better planning by providing advanced data. For the appraiser, since it enhances staff performance and eliminates or prevents problems. For the appraisee, because it allows the identification of weaknesses and strengths' areas, allows discussing complaints and issues, and to discuss and focus on development needs (Baruch & Harel, 1993).

Traditional performance appraisal, which consists of a performance evaluation through a single source - the hierarchical superior - evolved into multisource performance appraisal (Edwards, 1996; Loredana & Mirabela, 2015), being its use increasing significantly in and since the 90s (Atwater et al., 2002; Görün et al., 2018). This transition was due to the growing number of employees in the organizations, the development of information and expertise areas (Görün et al., 2018), changes in the organizational structure, processes and culture, discontent with the traditional performance appraisal system (Fletcher, 2011), and the recognition of multisource performance appraisal or MSF advantages, and benefits (Atwater et al., 2002; Entrekin & Chung, 2001; Görün et al., 2018).

MSF or multisource performance appraisal system refers to the formal appraisal of an employee's performance by more than one source and to the communication of the results to the employee, after comparing the ratings with predetermined organizational criteria or self-assessment (Edwards, 1996; Loredana & Mirabela, 2015). These multiple sources may be internal, for example, subordinates, peers, self or supervisors, or external, for example, clients or patients (Atwater et al., 2002; Loredana & Mirabela, 2015). The 360° degree feedback is a subtype of MSF (Loredana & Mirabela, 2015). It is seen as a "process in which subordinates, peers, supervisors, and/or customers provide anonymous feedback to recipients" (Atwater et al., 2002, p. 193), differing from MSF since not all sources can be applied to all employees (London & Smither, 1995).

Current trends emphasized the need to evaluate employee performance from multiple perspectives (Entrekin & Chung, 2001). MSF is typically used for development purposes, and its positive impact and advantages have been progressively recognized by managers (Atwater et al., 2002; Entrekin & Chung, 2001). It is seen as more objective, fair, reliable

and valid (Baruch & Harel, 1993; Flint, 1999; Görün et al., 2018). The quality of performance appraisal increases with the combination of multisource appraisals (Baruch & Harel, 1993), as supported by empirical research (cf. Harris & Schaubroeck, 1988; Weekley & Gier, 1989). In addition to what had been mentioned, several advantages and benefits are associated, but not limited to, total quality management, understanding by employees of others' opinions about their performance, clarification of employees and organization´s expectations, a brighter, fairer and more informative feedback and an effective communication (Görün et al., 2018; London & Smither, 1995), the discouragement of disagreements between employees, a more comprehensive and effective assessment, a support tool for continuous learning (Edwards, 1996), and the proven improvement in post MSF performance (Smither et al., 1995).

The current organizational context influenced and modified by social, political, economic and technological factors, has become more ambiguous, dynamic and demanding. To respond to current demands, employees are expected to be proactive and autonomous, giving emphasis and relevance to strategies such as job crafting, empowering leadership, and incremental employee involvement and participation in the various organizational processes. These are positively associated with performance (Rohlfer, 2018; Sharma & Kirkman, 2015; Srinivas & Ashok, 2018).

MSF in performance appraisal is closely linked to these strategies' success since it allows their correct implementation, facilitation, and the benefits enhancement associated with each one (Rohlfer, 2018; Sharma & Kirkman, 2015; Srinivas & Ashok, 2018), making it technically desirable. Due to the combination of different viewpoints, it empowers employees as it strives for their involvement and participation in decision-making processes, cultivating a feeling of competence, autonomy, and meaning of work in them (Rohlfer, 2018; Sharma & Kirkman, 2015). By creating conditions for intra-organizational communication and providing information on improvement areas, it nurtures feelings of control and ownership thus, enhancing the competencies and skills development of those involved (Atwater et al., 2002; Loredana & Mirabela, 2015; Rohlfer, 2018; Sharma & Kirkman, 2015). MSF in performance appraisal is also aligned with job crafting since it serves its purpose more effectively and objectively, that is, the reshape of job boundaries to respond to environmental demands. It allows the employee to understand his progress and improvement areas and enable the organization to guide

them in job design and to understand their progress and success (Atwater et al., 2002; Lee & Lee, 2018; Srinivas & Ashok, 2018).

With the changes occurred, work has been conquering a central and vital role in people's lives. The concept of DW has received several contributions throughout the International Labour Organization's history (ILO) and the United Nations action (dos Santos, 2019; Ferraro et al., 2016; Pereira et al., 2019) being defined through seven dimensions, which covers all 11 substantive elements (dos Santos, 2019; Ferraro et al., 2018). The DW approach is inclusive and universal, since it applies and extends to all types of organizations, is based on human rights at work and ethical claims, inserts work in a wide economic, political and social context, and is transposable to international policies and to the "global" (Sen, 2000). The workers' fair treatment is one of its components, considering the 11 substantive elements proposed by ILO and the first dimension of the Ferraro et al.'s (2018) model. While the seventh substantive element is named *equal opportunity and treatment in employment,* the first DW dimension is called *fundamental principles and values at work*. Both are concerned with fair treatment of all at work being aligned with one of MSF in performance appraisal objectives: to treat workers fairly (dos Santos, 2011).

The Sustainable Development 2030 Agenda, with its origins in the Millennium Development Goals and in the United Nations Global Compact, aims to guide human development until 2030 consisting of 17 sustainable development goals. The concept of DW is included in the eighth objective: *Promote sustainable, inclusive and sustainable economic growth, full and productive employment and decent work for all* (United Nations, 2000). A fair performance appraisal system is tuned with the requests and considerations of the treaties, declarations and principles set out above and its study, implementation and practice is essential to achieve the DW ideal and to achieve the 2030 goals. Thus, MSF in performance appraisal can contribute to that endeavor.

Performance appraisal systems have been studied for several decades. Numerous literature reviews have shown this (cf. Karabat et al., 2013; Reinhardt, 1985; Shahzileh & Aghajan, 2015; Stratton, 1988) however, none had focused on MSF in performance appraisal. Therefore, this work, besides being more recent, consequently updating

knowledge, will focus on MSF and aims to describe and further characterize the empirical research about MSF, in its various forms, in performance appraisal.

## II. Method

PRISMA Protocol is a guide to ensure methodological rigor in systematic reviews in the health care area (Moher et al., 2015). Although psychology is a different realm from health care, we found it inspiring and useful to adapt this protocol to guide our systematic literature review on MSF in performance appraisal. The studies presented and discussed in the subsequent sections were retrieved, until December 31$^{st}$, 2019, from the following academic publication databases: Academic Search Ultimate, Business Source Ultimate, Communication Source, Criminal Justice Abstract Full Text, ERIC, EDS (Ebsco Discovery Service), EDS Publication Finder, SocINDEX, Sage Premier, Academic Search Complete, American Chemical Society, American Institute of Physics, Annual Reviews, Association for Computing Machinery, Business Source Complete, Coimbra University Press, Current Contents (ISI), Elsevier, Essencial Science Indicators (ISI), Eric, IEEE, Institute of Physics, ISI Proceedings, Journal Citation Reports (ISI), LISTA, Nature, Sage, Royal Society of Chemistry, Springer, Society for Industrial and Applied Mathematics, Taylor & Francis, Web of Science, Wiley, Zentrablatt, Academic Search Complete, Annual Reviews, Cinahl, Business SourceComplete, Coimbra University Press, Dynamed, Health Business Elite, ISI Proceedings, Journal Citation Reports (ISI), Medline with full text, Psychology & Behavioral Science, Web of Science (ISI) included in B-ON search platform, Academic Search Ultimate, Business Source Ultimate, Communication Source, Criminal Justice Abstracts with Full Text, eBook Collection (EBSCOhost), ERIC, Library & Information Science Source, SocINDEX with Full Text; SPORTDiscuss with Full Text; EconLit with Full Text included in EBSCO host search platform and Ebook Central, Education Database, Psychology Database contained in ProQuest search platform.

The Boolean expression was: 'multisource OR multi-source' in abstract AND 'performance' in title AND 'appraisal OR evaluation OR assessment' in abstract. The intention was to avoid literature concerning other forms of assessment or evaluation

**SLR Questions**

- What were the study purposes?

- What kind of samples (employees of public or private organizations, doctors, teachers, managers, team leaders, project teams, engineers, nurses and students) have been the object of these studies?

- How has been multisource feedback operationalized in these studies?

- What other variables have been measured?

- Which research designs have been employed?

- What were the main findings of these studies?

- What kind of limitations, suggestions for future studies and practical implications have been pointed out?

**Eligibility Criteria**

- English papers only;

- Peer-reviewed academic journals;

- Focus on multisource feedback or on multisource feedback as a measurement tool;

- Only empirical research.

**Search strategy**

| Sources limitation | Boolean expression |
|---|---|
| - EBSCO host<br><br>- B-ON<br><br>- ProQuest | 'Multisource OR Multi-source' (in abstract) AND 'Performance' (in title) AND 'Appraisal OR Evaluation OR Assessment' (in abstract) |

**Total articles retrieved**

- Through EBSCO host $n = 64$

- Through B-ON $n = 78$

- Through ProQuest $n = 20$

**Articles excluded $n = 48$**

**Articles excluded on the first round $n = 34$**

- Not focused on multisource feedback or did not used it as a measurement tool $n = 4$

- Related to another area $n = 30$

**Articles after duplicates removed**

$n = 85$

**Articles screened**

$n = 85$

**Articles excluded on the second round $n = 14$**

- Not empirical studies $n = 14$

**Articles included in the review**

$n = 37$

Figure 1: *Flow diagram of articles selection*

focus, such as programs, objects or mechanisms. For this review, inclusion criteria were English publications in academic journals with peer reviewing.

The searches performed retrieved 162 articles. After applying the inclusion criteria and removing the duplicates, 85 articles were kept for the next round. Upon reading titles and abstracts, we have decided to exclude the articles that were not focused or did not used MSF as measurement tool or assessment, remaining 51 articles for analysis. From those, only empirical studies were retained, reducing to 37 articles. The remaining articles were not considered because they were related to another field study, did not focus on MSF in performance appraisal or did not used it as a measurement tool. The following sections are the result of the analysis of 37 articles (figure 1).

## III. Results

The articles were analyzed according to the following aspects: (a) the study purposes, (b) the samples (data sources) of the empirical studies, (c) the operationalization of multisource, (d) the variables measured besides performance, (e) research designs, (f) main findings, (g) limitations mentioned by the authors, (h) suggestions made by the authors concerning future research and (i) practical implications mentioned by the authors.

Concerning the studies that explore MSF in performance appraisal, Table 1 shows the study purpose, the data sources description, multisource operationalization, and other variables measured.

Table 1

*Study purpose, data sources description, multisource operationalization, and other variables measured*

| Authors and year of publication | Study purpose | Sample dimension ($n$) | Sample description | Multisource operationalization | Other variables measured |
|---|---|---|---|---|---|
| Anand et al. (2018) | To explore leader-member exchange as a mediator and within-group value congruence as a moderator of i-deals-performance outcomes relationship. | 289 | Software engineers from an Indian technology firm. | Self and managers (questionnaires). | Perceptions of i-deals and in-role performance, Leader-member exchange, Within-group value congruence, and Citizenship behavior. |

| | | | | | |
|---|---|---|---|---|---|
| Araújo & Taylor (2012) | To determine the influence of emotional and social competence on job performance through multisource feedback assessment. | 36 | Peruvian copper refinery employees. | Supervisors, peers, subordinates, and self [Emotional Competence Inventory 2.0 (Boyatzis et al., 2000; Boyatzis & Sala, 2004; Goleman et al., 2002; Sala, 2002; Wolff, 2005)]. | Emotional and social competences, and Job Performance. |
| Bindels et al. (2019) | To evaluate the Group Monitor's psychometric properties. | 254 | Dutch physicians. | Staff, peers, self, and hospital managers (Group Monitor tool). | Group Monitor's psychometric properties. |
| Buccieri et al. (2008) | | | | Phase 1 | |
| | To examine content validity of the assessment tool. | 91 | American clinical educators. | Not applied. | Behavioral competencies, and Demographic information. |
| | | | | Phase 2 | |
| | To design a tool for directors of clinical educators use after obtaining feedback on their performance. | 591 | American directors of physical therapist educators. | Center coordinators of clinical education, and clinical instructors (a tool developed in this study). | The tool developed. |
| Craig & Kaiser (2003) | | | | Phase 1 | |
| | To investigate the consequences of violating the independent rating assumptions in item response theory-based analysis under several conditions. | 1 000* | Managers enrolled in a leadership development program. | Direct reports subordinates [Benchmarks 360° leadership Assessment Instrument (Lombardo & McCauley, 1994; McCauley et al., 1989)]. | Sample size, and Test size. |
| | | | | Phase 2 | |
| | Equal to phase 1. | 200* | Equal to phase 1. | Equal to phase 1. | Equal to phase 1. |
| Darr & Catano (2008) | To compare developmental multisource feedback ratings against the ones obtained from a structured behavioral interview. | 77 | Canadian employees with experience in executive positions, and applied | Supervisors, peers, subordinates, and self (ad hoc questionnaire redesign by '20/20' Insight | Corporate/senior executive leadership, and Organization core competencies. |

| | | | | | |
|---|---|---|---|---|---|
| | | | | for a promotion. | Gold'' program). |
| deLeon & Ewen (1997) | To investigate the employees' attitudes towards a new appraisal system, namely the multi-source appraisal system. | 221 (pre-test) 216 (post-test) | Person operations office of an American federal agency. | Peers, self, direct-reports subordinates, and supervisor (ad hoc questionnaires by workers). | Attitudes toward the appraisal system. |
| Dupee et al. (2011) | To investigate the influence of multisource feedback on performance appraisal satisfaction in an inpatient nursing unit. | 11 (pre-test) 9 (post-test) | American nurses with performance appraisal experience. | Supervisor, professional peer, self, and workplace peer (survey). | Satisfaction with the new process. |
| Facteau & Craig (2001) | To test if a multisource performance appraisal instrument shows invariance across different groups of raters. | 1 883 | American managers. | Self, peer, supervisor, and subordinate (multisource appraisal form). | Variance across groups of raters. |
| Fang et al. (2013) | To examine if medical students enrolled via different programs had different performances. | 182 | Taiwan medical students. | Self, peer, nursing staff, visiting staff, and chief resident (questionnaire). | Admission programs, and Medical student's performances. |
| Gabriel et al. (2015) | Study 1 | | | | |
| | To understand how service familiarity operates as a boundary condition for the impact of employee positive emotional displays on service performance. | 114 | American employees. | Coworkers [ad hoc questionnaire based on Bettencourt et al' s (2001) work], and self-report [ad hoc questionnaire items based on Diefendorff et al.'s (2006) and Gutek's (1995) work]. | Employee positive emotional displays, Service familiarity, Service performance, and Control variables. |
| | Study 2 | | | | |
| | Equal to study 1. | 208 | Equal to study 1. | Costumers [ad hoc questionnaire items based on Gutek's (1995) work and survey], and trained research assistants (videotapes). | Equal to study 1. |
| | Pilot Study | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Grandey et al. (2018) | To confirm three assumptions about occupational-racial stereotype incongruence. | 124 | Mostly female and Caucasian, ranging from 18 to 70 years old. | Not applied. | Race influence on interpersonal warmth and fit for occupation perception, and Occupational stereotype. |
| | Study 1 | | | | |
| | To confirm three assumptions about occupational-racial stereotype incongruence via video stimuli. | 113 | Black and White participants. | Customers (survey). | Racial differences in performance, Interpersonal warmth, and Emotional labor. |
| | Study 2 | | | | |
| | To confirm three assumptions about occupational-racial stereotype incongruence via field surveys. | 311 | Equal do study 1. | Self, and supervisors (ad hoc rating tool done by a firm). | Equal to study 1. |
| Gutermann et al. (2017) | To investigate how leaders' work engagement can spread to followers. | 88 | Team leaders in a German organization. | Self, and subordinates (questionnaires). | Work engagement, Leader-member exchange, Turnover intention, and Performance. |
| Hoffman & Woehr (2009) | 1) To answer the following questions: "What are source effect measuring?", and 2) To investigate construct validity evidence for multisource performance rating. | 404 | American managers enrolled in an executive master of business administration program. | Supervisor, subordinate, and peer (multisource performance rating instrument). | Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980), California Psychological Inventory (Gough & Bradley, 1996), and Assessment Center. |
| Jong et al. (2019) | To determine if Queen's Simulation Assessment Tool can be used to provide multisource feedback using a standardized simulation case. | 34 | English resident osteopathic doctors in their $2^{nd}$ to $4^{th}$ postgrad year. | Self, peer, supervisors, and senior experts (ad hoc Queen's Simulation Assessment Tool applied using a simulation task). | Years of residence (1 to 4). |
| Lev & Koslowsky (2012) | To examine if on-the-job embeddedness is a potential mediator in predicting performance within an educational framework. | 115 | Israeli junior and senior-high-school teachers. | Self, school principal, and peers (questionnaires). | Job performance, On-the-job embeddedness, and Conscientiousness. |

| | | | | | |
|---|---|---|---|---|---|
| Li & Wong (2008) | To investigate the implementing processes of performance indicators and identify some effective practices that might be used in other societies. | 284 | Japanese kindergarten teachers or child-care workers. | Self-evaluation portfolios, classroom observations [Child Involvement Scale (Laevers, 1993), and Adult Engagement Scale (Laevers 1996)]. | Teachers improvement. |
| Lyde et al. (2016) | To determine and understand faculty perceptions of the multi-source method of evaluating teaching performance. | 13 | American tenured and tenure track faculty members. | Students, and self (student evaluations, instructor portfolio, and a reflection on formative external reviews). | Faculty perceptions of the multi-source method for performance evaluation. |
| Narayanan et al. (2018) | Group 1 | | | | |
| | 1) To test the reliability and validity of two multisource feedback questionnaires and the data derived from them, and 2) "Identify how doctors can use data for professional development purposes". | 2449 | Australian doctors that undertook multisource feedback as a vocational training requirement. | Patients, colleagues, and self (patient questionnaire, colleague questionnaire, self-evaluation using colleague andpatient questionnaires). | Doctor groups, Professional behavior and values, Clinical competence, and Knowledge of technical procedures. |
| | Group 2 | | | | |
| | Equal to group 1. | 1890 | Australian doctors that undertook multisource feedback as a continuing professional development activity (voluntarily) | Equal to group 1. | Equal to group 1. |
| | Group 3 | | | | |
| | Equal to group 1. | 375 | Australian doctors that undertook multisource feedback compulsory. | Equal to group 1. | Equal to group 1. |
| | Group 4 | | | | |
| | Equal to group 1. | 1888 | Australian doctors that undertook | Equal to group 1. | Equal to group 1. |

| | | | | | |
|---|---|---|---|---|---|
| | | | multisource feedback as a regulatory requirement. | | |
| Noonan et al. (2011) | To develop a multi-source tool to use with consultant anaesthetists. | 15 | Anaesthetists. | Chief anaesthetist, assistants, trainees, and chief nurse (ad hoc survey). | Categories of behaviors. |
| Nuryanti et al. (2017) | Phase 1 | | | | |
| | To develop a nurses' multisource performance assessment instrument. | 53 | Indonesian nurses. | Not applied. | Instrument psychometric properties. |
| | Phase 2 | | | | |
| | Equal to phase 1. | 47 | Equal to phase 1. | Administrator, head of nurses, self, and patients (ad hoc survey). | Instrument quality value. |
| Overeem et al. (2012a) | To determine which factors impact on specialists' reported change in response to multisource feedback. | 456 | Dutch medical specialists. | Colleagues, coworkers, patients, and self (reflective portfolio, interview and questionnaires). | Self-report change in response to multi-source feedback. |
| Overeem et al. (2012b) | To assess the psychometric properties of three new multisource feedback instruments, the influence of sociodemographic variables, and the link between self and others' appraisals. | 146 | Dutch surgical and medical specialty physicians. | Peers, non-physician co-workers, patients, and self (reflective portfolio, interview, and questionnaires). | Initial psychometric properties three new instruments based on multisource feedback instruments. |
| Sargeant et al. (2011) | To understand how physicians used Nova Scotia Physician Achievement Review medical colleague questionnaire. | 23 | Canadian family and specialist physicians who were raters more than once. | Peers, patients, coworkers, and self (ad hoc questionnaire included in Physician Achievement Review). | Sources of data accessed, and Examples of behaviors given by raters. |
| Schuh et al. (2018) | Study 1 | | | | |
| | To examine the effect of employee innovative work behavior and the social context between supervisors and employees and its influence on performance evaluations. | 29 | Chinese engineer team leaders. | Subordinates, colleague, and supervisor (survey). | The link between innovative work behavior, leader-member exchange (LMX), performance ratings, and supervisors/subordinates' perceptions of LMX. |

| | | | | | |
|---|---|---|---|---|---|
| | | | Study 2 | | |
| | Equal to study 1. | 177 | Chinese full-time employees. | Self (survey). | Equal to study 1. |
| Seaburg et al. (2016) | To examine associations between the number of internal medicine resident's publications and their clinical performance. | 308 | American 3rd year medicine residents. | Peers, senior medical residents, faculty, and non-physician professionals (survey). | Internal medicine resident's publications. |
| Selvarajan & Cloninger (2012) | To examine the relationship between appraisal characteristics, perceived reactions to them, and appraisal outcomes. | 203 | Mexican full-time workers in an executive program. | Participant´s declaration that had been assessed by more than one source. | Appraisal characteristics and perceived reactions to them, and Appraisal outcomes. |
| Sung & Choi (2018) | To examine the effects of training and development in employee outcomes, and firm innovative performance considering situational boundaries and contingencies. | 325* | Korean employees, directors, and managers. | Self and supervisors (survey). | The link between firm innovative performance, workers' perceptions of training and development (T&D), and voluntary participation in T&D. |
| Treadway et al. (2013) | | | Study 1 | | |
| | To examine the moderating role of political skill on the relationship between performance and interpersonal power. | 97 | Canadian employees. | Coworkers (single-item measure), and self-evaluation [18-item Political Skill Inventory (Ferris et al. (2005)]. | Power, Performance, and Political skills in organizations. |
| | | | Study 2 | | |
| | Equal to study 1. | 384 | American employees. | Organization's human resources department. | Equal to study 1. |
| van der Meulen et al. (2017) | To examine the validity, reliability, and feasibility of Inviting Co-workers to Evaluate Physicians Tool. | 218 | Dutch physicians. | Peers, self, residents, and coworkers (ad hoc INviting Co-workers to Evaluate Physicians Tool). | Psychometric properties of INviting Co-workers to Evaluate Physicians Tool. |

| | | | | | |
|---|---|---|---|---|---|
| van Hooft et al. (2006) | To investigate the psychometric properties of multisource ratings using Intelligence Test, In-Basket Exercise, and Personality Questionnaire. | 195 | Dutch public employees. | Self, supervisor, and peer (ad hoc survey based on Thornton and Byham's (1982) dimensions). | Psychometric properties of the multi-source rating instrument. |
| van Veelen & Ufkes (2019) | To investigate circumstances under which demographic diversity facilitates performance. | 22 | Project teams in a Dutch undergraduate psychology program. | Expert supervisors (ad hoc survey based on an external grading system), and self (survey). | Diversity in gender and nationality. |
| Violato et al. (2008) | 1) To evaluate validity and reliability of Physician Achievement Review (PAR) instruments for family medicine and general practice, 2) To investigate changes in performance between the two PAR assessments, and 3) To determine if changes are associated with initial assessments and socio-demographic characteristics. | 250* | Canadian family doctors or general practitioners. | Patients, medical and non-medical colleagues, and self (ad hoc questionnaires developed and tested by the College of Physicians and Surgeons of Alberta-Physician Achievement Review program). | Psychometric properties, Performance change, and Socio-demographic data. |
| Weigl et al. (2019) | 1) To determine provider's teamwork evaluation during postsurgical care, 2) To determine agreement level between and within rater and ratee, 3) To determine the link among teamwork and provider outcomes, and 4) To determine the link between intra team agreement about performance and provider outcomes. | 137 | Postsurgical providers included in postsurgical teams. | Self and expert-observer (survey and expert observation). | Provider characteristics, Expert-observer evaluation, and Perceived demands, disruptions and stress during handover. |
| Whitaker (2012) | 1) To investigate the invariance of the task and contextual performance ratings across rating sources, and 2) To examine the effects of covariates on latent task and contextual performance variability. | 252 | American undergraduates working students. | Supervisor and self (survey). | Performance perceptions, and Demographic information. |
| Williams et al. (2017) | To determine the agreement level between a measure based on Dreyfus | 264 | American physicians in supervisory roles. | Leaders, peers, self, and reports (survey). | Agreement measure between the two instruments. |

| | | | | | |
|---|---|---|---|---|---|
| | and Dreyfus framework and a 360º assessment. | | | | |
| Wyland et al. (2016) | To examine how the work-school interface relates to work outcomes, and how work-school resources demand influence work-school interface. | 170 | American undergraduate working students. | Self and supervisor (survey). | Work-school and school-work conflict, facilitation, related demands and resource, Control variables, and Outcomes. |

Note 1: The samples identified with ´*´ were retrieved from already existing databases.

Note 2: The core concept is written in the same way as the authors wrote it in their studies.

The specificities of each study are mirrored in each study's purpose. However, it can be grouped in general study purposes namely, (1) the ones that investigate (a) the links between two or more variables, their mediators, moderators or facilitators, and (b) the influence, impact or effect of one or more variables, being one of them MSF, into others (22 of the total, for example (e.g.), Dupee et al., 2011), (2) the ones which evaluate psychometric properties, and develop tools such as, for example, invariance across constructs, validity and reliability, and MSF instruments or tools (18 of the total, e.g., Bindels et al., 2019), (3) the ones which examine specifically performance changes due to MSF and what factors facilitate it (three of the total, e.g., Violato et al., 2008), (4) the ones which examine how to implement performance appraisal, if an existing instrument can be applied or adapted, and how can serve multisource rating purposes (three of the total, e.g., Jong et al., 2019), and (5) the ones that assess an individual or group perceptions or attitudes towards a new or recently implemented multisource appraisal system (two of the total, e.g., Lyde et al., 2016). One of Narayanan et al. (2018)'s aims was to understand how MSF can be used for professional development purposes. Darr and Catano (2008) wanted to compare performance ratings obtained through MSF and structured behavioral interviews. Most researches were led upon one single purpose (30 of the total), belonging mostly to group 1 (13 of the total, e.g., Anand et al., 2018) and 2 (nine of the total, e.g., Bindels et al., 2019). Five investigations (e.g., Whitaker, 2012) were conducted upon two study purposes belonging mostly to group 2. Two investigations (e.g., Violato et al., 2008) were led upon three study purposes mainly belonging to group 2 (three of the total) and 3 (two of the total). Weigl et al. (2019) led their investigation upon four study purposes, belonging to group 2.

Studies were carried out with very diverse participants. Sample dimension, here understood as the individuals evaluated by others or themselves, ranges from $n = 9$ (Dupee et al. (2011) in post-test sample) to $n = 2449$ (Narayanan et al. (2018) in Group 1 sample), being the most frequent dimension in the order of hundreds (31 of the total, e.g., Fang et al., 2013). The majority of samples (47 of the total, e.g., Nuryanti et al., 2017) were collected by authors resorting or not to firms specialized in data collection however, three authors (e.g., Violato et al., 2008) extracted their samples from already existing databases. Eleven studies (e.g., deLeon & Ewen, 1997) enclose within more than one study or phase. It contained two to four samples having the majority (eight of the total, e.g., Buccieri et al., 2008) two samples, used as phases 1 and 2 samples (four of the total, e.g., Gabriel et al., 2015), pre and post-test samples (three of the total, e.g., deLeon & Ewen, 1997), or as study 1 and 2 samples (Craig & Kaiser, 2003).

The samples were retrieved in its majority from America (22 of the total), especially from the United States (16 of the total, e.g., Seaburg et al., 2016) and Canada (four of the total, e.g., Darr & Catano, 2008) followed by Mexico (Selvarajan & Cloninger, 2012), and Peru (Araújo & Taylor, 2012). Samples were also collected from European countries (nine of the total), especially from Netherland (six of the total, e.g., Bindels et al., 2019), United Kingdom (Jong et al., 2019), and Germany (Gutermann et al., 2017). Moreover, nine samples were retrieved from Asian countries namely, India (Anand et al., 2018), Taiwan (Fang et al., 2013), Israel (Lev & Koslowsky, 2012), Japan (Li & Wong, 2008), Indonesia (Nuryanti et al., 2017), China (Schuh et al., 2018), and Korea (Sung & Choi, 2018). Samples contemplated in Narayanan et al. (2018)'s investigation are from Australia. The ones collected by Craig and Kaiser (2003), Grandey et al. (2018), Noonan et al. (2011) and Weigl et al. (2019) have unknown origins.

Participants' professional occupations can be grouped in (1) physicians (25 of the total, e.g., Williams et al., 2017), (2) employees on a public or private organizations (11 of the total, e.g., Selvarajan & Cloninger, 2012), (3) managers, team leaders, and project teams (six of the total, e.g., Facteau & Craig, 2001), (4) teachers (four of the total, e.g., Lyde et al., 2016), (5) students (four of the total, e.g., Seaburg et al., 2016), (6) nurses (three of the total, e.g., Nuryanti et al., 2017), and (7) engineers (two of the

total, e.g., Schuh et al., 2018). Three studies used subjects enrolled in leadership or executive programs (e.g., Selvarajan & Cloninger, 2012), and three others used subjects already submitted to performance appraisals (e.g., Sargeant et al., 2011). Grandey et al. (2018) used three samples with Caucasian and Black participants.

Table 1 shows that performance was rated by internal (105 of the total) and external sources (12 of the total). Internal rating sources included peers (37 of the total) namely peers, colleagues or coworkers (27 of the total, e.g., Bindels et al., 2019), non-medical peers (seven of the total, e.g., Seaburg et al., 2016), medical peers (three of the total, e.g., Dupee et al., 2011) and residents (van der Meulen et al., 2017), and self (31 of the total, e.g., Whitaker, 2012). Moreover, it included hierarchical superiors (27 of the total) namely, supervisors (19 of the total, e.g., Williams et al., 2017), chief residents (three of the total, e.g., Fang et al., 2013), managers (two of the total, e.g., Selvarajan & Cloninger, 2012), senior experts (two of the total, e.g., Jong et al., 2019), and school principals (Lev & Koslowsky, 2012). Subordinates (10 of the total) namely, subordinates (seven of the total, e.g., Hoffman & Woehr, 2009), direct report subordinates (two of the total, e.g., deLeon & Ewen, 1997), and students (Lyde et al., 2016) were used as internal rating sources. External rating sources include patients (six of the total, e.g., Overeem et al., 2012a), external observers (two of the total, e.g., Gabriel et al., 2015), organization´s human resources department (three of the total, e.g., Treadway et al. (2013) in Study 3), and costumers (Gabriel et al., 2015). In Selvarajan and Cloninger's (2012) study, the performance was rated through a participant´s declaration that they had been assessed by more than one source, and in Grandey et al.'s (2018) pilot study, the performance was not evaluated. Each author, except Grandey et al. (2018) in theirs pilot study, resorted to one (five of the total, e.g., Craig & Kaiser, 2003), two (14 of the total, e.g., Weigl et al., 2019), three (six of the total, e.g., van Hooft et al., 2006), four (16 of the total, e.g., Overeem et al., 2012b), or five rating sources (Fang et al., 2013). Who resorted to a single rating source did it because (1) the investigation's results were the combination of two or more studies results (three of the total, e.g., Treadway et al., 2013), (2) participants declared that were rated by more than one source (Selvarajan & Cloninger, 2012), and (3) raters used a 360º instrument to assess participants (Craig & Kaiser, 2003). Studies that used two evaluators obtained their data from (a) self and hierarchical

superior's ratings (six of the total, e.g., Anand et al., 2018), (b) self and peer´s ratings (two of the total, e.g., Gabriel et al., 2015), (c) self and external sources' ratings (two of the total, e.g., Li & Wong, 2008), (d) self and subordinate's ratings (Lyde et al., 2016), (e) two peers (Buccieri et al., 2008), and (f) the combination of two external sources' ratings (Gabriel et al., 2015). Who used three rating sources obtained their data through (a) a combination of self, hierarchical superior and peer's ratings (three of the total, e.g., van Hooft et al., 2006), (b) subordinate, hierarchical superior and peer's ratings (Hoffman & Woehr, 2009), (c) external source, peer and self (Narayanan et al., 2018), (d) external source, peer and hierarchical superior (Williams et al., 2017), and (e) supervisor, non-medical peer and medical peer (Dupee et al., 2011). Who used four rating sources resorted to (a) the combination of hierarchical superior, peer, subordinate and self-ratings (five of the total, e.g., Araújo & Taylor, 2012), (b) self, two different types of peers and external source's ratings (four of the total, e.g., Overeem et al., 2012a), (c) hierarchical superior, peer, external rating source and self-ratings (Williams et al., 2017), (d) two different types of hierarchical superiors, self and peer's ratings (Jong et al., 2019), (e) two different types of hierarchical superiors, self and external sources' ratings (Nuryanti et al., 2017), (f) two different types of hierarchical superiors, subordinate and peer's ratings (Noonan et al., 2011), (g) three different types of peers and self-ratings (Violato et al., 2008), (h) three different types of peers and subordinate's ratings (Seaburg et al., 2016), and (i) two different typr of peers, self and hierarchical supervisor's ratings (Bindels et al., 2019). Fang et al. (2013) used five rating sources two obtain their data, namely, self, hierarchical superior and three types of peers.

Most studies (33 of the total, e.g., Sung & Choi, 2018) applied questionnaires, surveys, multisource appraisal forms or rating tools to operationalize multisource, followed by other instruments (10 of the total, e.g., Jong et al., 2019), portfolios (three of the total, e.g., Lyde et al., 2016), single-items (two of the total, e.g., Selvarajan & Cloninger, 2012), interviews (two of the total, e.g., Overeem et al., 2012a), videotapes or observations (two of the total, e.g., Gabriel et al. (2015) in Study 2), self-reflections (Lyde et al., 2016), and data provided by organization's human resource department (two of the total, e.g., Treadway et al. (2013) in Study 2). Fifteen studies used ad hoc tools, being 10 questionnaires (e.g., Noona et al., 2011), four instruments (e.g., van

der Meulen et al., 2017), and one rating tool (Grandey et al. (2018) in Study 2). The multisource operationalization of 35 studies was through one method (e.g., Anand et al., 2018). Five studies combined two methods, namely, a questionnaire plus a videotape or external observation (two of the total, e.g., Weigl et al., 2019), a combination of two different questionnaires (two of the total, e.g., Gabriel et al. (2015) in Study 1), and a single-item plus instruments (Treadway et al. (2013) in Study 1 and 2). Three studies used a combination of three methods specifically, portfolio, interview and questionnaires (two, e.g., Overeem et al., 2012b), and students evaluation, portfolio and self-reflection (Lyde et al., 2016). Narayanan et al. (2018) used four different types of questionnaires. Buccieri et al.'s (2008) phase 1, Grandey et al.'s (2018) pilot study, and Nuryanti et al.'s (2017) phase 1 did not use any method. The studies which resorted to instruments used one (seven of the total, e.g., Araújo & Taylor, 2012), combined with other methods or not, or two different instruments (Li & Wong, 2008). Instruments measured (1) emotional competence (cf. Araújo & Taylor, 2012), (2) child involvement and adult engagement (cf. Li & Wong, 2008), (3) political skills (cf. Treadway et al. (2013)'s Study 1 and 2), and (4) performance (cf. Craig & Kaiser, 2003).

The variables also measured varied according to each study's purpose. The most commonly analyzed, but not exclusively, were the work-related competences (14 of the total, e.g., Gutermann et al., 2017) such as behavioral competencies, conscientiousness, emotional labor, work engagement, corporate/senior executive leadership and professional behavior, and values. Fourteen studies (e.g., Weigl et al., 2019) analyzed demographic and control variables, and ten (e.g., Gabriel et al., 2015) analyzed performance such as medical student´s performance, and racial differences in performance. Psychometric properties (10 of the total, e.g., Craig & Kaiser, 2003) such as sample size, test size or psychometric properties of one instrument were analyzed in addition to relationships between two or more variables (five of the total, e.g., Grandey et al., 2018), and individual perceptions and attitudes towards a specific construct or instruments (four of the total, e.g., Anand et al., 2018). Hoffman and Woehr (2009) measured three tools, Sargeant et al. (2011) examined data sources accessed and examples of behaviors given by raters, and Wyland et al. (2016)

investigated work-school and school-work conflict, facilitation, related demands and resources, as well as school relevance and study outcomes.

Concerning the studies that explore MSF in performance appraisal, Table 2 shows the research design (type and description).

Table 2

*Research design*

| Authors and year of publication | Research design | Procedure description |
|---|---|---|
| Anand et al. (2018) | Cross-sectional: Correlational. | Data were collected via questionnaires. |
| Araújo & Taylor (2012) | Cross-sectional: Correlational. | Data were collected through the application of questionnaires. |
| Bindels et al. (2019) | Cross-sectional: Correlational. | Data were retrieved through Group Monitor administration. |
| Buccieri et al. (2008) | | Phase 1 |
| | Cross-sectional: Descriptive. | Data were retrieved through an online survey. |
| | | Phase 2 |
| | Equal to phase 1. | Data were collected through versions A and B of the tool developed. |
| Craig & Kaiser (2003) | Cross-sectional: Correlational. | Data were collected from direct reports´ ratings. |
| Darr & Catano (2008) | Cross-sectional: Descriptive / Correlational. | Data were collected through the 360º feedback program and a post-implementation interview. |
| deLeon & Ewen (1997) | Pre-experimental. | Data were collected with a pre and post-implementation of 360º system survey. Interviews were also done. |
| Dupee et al. (2011) | Pre-experimental. | Data were collected with a pre and post-test. |
| Facteau & Craig (2001) | Cross-sectional: Correlational. | Data were collected via a multisource appraisal form. |
| Fang et al. (2013) | Longitudinal: Expo-facto. | Data were collected through different methods being repeated for one year. |
| Gabriel et al. (2015) | | Study 1 |
| | Cross-sectional: Correlational. | Data were retrieved via questionnaires. |
| | | Study 2 |
| | Equal to study 1. | Data were retrieved via observation of videotapes |
| Grandey et al. (2018) | | Pilot Study |
| | Experimental. | There were four between-subject conditions. It was provided a photo with a description (held constant) and employee race, and occupation was manipulated. |
| | | Study 1 |
| | Cross-sectional: Correlational. | Data were collected via a customer survey. |
| | | Study 2 |

|  | Equal to study 1. | Data were collected via a self and a supervisor survey. |
|---|---|---|
| Gutermann et al. (2017) | Cross-sectional: Correlational. | Data were retrieved via questionnaires. |
| Hoffman & Woehr (2009) | Cross-sectional: Correlational. | Data were collected through the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980), the California Psychological Inventory (Gough & Bradley, 1996), an assessment center, and a multisource performance rating. |
| Jong et al. (2019) | Cross-sectional: Descriptive / Correlational. | Data were retrieved via a standardized resuscitation after overdose case with anchors used by the Queen's Simulation Assessment Tool, and comparison between different assessors. |
| Lev & Koslowsky (2012) | Cross-sectional: Descriptive. | Data were collected at time 1 through questionnaires and at time 2 through self-reports. |
| Li & Wong (2008) | Longitudinal: Descriptive. | Data were collected through document analysis, classroom observations, and group interviews (at the beginning and end of the project). |
| Lyde et al. (2016) | Cross-sectional: Descriptive. | Data were collected through post-implementation interviews. |
| Narayanan et al. (2018) | Cross-sectional: Descriptive. | Data were collected over time through questionnaires. |
| Noonan et al. (2011) | Cross-sectional: Descriptive. | Data were retrieved through a survey and a report. |
| Nuryanti et al. (2017) | | Phase 1 |
| | Cross-sectional: Descriptive. | Data were retrieved through a focus group discussion and questionnaires. |
| | | Phase 2 |
| | Equal to phase 1. | Data were retrieved using the instrument. |
| Overeem et al. (2012a) | Pre-experimental. | Data were collected through a reflective portfolio, a facilitative interview, and online questionnaires. |
| Overeem et al. (2012b) | Cross-sectional: Descriptive / Correlational. | Data were collected via questionnaires. |
| Sargeant et al. (2011) | Cross-sectional: Descriptive. | Data were collected through two four-hours focus groups: one with family physicians and one with specialist physicians. |
| Schuh et al. (2018) | | Study 1 |
| | Cross-sectional: Correlational. | Data were collected through surveys. |
| | | Study 2 |
| | Equal to study 1. | Equal to study 1. |
| Seaburg et al. (2016) | Longitudinal: Correlational. | Data were retrieved through the American Board of Internal Medicine certification examination, a mini clinical evaluation exercise, and clinical performance assessment. |
| Selvarajan & Cloninger (2012) | Cross-sectional: Correlational. | Data were collected through questionnaires. |
| Sung & Choi (2018) | Cross-sectional: Correlational. | Data were collected in three time points over five years via questionnaires. |
| Treadway et al. (2013) | | Study 1 |
| | Cross-sectional: Correlational. | Data were collected via questionnaires. |

|  | | Study 2 |
| --- | --- | --- |
|  | Equal to study 1. | Data were collected in time 1 through organizations' human resources records, and in time 2 through self-reports and coworkers' ratings. |
| van der Meulen et al. (2017) | Cross-sectional: Correlational. | Data were collected through questionnaires. |
| van Hooft et al. (2006) | Cross-sectional: Correlational. | Data were collected through a multi-source instrument, an intelligence test (MBS-Brain-H), an In-Basket exercise ("Zeezicht"), and a personality questionnaire (MBS-Quest). |
| van Veelen & Ufkes (2019) | Cross-sectional: Descriptive / Correlational. | Data were collected through admission records, external grading system, and through a survey. |
| Violato et al. (2008) | Longitudinal: Pre-experimental. | Data were collected at two moments over time through questionnaires. |
| Weigl et al. (2019) | Cross-sectional: Correlational. | Data were retrieved through surveys and observations. |
| Whitaker (2012) | Cross-sectional: Correlational. | Data were retrieved through a self-assessment, and a survey completed by the supervisor. |
| Williams et al. (2017) | Cross-sectional: Correlational. | Data were collected through a 360º survey. |
| Wyland et al. (2016) | Cross-sectional: Correlational. | Data were obtained at time 1, 2 and 3 via surveys and work-related performance data. |

Note: The core concept is written in the same way as the authors wrote it in their studies.

The research designs of the studies are diverse. Thirty-six studies have cross-sectional research designs (e.g., van der Meulen et al., 2017), four have longitudinal research designs (e.g., Li & Wong, 2008), three have pre-experimental research designs (e.g., deLeon & Ewen, 1997), and Grandey et al.'s (2018) pilot study has an experimental research design.

Twenty-three cross-sectional research designs are correlational (e.g., Whitaker, 2012), eight are descriptive (e.g., Lyde et al., 2016), and four are descriptive and correlational (e.g., Jong et al., 2019). The cross-sectional correlational studies collected data mostly through questionnaires, surveys or multisource appraisal forms (18 of the total e.g., van der Meulen et al., 2017), by instruments (three of the total, e.g. van Hooft et al., 2006), multiple rating sources (two of the total, e.g., Craig & Kaiser, 2003), the organization (two of the total, e.g., Treadway et al., 2013), and through videotapes or observations (two of the total e.g., Gabriel et al., 2015). It was used one retrieving method (18 of the total, e.g., Facteau & Craig, 2001), or a combination of two (five of the total, e.g., Bindels et al., 2019). The cross-sectional descriptive studies collected data mostly through questionnaires (five of the total, e.g., Narayanan et al., 2018), self-reports (two of the total, e.g., Lev & Koslowsky, 2012), focus-groups (two of the total, e.g., Sargeant et al., 2011), instruments (two of the total, e.g., Buccieri et al. (2008) in Phase 2), and

interviews (Lyde et al., 2016). It resorted to one (six of the total, e.g., Lyde et al., 2016), or a combination of two retrieving methods (three of the total, e.g., Lev & Koslowsky, 2012). The cross-sectional descriptive and correlational studies collected data mostly through questionnaires or surveys (two of the total, e.g., van Veelen & Ufkes, 2019), 360° programs (Darr & Catano, 2008), interviews (Darr & Catano, 2008), ad hoc instruments (Jong et al., 2019), admission records and external grading systems (van Veelen & Ufkes, 2019). It was used one (two of the total, e.g., Overeem et al., 2012b), a combination of two (Darr & Catano, 2008), or of three methods (van Veelen & Ufkes, 2019). The longitudinal pre-experimental study (Violato et al., 2008) collected data through questionnaires. The longitudinal expo-facto study collected data through instruments (Fang et al., 2013), the descriptive collected data through document analysis, interviews and observations (Li & Wong, 2008), and the correlational collected data through a combination of three tests (Seaburg et al., 2016). Longitudinal studies collected data using one (two of the total, e.g., Fang et al., 2013), or a combination of three methods (two of the total, e.g., Li & Wong, 2008).

Pre-experimental studies collected data through tests (Dupee et al., 2011), interviews (two of the total, e.g., deLeon & Ewen, 1997), portfolios (Overeem et al., 2012a), and surveys (two of the total, e.g., Overeem et al., 2012a), using one (Dupee et al., 2011), a combination of two (deLeon & Ewen, 1997), or of three methods (Overeem et al., 2012a). In Grandey et al.'s (2018) experimental investigation, data was collected through a photo description held constant, and through employee race and occupation, which were manipulated.

Concerning the studies that explore MSF in performance appraisal, Table 3 shows the main findings.

Table 3

*Main findings*

| Authors and year of publication | Main findings of the studies |
| --- | --- |
| Anand et al. (2018) | Leader-member exchange mediated the positive link between i-deals, in-role performance and organizational citizenship behavior. Within-group value congruence moderated the link between i-deals and in-role performance. When within-group value congruence was higher, the link was weaker. |
| Araújo & Taylor (2012) | Job performance, when rated by others, was influenced and predicted by emotional and social competences (ESC). Self-rated ESCs were positively correlated with self-rated transparency. |

| | |
|---|---|
| Bindels et al. (2019) | Group Monitor was internally consistent for all raters and had good construct validity. For staff, peers, and managers, seven to nine raters, eight to 15 raters, and six to seven raters were needed, respectively. The interpretation of the subscales differed for all rater classes. Four performance subscales were identified for all raters and a fifth was identified for self and staff. |

**Buccieri et al. (2008)**

<div align="center">Phase 1</div>

Raters were diverse in practice settings, roles, and years of experience. They could give moderate to significant feedback in the Administration and Teaching categories. They preferred to complete the performance assessment tool electronically, annually, or coinciding with each student placement. Center coordinators of clinical education were able to provide more feedback than clinical instructors or both.

<div align="center">Phase 2</div>

The tool was comprehensive and could be completed in a reasonable time.

**Craig & Kaiser (2003)**

<div align="center">Phase 1</div>

A large sample and test produced no effects on parameter estimates. All of the chi-squares tests were significant, except task orientation and relations-orientation items. There was no significant item level or scale-level differential functioning.

<div align="center">Phase 2</div>

A small sample and test produced no effects on parameter estimates. The proportion of chi-square statistics was lower than phase 1 statistics.

**Darr & Catano (2008)**

Planning and organizing skills were rated higher by supervisors, and thinking skills were rated higher by supervisors and peers. The correlations peer-supervisor and peer-subordinate were significant being the last the highest and the first, correlated with the interview. There was a common factor for 360° and a behavioral interview.

**deLeon & Ewen (1997)**

Employees approved the new system. Only training was distributed fairly. Ratees should select their raters. The system gave useful information, increased trust, communication with supervisors, cooperation, and motivation. Supervisors valued knowing how workers were perceived. The new system allowed a better customer service. Workers understood better the system and its advantages. Managers were rated higher by their bosses and lower by subordinates and peers. Non-supervisory peers were rated lower by colleagues.

**Dupee et al. (2011)**

No statistical difference between pre and post-intervention was found but, scores regarding satisfaction with the appraisal process, the absence of bias, useful feedback, concrete examples and clear goals, dropped over time. The system was well received by nurses.

**Facteau & Craig (2001)**

The baseline and the invariant models were indistinguishable in how they fit the data. Three items and one scale met the necessary criteria, and all of those were confined to the Motivating Others scale. Latent performance variables varied across all groups.

**Fang et al. (2013)**

There was a difference between genders and grades in the two groups. Who enrolled through National College University Entrance Examination (NUCEE) pointed teamwork as the most crucial aspect of professionalism, who enrolled through school recommendations (SR) pointed out teamwork. Students enrolled through SR, rated their improvement and continual improvement of medical technology knowledge higher, as well as peers, for the last one. Students enrolled through NCUEE, were rated higher for continual improvement of medical technology knowledge by visiting staff (VS) and chief resident. VS, chief resident and nursing staff rated higher work attitude.

**Gabriel et al. (2015)**

<div align="center">Study 1</div>

Employee gender correlated with service performance, occupational service familiarity, and service familiarity. Employee positive emotional displays and service familiarity are positively related to coworker ratings. Positive emotional displays were positively associated with judgments in low-familiarity contexts.

<div align="center">Study 2</div>

Employee positive emotional displays and service familiarity were predictors of transaction satisfaction and employee friendliness.

**Grandey et al. (2018)**

<div align="center">Pilot Study</div>

Occupation predicted occupation warmth. Expectations for warmth were the highest. Employee race affected perceived warmth and perceived fit, such that Whites were perceived as a warmer and better fit. Men showed more racial bias than women.

| | Study 1 |
|---|---|
| | Employee race affected warmth judgments with White employees being rated higher. Emotional labor improved only evaluations of the Black's perceived warmth and performance. The respondent race was a moderator only for emotional labor. |
| | Study 2 |
| | Interpersonal warmth judgments explained racial disparity in overall job performance such that Blacks were rated lower. Expressive regulation mitigated the performance rating race disparity by improving warmth judgments specifically for Black employees. |
| Gutermann et al. (2017) | Women and older followers reported higher work engagement. Women leaders reported higher work engagement. Leader-member exchange was positively related to leaders' and followers´ work engagement. Followers' work engagement was positively related to their performance and negatively related to their turnover intentions. |
| Hoffman & Woehr (2009) | Peer and subordinate rating sources correlated with externally measured constructs. Assessment centers (AC) dimensions related to source factors. AC interpersonal skills factor was related to peer, and AC leadership skills factor was related to subordinate source. AC technical/administrative performance factor was related to all sources. |
| Jong et al. (2019) | Faculty members provided the highest Queen's Simulation Assessment Tool (QSAT) scores. The agreement between faculty members and nurse evaluations was lower comparing to peer or emergency medical services (EMS) provider evaluations. Self-evaluations were the lowest in all categories. Physicians and EMS providers agreed the most regarding QSAT scores. |
| Lev & Koslowsky (2012) | Conscientiousness predicted task and contextual job performance. On-the-job embeddedness (ONJE) was positively related to both performance dimensions. All ONJE components related to contextual performance but only links and fit were significant. The links between conscientiousness, ONJE and contextual performance were significant. |
| Li & Wong (2008) | The setting size and the establishment of clear roles and responsibilities were crucial to implement the evaluation and improvement process. Time, space, resources, workload, and teachers' motivation and preparation were pointed out by teachers as obstacles. The motivated and prepared staff ensured the procedure validity. The parent's resistance was pointed out by the research team as obstacles. Teachers improved in most settings. |
| Lyde et al. (2016) | There was a development benefit from completing the portfolio, decreasing it over subsequent submissions. The tool structure and multisource nature were useful to motivate teachers to reflect on their teaching. Faculty members were discontent with the portfolio timing but approved the tool. The tool was not good enough and suggestions were made. |
| Narayanan et al. (2018) | Interpersonal communication and impediments to care access affected patient evaluations. Patient and colleague questionnaires had good external and construct validity. Communication with patients was positively associated with performance. Self-evaluation was lower than others' ratings. Colleagues-self-scores and patients-self-scores converged. Using a patient questionnaire, group 1 and 4 scored themselves higher. Using colleague questionnaire, group 4 scored themselves higher. All questionnaires measured different aspects. There was a high agreement in item interpretation and internal consistency of patients' and colleagues' questionnaires. |
| Noonan et al. (2011) | The initial process of developing the database was labour intensive. The implementation became easier as successive sets of appraisals were performed. The lowest response rates were for items dealing with technical proficiency. The anaesthetic trainees had the highest non-response rates, and the incharge anaesthetists had the lowest. Assessments were considered valid, reliable and internally consistent. |
| Nuryanti et al. (2017) | Phase 1 |
| | The instrument was valid and reliable for all questions. |
| | Phase 2 |
| | Instrument revealed to have good quality being highly functional, reliable, efficient, usable and portable. |
| Overeem et al. (2012a) | Self-ratings were lower than colleagues´ ratings. Performance improved due to multisource feedback assessments. Reported change was positively associated with |

| | |
|---|---|
| | perceived quality of mentoring and feasibility of the webbased service. Colleagues and self-ratings were negatively correlated with reported change. |
| Overeem et al. (2012b) | Peer, co-worker and patient instruments had high internal consistency. Peers gave to their peers higher ratings. Experienced physicians were scored lower by peers and co-workers. Male co-workers scored lower. Patients' age was correlated with the ratings given. Five peers, five co-workers, and 11 patients were required, respectively, to reach reliable results. Performance variance could be explained by raters´ gender and length of the relationship with the ratee. |
| Sargeant et al. (2011) | Family (FP) and specialist physicians (SP) provided examples of high and low scoring behaviors. Direct contact or close working relations were necessary to assess accurately. More objective means were required. To assess Clinical Competence, SP used referral letters and SP and FP evaluated the pertinence of diagnostic tests, medications prescribed and patient follow-ups. To assess Psychosocial Management of Patients, were used colleagues and patients' feedback, and it was difficult for FP. |
| Schuh et al. (2018) | **Study 1** |
| | Innovative work behavior related positively to supervisor ratings and negatively with leader-member exchange. Supervisors rated higher employees engaged in innovative work behavior. |
| | **Study 2** |
| | Employee innovative work behavior related positively to supervisor ratings. Leader-member exchange (LMX) influenced innovative work behavior and job performance ratings, being these stronger when employees had a high LMX relationship. |
| Seaburg et al. (2016) | Multi-source evaluations were positively linked with the number of scientific publications. Half of the subjects wanted to complete the tool annually, and the other half wanted to do it after each rotation. |
| Selvarajan & Cloninger (2012) | Multisource appraisals were perceived as been the most distributive, procedural and interactional fair. Feedback richness was related to procedural and interactive justice. Appraisals used for administrative and developmental purposes, and with more feedback richness, was perceived as the most accurate and fair. Perceptions of procedural and interactive justice were related to appraisal satisfaction, and motivation to improve. |
| Sung & Choi (2018) | The manufacturing industry predicted the firm innovative performance, and related to market demand, technological change, and firm investment in training and development (T&D). Employee competence and commitment positively predicted the firm investment effects and employee T&D positive perceptions on firm innovative performance when they voluntarily participate in T&D, and when firms did not implement T&D. |
| Treadway et al. (2013) | **Study 1** |
| | Age, gender, organizational tenure, performance, political skills, and the interaction between them predicted interpersonal power. Employees with higher levels of political skill had more interpersonal power. |
| | **Study 2** |
| | Race, position, performance, political skill and the interaction between them predicted power. Higher performers exerted more influence when their political skill was high. |
| van der Meulen et al. (2017) | A minimum of three peers, two to three residents, and three to four coworkers guaranteed reliably assess. The instrument provided reliable and valid information about performance. The association between narrative and numerical feedback was significant. |
| van Hooft et al. (2006) | The agreement was higher for supervisor-peer and self-peer. The correlation between peer-rating, organizing, planning and the In-Basket score was significant. In the intelligence test, the total rating on judgment correlated with general intelligence. In the personality test, correlations for similar dimensions were the highest, and self-ratings correlated the most. Supervisor-ratings were the lowest or equal to other's ratings. |
| van Veelen & Ufkes (2019) | Team members assessed their subjective performance as 'sufficient' and their objective performance as 'good'. Teams with weak group identity scored lower on team task learning, team efficacy, and subjective performance. Teams with a strong group identity, objective potential diversity, and team efficacy positively related to objective and subjective performance. For teams with a strong group identity, objective diversity had an effect on objective and subjective performance via team learning and efficacy. |

| Violato et al. (2008) | Instruments were internally consistent over time. Groups of eight assessors and 25 patients were reliable. There were changes between the two medical colleagues' ratings, due to years of practice, co-workers' ratings, due to gender and urban practice location, and patients' ratings, due to communication, location of practice and years of practice. Professionalism was a key variable in all regressions. |
|---|---|
| Weigl et al. (2019) | Perceived teamwork had positive evaluations. Cooperation and resource management received the lowest ratings. Intra team agreement was better in sub-teams. Self and observer disagreed in ratings, being the last lower. Perceived stress was higher in sub-teams. Intra team disagreement on perceived stress was related to team performance. |
| Whitaker (2012) | Supervisor and subordinate's ratings differed. Ratees' age was positively correlated with others' ratings of contextual performance. Female raters assessed contextual performance higher. Ratees' tenure was positively associated with self-ratings of task performance and raters' ratings of the task and contextual performance. |
| Williams et al. (2017) | The items in the Dreyfus and Dreyfus measure focused on the same construct and had a common factor with 360° instrument. Dreyfus and Dreyfus could capture the same information contained in the 360° instrument; however, there were competencies only capture through the last one. Dreyfus and Dreyfus correlated with other's assessments. |
| Wyland et al. (2016) | School-work facilitation was positively related to school demands, and interpersonal interest/support received at school and school relevance. The school-work conflict was positively associated with school demands and negatively to task performance. Work-school facilitation was positively related to job demands, job control and interpersonal interest/support received at work, and predicted job satisfaction, task performance and interpersonal facilitation. The work-school conflict was negatively related to job control, and interpersonal interest/support received at work, and positively with job demands. School control was negatively related to school-work control. |

Note: The core concept is written in the same way as the authors wrote it in their studies.

Table 3 exhibits the diversity but also the idiosyncrasy of each relationship presented. Because of our study purpose, we chose only to report significant findings related to MSF.

Seventeen investigations' results were related to multisource rating sources (e.g., deLeon & Ewen, 1997) namely, the existence of internal sources such as self, hierarchical superior, subordinate and peer, and external sources such as patients, parents and costumers (e.g., Darr & Catano, 2008). Each rating source evaluates different performance aspects and domains of practice (e.g., Overeem et al., 2012b). It was found that ratees assessed themselves lower (e.g., Overeem et al., 2012a) having the non-supervisory or supervisory role influence, being lower for non-supervisory roles (e.g., deLeon & Ewen, 1997). Results show a disagreement between each rating sources' pair (e.g., Jong et al., 2019), and that variables such as length of relationship between ratee and rater, age, innovative work behavior, gender, and work experience have an impact in how low or high ratings are given and received (e.g., Whitaker, 2012). Moreover, setting size, proper training, clear roles and responsibilities were crucial keys to the evaluation and improvement processes' implementation (e.g., Li & Wong, 2008). Ratees should help to select their raters (e.g., deLeon & Ewen, 1997), raters resort to several information sources to rate (e.g., Sargeant et al., 2011), and supervisor's ratings can predict

performance on an interview (e.g., Darr & Catano, 2008). Although the initial implementation process is complicated, it became easier with the successive sets of appraisals (e.g., Noonan et al., 2011). MSF should be administrated annually or electronically (e.g., Buccieiri et al., 2008).

Regarding multisource tool's characteristics and psychometric properties, 16 investigations revealed that instruments were internally consistent over time, however, there were changes between assessments due to, for example, years of practice, gender and practice location, being professionalism a key factor (e.g., Violato et al., 2008). Sample and test sizes had no influence, and there was a common factor between behavioral interviews and MSF (e.g., Craig & Kaiser, 2003). The agreement level between supervisor-peer and self-peer was higher, and between self-subordinate was lower (e.g., van Hooft et al., 2006). There was a disagreement among self-rating and external source's rating, being the last lower (e.g., Weigl et al., 2019). To achieve reliable performance appraisals a minimum of two to 15 co-workers, three to nine peers, six to seven hierarchical superiors, and eleven to 25 external sources are needed (e.g., Overeem et al., 2012b). MSF tools were considered reliable, valid, efficient, usable, portable, comprehensive, and highly functional (e.g., Nuryanti et al., 2017).

Ten studies showed results regarding other variables, which are not performance, its relationships and outcomes. Sociodemographic variables such as age, gender and race, positively predicted work engagement, perceived warmth and fit, emotional labor, interpersonal warmth, expression regulation, and interpersonal power (e.g., Grandey et al., 2018). Work engagement was positively predicted by leader-member exchange (e.g., Gutermann et al., 2017). Work-school facilitation was positively related to interpersonal interest or support received at work, and job demands and control (e.g., Wyland et al., 2016). Emotional displays and service familiarity predicted employee friendliness (e.g., Gabriel et al., 2015). Interpersonal power was positively related to organizational tenure, political skills, performance and position (e.g., Treadway et al., 2013).

Ten investigations revealed that performance and its subtypes, can be predicted, mediated or moderated by several variables (e.g., Lev & Koslowsky, 2012). The firm innovative performance was predicted and mediated by the industry type, external demands, training and development (T&D), employee competence and commitment, and

employees' positive perceptions (e.g., Sung & Choi, 2018). Subjective, objective, contextual, service and in-role performance was positively related with group identity, objective potential diversity, team efficacy and learning, communication with patients, raters' gender and length, emotional labor, perceived warmth, employee gender, conscientiousness, service familiarity, within-group value, organizational citizenship behavior, work-school facilitation, leader-member exchange, emotional and social competences and on-the-job embeddedness (e.g., Araújo & Taylor, 2012).

Six studies reported an overall approval of MSF tools implemented or used. It was described the development benefits, recognized the usefulness of information, perceived the increase of trust and communication with supervisors and cooperation, motivation and MSF structure and process comprehension (e.g., deLeon & Ewen, 1997). MSF appraisal was perceived as more fair and rich than single-source appraisals (e.g., Selvarajan & Cloninger, 2012). Portfolios and mentoring were useful practices (e.g., Lyde et al., 2016). The need for MSF improvement was emphasized (e.g., Lyde et al., 2016).

Regarding the studies that explore MSF in performance appraisal, Table 4 shows the limitations, suggestions for future studies and practical implications made by the authors.

Table 4

*Limitations, suggestions for future studies and practical implications*

| Authors and year of publication | Limitations | Suggestions for future studies | Practical implications |
|---|---|---|---|
| Anand et al. (2018) | Cross-sectional data. Lack of sample diversity. | Other research designs. Other i-deals forms and factors, if relationships patterns differ across i-deals types, leader-member exchange links. | Managers are a link in i-deals manifestation. Leaders should be trained and active in i-deals executing for employees with low group perception. I-deals should be carefully designed and implemented. Collaboration is vital to the organization's goals. |
| Araújo & Taylor (2012) | Same raters for all ratees. Raters were also ratees. Reticence was not controlled. Performance appraisal criterion differed from originals. Feedback was not given. | Larger samples. To examine the relationship of emotional and social competences with job performance through a variety of measures and professional contexts. | Self-assessments barely predict job performance. Emotional and social competences, when evaluated by others, have a partial influence on job performance being the last explained mostly by four competencies. |

| | | | |
|---|---|---|---|
| Bindels et al. (2019) | The tool is context-specific. No follow-up research. | To examine Group Monitor (GM) scores in relation to other measures of group performance and measures of other domains. To investigate GM effects on practice. | Group Monitor is a uniform multisource tool. Raters should be treated as separate information sources. Rater's perceptions are influenced by culture and the physician's specialty. It is crucial to have enough raters. Validity should be regularly updated. |
| Buccieri et al. (2008) | Low response rate and inability to calculate it correctly. The tool may not be comprehensive. | To design and validate student, program director, and program faculty versions of the tool. | The information retrieved via this tool may be used in promotion portfolios and to implement changes in performance. |
| Craig & Kaiser (2003) | Scales construction. Sample size and characteristics. No explanation of why parameter estimates were unaffected by independence assumption violation. | Findings replication with multisource ratings in other contexts. Research design. To assess the influence of characteristics of raters on leadership performance ratings. | The inclusion of multiple raters per ratee does not affect the estimation of item response theory item parameters. Incurring the associated sample size reduction may yield no benefits regarding parameter estimation. |
| Darr & Catano (2008) | Small sample size. Use of a single criterion to validate ratings. | Larger samples and a variety of criteria. | When multisource feedback has development purposes, managers value more their supervisor' and peer's evaluations. |
| deLeon & Ewen (1997) | Only one pre and post-test. It was not possible to rule out certain hypotheses. | Research design. To investigate the effect of rival hypotheses and the conditions crucial to confirm 360º as the best system. | Multisource assessment is crucial for employees' perceptions of performance appraisal fairness. Employees should develop performance appraisal criteria. Evaluating a team strengthens the appraisal process. |
| Dupee et al. (2011) | No control group. Problems with the procedure execution. | Research design. Larger sample. Control group. | Performance appraisal satisfaction must be taken into account before and during its implementation. |
| Facteau & Craig (2001) | Managers chose their raters. Focus on a single multisource feedback system. Sample. | Examine the results generalizability to other types of performance rating instruments and contexts. | Rating discrepancies cannot be attributed to differences between rater groups and may occur due to managers´ behavior. |
| Fang et al. (2013) | Not mentioned. | To explore if different course characteristics and ethics make a difference in performance in professionalism. | It helps to improve medical education, determine how many students enroll through each method, develop course feedback and enhance teachers, tutors, lecturers and counselors' abilities. |
| Gabriel et al. (2015) | Study 1 | | |
| | Focused only on positive emotions. The omission of some variables. Customer–employee familiarity measures. Coworkers and customers ratings were not included. | Other research design. Consider culture. If familiarity predicts positive displays when employees are not monitored. Employees' characteristics impact on customers' perceptions. | Consistently displaying positive emotions toward customers with service familiarity is not vital and may have negative outcomes. Selection, training and employees' emotional displays monitoring and motivation should exist in some contexts. |
| | Study 2 | | |

| | Use of one item to measure friendliness. | Equal to study 1. | Equal to study 1. |
|---|---|---|---|
| Grandey et al. (2018) | Pilot study | | |
| | Based on first impressions. Occupation-racial stereotype incongruence was not tested as a mechanism. Racial bias against Black employees. | Demonstrate that race goes beyond personality predictors. Use contexts where racial stereotypes are congruent or not with occupational requirements. Determine how Blacks' emotional labor can be enhanced. | Expressive and mood regulation are effective strategies to overcome stereotypes and the implications for service performance judgments, but can not be always employed. Managing the workplace climate helps to attenuate the customer's racial biases impact. |
| | Study 1 | | |
| | Sample, stimuli and context characteristics. Employee effort was only observed. Based on first impressions. | Equal to the pilot study. | Equal to the pilot study. |
| | Study 2 | | |
| | Cross-sectional data. Other explanatory factors can exist. Sample characteristics. | Equal to the pilot study. | Equal to the pilot study. |
| Gutermann et al. (2017) | Sample characteristics. Self-reported performance. Leader-member exchange assessed as a mediator. Team dynamics related to leader engagement and its effects were not assessed. | Replication in different contexts, with other research designs and considering other variables. Occupation type or industry and the link between employees´ turnover intentions and performance. | Organizations should promote work engagement at the managerial level to foster an organizational culture of engagement. In times of cost-efficient training developments, organizations could use the multiplication effects of manager trainings. |
| Hoffman & Woehr (2009) | Instrument based on one facet. Impact of individual raters. Traditional performance appraisal approach. | Replication with other instruments. To investigate the psychometric properties and theoretical underpinnings. | This study is the first step to disentangling the meaning of multisource performance rating sources and dimension effects. |
| Jong et al. (2019) | Research design. Raters attribution to ratees. Convenience sample. Participants were known. No 1st postgraduate year´s students participated. Lack of training to use the tool. | Not mentioned. | The Queen's Simulation Assessment Tool is a viable tool, but, self-rating should not be included. Emergency medical services providers should also be a source. To include nurses as sources training, experience and gender have to be considered. |
| Lev & Koslowsky (2012) | Specific sample. Use of a common tool for all outcomes measures. | Other models. Non-Big Five individual difference variables. Psychometric properties. Other research design. | On-the-job embeddedness (ONJE) can enhance organizational productivity. Workshops or open discussions enhance ONJE levels. 360º, when used to other purposes, is a more comprehensive and objective alternative. |

| | | | |
|---|---|---|---|
| Li & Wong (2008) | Settings representativeness. Few quantitative evidences. | Larger and more representative evidences. | There is a need for more governmental resources in teacher education. The criteria for preschool teachers should be higher. Teacher professionalism is vital for the quality of early childhood education. |
| Lyde et al. (2016) | Not mentioned. | Establish a schedule of due dates, and ongoing reflection, a development cycle, criteria to performance score levels, a reflection and justification about professional progress. | Timing of reflections and accountability from year to year should be addressed. Administration and senior faculty should assume mentoring responsibilities to promote this on faculty culture. |
| Narayana n et al. (2018) | Variable number of doctors. Large sample. Doctors selected their raters. | Not mentioned. | Multisource feedback (MSF) enhances professional development. Mechanisms to help doctors be more comfortable about self-rating are needed. The questionnaires can be used for MSF. |
| Noonan et al. (2011) | Raters characteristics. High non-response rates. | To implement and develop this system. | The assessment system provided a focus on anaesthetist's behaviors and interactions. It allows for objective discussion with staff. |
| Nuryanti et al. (2017) | Not mentioned. | To develop the web with quantitative indicators for nurse assessment. | The hospital should invest in IT, involve raters of other professional areas, and use multisource feedback. Feedback should include verbal, for example, communication, listening, problem-solving, and agreement. A formal report is needed. |
| Overeem et al. (2012a) | Other predictors, specialists' age, hospital and groups, comparison of responders with non-responders and external assessments were not analyzed. The response rate. | Investigate other self-reported change predictors, real change in practice, the effect of several combinations of gender, and specialty. If specialists would be satisfied with narrative comments. | Coworkers and patients' ratings are not decisive in specialists' change. Mentoring should be part of multi-source feedback programs. Mentors should be equipped with strategies such as posing reflective questions, and goal setting collating. |
| Overeem et al. (2012b) | Sample selection. Not all peer responses were used. Research design. Data unavailability. | Examine the meaning of missing values, assess the impact of several hospitals and specialty groups on reported change. Establish the validity of the tools. | Multi-source feedback is valid and reliable. A minimum of five peers, five co-workers, and 11 evaluations of patients are needed. Self-reports can be used for reflection and development. |
| Sargeant et al. (2011) | Research design. Performance standards and findings may not be generalizable. | If performance standards can be generalized. The usefulness of the rubric as an aid to informed assessment and feedback. | A rubric was developed. Concrete examples can stimulate critical thinking about one's practice and improvements. |

Study 1

| | | | |
|---|---|---|---|
| Schuh et al. (2018) | Focus on only two central assumptions. Measures of employee innovative work behavior measures. Sample characteristics. | The links between workers' behaviors, leader-member exchange (LMX) and performance ratings. Other variables and measures. Boundary conditions for LMX influence on innovative work behaviors appraisal. | To increase objectivity, human resources professionals can provide rating scales and frequent rater training. Organizations might also invest in efforts that objectively assess innovative performance whenever there are multiple leaders in a single environment. |
| | Study 2 | | |
| | Equal to study 1. | Equal to study 1. | Equal to study 1. |
| Seaburg et al. (2016) | Research design. The system is not generalizable to all contexts. | Evaluate if there is a link between publication during residency and future accomplishments. | Providing education in research during residency training can improve resident's core competencies. |
| Selvarajan & Cloninger (2012) | Research design. Self-reported questionnaires from the same source. Multiple source information was not possible. Cultural variables were not explicitly included. | The influence of cultural factors on fairness and appraisal outcomes and if this link is due to employee concern with performance appraisal process or psychological contract. To examine the effectiveness of paternalistic leadership in Mexico. | American bureaucracies do not work in Mexico. A paternalistic leadership style can be effective. Organizations should adopt multisource appraisals. The feedback provided needs to be rich. Raters should be trained and encouraged to provide frequent and specific feedback and keep diaries of employee performances. |
| Sung & Choi (2018) | Measures used. The intermediate psychological processes involved were not tested. Sample characteristics. Research context. | Objective indicators of firm innovation. The multilevel dynamics involving training and development and its dimensions or purposes. Investigate industry-specific dynamics and internal firm environments. | Bottom-up employee-centered approaches to training and development (T&D) may be highly effective. Managers should promote high-quality, practical programs and positive attributions of T&D via human resources practices and participation in decision making. |
| Treadway et al. (2013) | Study 1 | | |
| | The interpersonal power levels influence on performance, and the links between organizational position and interpersonal power were not explored. Definition of power. Unavailability of data. | Examine the link between performance, political skill, other workplace behaviors, and work social networks. Examine the political skill construct. The individuals' abilities to manage resources and others' impressions. More data collection. | Fostering the development of political skills in managers may result in positive outcomes. The lack of recognition may result in perceived inequity or injustice, which may lead to lower levels of job satisfaction and organizational commitment. |
| | Study 2 | | |
| | Use of one item to measure performance. | Equal to study 1. | Equal to study 1. |
| van der Meulen et al. (2017) | Ratees selected the raters. Dichotomization of narrative feedback. | If the instrument is approved. How can contribute to professional performance. Study random sampling effects and its validity in other contexts. | To maintain physician commitment to the process, they should not be overloaded with extensive tools. |

| | | | |
|---|---|---|---|
| van Hooft et al. (2006) | Number of raters used. No subordinate ratings. Psychometric properties. | Other research design. To assess construct validity when using external measures that evaluate the same dimensions. | Organizations should be careful when using 360° performance appraisals for more than developmental purposes. |
| van Veelen & Ufkes (2019) | Sample dimension. Power issues. Focus on team diversity-performance-intragroup dynamics interactions. Project teams were short-lived. | Larger samples. If leadership or coaching skills support team building in diverse teams. Focus on intra and inter-team learning. Research design. | Supervisors should create an environment where members can build their own team identity, their interdependency and facilitate joint goal setting. To prevent misperceptions, intermediate feedback should be given. |
| Violato et al. (2008) | Time gap between appraisals. Few feedback formal moments. | Which changes occur over shorter and longer periods. | In this setting, formal mentoring or coaching system in union with reports can alter self-rating. |
| Weigl et al. (2019) | Research design. Variables and not fully captured. Sample. Tool structure. Observer bias. Focus on professionals' teamwork skills. | Consider measures of provider attendance for handover duration. Consider what kind of assessment practice is chosen. Use videotapes for multi-professional reviews. | Checklists are an effective tool for information transfer. Work and process redesign is necessary good handover processes. Disruption of the environment helps the caregiver to plan subsequent steps of care. |
| Whitaker (2012) | Sample characteristics. Supervisors' characteristics. | Distinguish task and contextual performance. Replicate the study with other job performance measures and an older sample. How contextual performance can be better included in multi-source feedback. What means performance for jobs gender-dominated. | Demographic variables of ratees and raters influence job performance ratings, having more impact on contextual performance. Rater training should take demographic variables influence into account. |
| Williams et al. (2017) | Bias. Data came from a single center. Inability to generalize results. No demographic information. | To assess the relationship between 360°, a general measure of mastery, and their purposes. | The use of the two measures leads to a more efficient evaluation process and to richer data, which gives insight into more effective methods and content. |
| Wyland et al. (2016) | Sample origins and size. The research design was not longitudinal. | Other research design. A diverse and different sample. Antecedents and outcomes of work-school facilitation and other bi-directional forms of conflict and facilitation. The role that work-school and school-work conflict play on work outcomes. | Work-school facilitation has a crucial role for both working students and the organization. Employers must provide working students autonomy, challenges, and responsibilities in their jobs. Employers must encourage organizational interest in their colleagues´ school endeavors and create opportunities for work–school synergies. |

Note: The core concept is written in the same way as the authors wrote it in their studies.

The limitations pointed are diverse. Thirty-six studies (e.g., Wyland et al., 2016) highlighted limitations concerning the study design, for example, the procedure execution, research design type, the means of raters' selection, bias or response rate not

controlled. Sixteen studies (e.g., Darr & Catano, 2008) stressed limitations related to the samples used such as sample characteristics, sample origin, sample size or the way of sample retrieving. Twelve studies (e.g., Craig & Kaiser, 2003) mentioned variables/relationships not contemplated in the studies/not explained, or the exclusive focus on certain variables. Thirteen studies (e.g., van Hooft et al., 2006) stressed limitations concerning tools or measures such as its psychometric weaknesses, the use of a single measure or criterion, power issues, or the inability to generalize it to all contexts. Five studies (e.g., Anand et al., 2018) highlighted limitations data related, namely, the unavailability of data or the use of cross-sectional data. Treadway et al. (2013) mentioned a limitation regarding power definition. Three studies (e.g., Lyde et al., 2016) did not mention limitations.

Concerning suggestions for future studies, it was proposed to investigate or assess other variables, predictors, psychometric properties, the importance of narrative comments, the findings generalizability and the impact, influence or effect of one variable into another, of a variable in a relationship or a relationship into a variables (31 of the total, e.g., Seaburg et al., 2016). Seventeen studies (e.g., van Hooft et al., 2006) suggested a change, alteration or re-conceptualization of the research through, for example, different research design, the use of different samples, or by considering and introduce other variables, indicators or relationships. Three studies pointed to further development, implementation and validation of the tools created to that context or others (e.g. Buccieri et al., 2008). Four studies (e.g., Hoffman & Woehr, 2009) suggested replicating its studies within other conditions. DeLeon and Ewen (1997) suggested a repetition of their test administrations but with an improved procedure. Whitaker (2012) proposed to distinguish between task and contextual performance and what means performance for jobs gender-dominated. Two authors (e.g., Jong et al., 2019) did not mention any suggestions for future studies.

Concerning the practical implications, 25 studies (e.g., Narayanan et al., 2018) emphasized the role of MSF, feedback characteristics, the need to include checklists and forms, its effects, the characteristics that boost its advantages and its importance in and within organizations. It was advised its incorporation in organizations, a regular update to maintain validity, stressed out its reliability and validity in medical contexts, highlighted the importance of self-assessment, that prevent and detect teams

misperceptions, that allows a focus on rate behavior, its developmental purpose and that, when applied to other purposes, is a comprehensive and objective tool. It was recommended the use of MSF and a measure based on the Dreyfus and Dreyfus framework to turn evaluation more efficient and obtain richer data (Williams et al., 2017). Moreover, it was suggested the number of raters needed per category to assure process validity and reliability, that ratees pay more attention to their supervisors and peers' evaluation and that raters perceptions are influenced by culture and rate characteristics. Seventeen studies (e.g., van Veelen & Ufkes, 2019) made inferences related to the crucial role of managers, team leaders, and organizations on employees' job performance improvement and, especially, on the implementation or improvement of MSF and performance appraisal process. They pointed out the need of training, suggested the inclusion of mentoring as part of performance appraisal system, pointed out the need for the promotion of programs that allow the development of employees' competencies, the need of having in account individual variables, and to be careful when implementing MSF. Four studies (e.g., Anand et al., 2018) discussed implications on performance, on employers, and on employees of variables such as emotional and social competences, service familiarity, i-deals and work-school facilitation. Two studies (e.g., Facteau & Craig, 2001) highlighted implications regarding MSF tool's psychometric properties.

## IV. Discussion

The introduction and practice of MSF in performance appraisal only has 30 years (Atwater et al., 2002; Görün et al., 2018). The small number of empirical articles eligible for this literature review confirms that this subject is still in the early stages of development in the psychology research area. The relevance that this concept has gained in the current organizational, political and strategic conjuncture (Ferraro et al., 2016; Görün et al., 2018; London & Smither, 1995) emphasizes the need of a more in-depth knowledge about it in the form of more empirical research on this subject (Table 5).

The MSF conceptualization, congruent with the democratic ideal, is based on the idea that the existence and joining of different viewpoints on the same object enrich the whole (Edwards, 1996; Loredana & Mirabela, 2015; Sen, 2000). Similarly, the multiplicity of study purposes expresses the growing attempt to understand the various MSF aspects. Several viewpoints came together in this research. Despite the fact that some

improvements were suggested on the way MSF should be put into practice, none of the studies concluded that it was useless or had detrimental effects on outputs or workers (Table 5). However, there are potential negative effects and disadvantages associated with MSF use (Hosain, 2016). That is confirmed by mixed results obtained in the studies that aimed to establish the positive impact of MSF (Bailey & Fletcher, 2002). Thus, it should be determined, confirmed and understood the potential negative effects associated with MSF use (Table 5).

Table 5

*Gaps in the existent MSF-performance appraisal literature*

| Gaps | Suggestions | Assumptions |
| --- | --- | --- |
| Low number of empirical studies about MSF in performance appraisal. | More empirical studies about MSF in performance appraisal. | MSF in performance appraisal has gained relevance in the current organizational, political and strategic situation being vital to deepen the knowledge about it. |
| Lack of studies about possible negative effects of the use of MSF in performance appraisal. | More studies about the possible negative effects or the "dark side" of MSF in performance appraisal. | There are possible negative effects of the use of MSF in performance appraisal. |
| Low number of variables related to using MSF and other variables. | Expanding the MSF nomological network by researching its relationships with new variables. | The knowledge on the relationships of MSF with other variables is useful to the understanding of the organizational phenomena and to intervene in organizations. |
| Low number of empirical studies focusing on MSF psychometric properties. | More empirical studies focused on MSF' psychometric properties. | The usefulness of MSF depends on its instruments' psychometric properties. |
| Lack of studies in African countries. | Developing future studies in African countries. | Cultural differences affect the use of MSF in performance appraisal. |
| | | It may exist cultural specificities that prevent the direct generalization of results obtained in one cultural context to another. |
| Few studies in non-western cultures. | More studies about the cultural sensibility of MSF in performance appraisal. | Cultural specificities may influence the effectiveness of MSF in performance appraisal. |
| | | MSF in performance appraisal may be more adapted to one culture than to another. |
| Almost exclusive use of questionnaires in MSF operationalization. | Develop more studies that operationalize MSF through different tools. | Data collected through tools other than questionnaires will enrich the usefulness of MSF and will deepen understanding of organizational phenomena. |

| Low number of longitudinal research designs. | Developing future studies featuring longitudinal research designs. | Longitudinal research designs may print more accuracy on the new knowledge by assessing temporal links about MSF in performance appraisal. |
|---|---|---|
| Lack of studies about MSF' sensitivity to organizational political factors and their impact. | Develop studies which investigate MSF' sensitivity to organizational political factors and their impact. | Raters may manipulate the performance appraisals given in order to respond to their own personal hidden political agenda. |
| Lack of studies about MSF negative impact at the individual and organizational level. | Develop studies which investigate MSF negative impact at the individual and organizational level. | MSF can be used can be used as an instrument of moral harassment. |
| Lack of studies relating the use of MSF and organizational strategy, policy, and philosophy. | Undertaking research relating organizational strategy, policy and philosophy to MSF inclusion in performance appraisal. | Relating the several levels of organizational leadership is useful to make organizations more effective and to understanding organizational phenomena. |

Knowledge and understanding of MSF and its relationship to other variables as well as its impact are vital to understand the organizational phenomenon, enhance its use and benefits, and to intervene suitably in organizations. However, most authors focused on assessing the links between two or more variables or studying the influence, impact or effect of one variable into another, not being neither MSF. Future studies may expand the MSF nomological network by researching their relationships with new variables (Table 5). The results point to a small number of empirical investigations assessing the psychometric properties of the instruments or tools that allow MSF operationalization (Table 5) however, the usefulness of MSF depends on its instruments' psychometric properties. Therefore, future research should focus on this crucial aspect (Table 5).

The MSF introduction in performance appraisals responds to organizational, political and strategic needs. Organizational needs, since its advantages and contributions to sustainable and successful organizational development, have been recognized (Baruch & Harel, 1993; Görün et al., 2018). Political needs, given its close connection with the DW concept and its importance to the achievement of the goals set for 2030 (Ferraro et al., 2016). Strategic needs, as the inclusion, empowerment and involvement of workers in internal processes such as performance appraisal, is linked to a performance increasing, enabling organizational development (Baruch & Harel, 1993; Görün et al., 2018; Rohlfer, 2018; Sharma & Kirkman, 2015). The significant mismatch found in samples - units vs

thousands of participants -, and the progressive increase of subjects by year of publication, reinforce the gradual attempt by and in organizations to introduce MSF in performance appraisal and the success of it. Still, the existence of several samples with hundreds of subjects reflects an already cemented MSF practice in organizations.

Given the vital contribution of the MSF concept from an organizational, strategic and political viewpoint, its integration and use in performance appraisal system is expected to expand and globalize. Sampling on several continents except for Africa (Table 5) reinforces the strong organizational adherence to MSF and the assumption that it is not a context-specific phenomenon given the visible geographical and organizational expansion of its use. Moreover, the significant predominance of North-American samples confirms that they are still a major pole of scientific research. However, the emergence of scientific research across Asia points to a progressive shift of the scientific knowledge pole.

Culture is a set of ideas, values, or assumptions shared by people from a social group or category, organization, or community that encode the way they understand, feel, behave, perceive, or expect something (dos Santos, 2019). Although we tend to have a global and universal identity underpinned by common cross-cultural specificities, different cultural backgrounds continue to determine and shape the ways we understand, feel, behave, perceive or expect something (dos Santos, 2019). The results reveal the lack of studies in African countries (Table 5). Thus, further studies in that continent should be developed since cultural specificities and differences may affect how MSF in performance appraisal is perceived and applied and may prevent the generalization of results obtained from one cultural context to another (dos Santos, 2019; Entrekin & Chung, 2001; Table 5). Cultural characteristics differ substantially between eastern and western countries (Entrekin & Chung, 2001). That is, the four value dimensions identified by Hofstede (1984) vary according to the culture we are dealing with, influencing the encoding of the way people understand, feel, behave, perceive, or expect something (dos Santos, 2019; Entrekin & Chung, 2001). Since human resource management practices are the most vulnerable to cultural relativity, practices carried out and accepted in one culture may not be easily transferable to another (Entrekin & Chung, 2001). Since the effectiveness and strength of MSF depends on its correct implementation and practice, as well as attitudes towards it, further studies are needed to understand the applicability,

validity and effectiveness of MSF in performance appraisal in eastern countries and in which culture is more adaptable and effective (Table 5).

The MSF in performance appraisal, the DW concept and approach, and the 2030 Sustainable Development Goals are closely linked since the former allows the full realization of the latter (Ferraro et al., 2016; Pereira et al., 2019). The DW approach, despite its complexity, has some important features that are emphasized by Sen (2000), one of them being its universality and inclusion, that is, its application to all sectors, including informal and irregular sectors and to all workers, including self-employed and domestic workers. In order to meet the 2030 Sustainable Development Goals, MSF should be implemented and applied to all sectors, including informal and irregular sectors, and to all workers, including domestic workers and self-employed workers. Although none of the investigations covers the above-mentioned sectors and worker groups, the results point to the use of MSF in performance appraisal across several organizational sectors and professional groups, confirming that MSF implementation is becoming widespread.

According to Tureta et al. (2006), there are multiple approaches and viewpoints on organizations. The contingency approach stands out, presupposing the absence of a rigid, effective and common organizational structure, depending on its best optimization of contingency factors such as strategy, size, task uncertainty and technology, that is, environmental factors. Each organizational structure becomes a reflection of the environment in which it is inserted (Tureta et al., 2006). The variety of instruments and tools used to operationalize MSF and, in particular, the adaptation of some of them, emphasizes the contingency nature of organizational structures, allowing MSF in performance appraisal to be assumed as having a flexible nature and be adaptable to various organizational contexts. These characteristics combined with the advantages found in the literature justify why its implementation and practice increased (Baruch & Harel, 1993; Flint, 1999; Görün et al., 2018; London & Smither, 1995).

Most authors chose to operationalize MSF exclusively through questionnaires, surveys or forms (Table 5). Since one of the advantages and objectives of the MSF is to provide more fairness, objectivity and validity to performance appraisals performed by and within organizations (Görün et al., 2018), it is only natural to choose an instrument

that is guided and reflects the objectives mentioned (Alferes, 1997). The MSF concept reflects the idea of enriching the whole by bringing together different viewpoints (Edwards, 1996; Loredana & Mirabela, 2015). The MSF operationalization through other instruments or tools that are also advantageous (Alferes, 1997) may allow the collection of different types of data. That data can enrich the usefulness of MSF and allow a deeper understanding of the organizational phenomena. Further studies that operationalize MSF through different tools or instruments are recommended (Table 5).

Of the studies collected and analyzed, few were dedicated to assessing in a long-term the attitudes, perceptions or consequences of MSF in performance appraisal. Most studies, to achieve its study purposes, implemented cross-sectional research designs, reflecting the organizational contingencies as well as the inherent advantages of this research design type (Alferes, 1997). Performance appraisal systems have been studied for several decades, however, there are a lack of investigations focusing on MSF in performance appraisal systems. There is a lack of longitudinal studies that are crucial to validate the temporal constancy of the existing knowledge such as, its advantages, its psychometric properties or the attitudes and perceptions regarding MSF. On the other hand, this research design is important to make a new knowledge more accurate (Table 5). Thus, it is desirable to analyze temporal relationships concerning MSF in performance appraisal through the development of more studies that adopt longitudinal research designs (Table 5).

MSF refers to the performance evaluation of a given employee by more than one rating source, internal or external (Atwater et al., 2002; Loredana & Mirabela, 2015). The evidence of internal and external rating sources and their use to operationalize MSF is in line with the literature. There was a higher incidence of internal rating sources, namely, self, peers, hierarchical superiors and subordinates, and three or more rating sources were combined to evaluate performance. As advocated, the involvement and participation of all employees, regardless of their hierarchical level, in decision-making processes such as performance evaluation, is desirable (Sharma & Kirkman, 2015). The strategic introduction of MSF in performance appraisal, operationalized through many and diverse rating sources, empower employees, contributing to an increase in performance, in their motivation, resulting in organizational development (Atwater et al., 2002; Sharma & Kirkman, 2015). Raters, both internal and external, have access to a partial portion of

ratees' performance since they contact whit it in different ways and contexts (Entrekin & Chung, 2001). The MSF inclusion in performance appraisal and the combination of multiples viewpoints on someone's performance prints more objectivity and fairness to the process allowing organizations to get closer to the decent work concept and to the objectives to be reached globally by 2030 (Atwater et al., 2002; dos Santos, 2019; Ferraro et al., 2016; Görün et al., 2018).

Atwater and Yammarino (1992), and Yammarino and Atwater (1993) proposed that self-ratings cannot be considered isolated when evaluating performance since they tend to underrate themselves. To maintain and/or increase the objectivity, validity and fairness of the process, it is recommended the combination of self-ratings with others' ratings of different hierarchical levels. The results revealed that in general, self-ratings were lower than other's ratings, regardless they are internal or external. Interrater agreement was distinct being higher between supervisor and peer, and between self and peer. Moreover, it was found that employees with non-supervisory roles were rated lower than employees with supervisory positions. This evidence, together with those mentioned above, demonstrates the vitality of the involvement of multiple organizational actors and the joining of different perspectives to impress justice and validity to the process enhancing consequently individual and organizational development.

Organizations are viewed by Ferris et al. (2007) as political arenas. This approach asserts the existence of relationships between various organizational, environmental, work and personal factors, and the subjective individual agendas (Ferris et al., 2007). Performance management is a complex activity that encompasses a wide range of factors that influence the success of performance appraisal, including political factors (Baruch & Harel, 1993; Garg, 2018; Russo et al., 2017). The intentional manipulation of the appraisals given by supervisors to carry on with the execution of their hidden political agendas is an already known organizational phenomenon (Russo et al., 2017). The MSF brings more objectivity to performance appraisals since the unification of several perspectives reduces the subjectivity inherent to the traditional appraisal processes (Baruch & Harel, 1993; Görün et al., 2018). Despite the advantages, MSF can also reinforce and support the feedback provided by the supervisor, leading to the predominance of a single viewpoint, and thus, facilitating the accomplishment of hidden agendas, goals and wills (Atwater et al., 2002; Russo et al., 2017). The ratees' perceptions

of injustice and subjectivity, consequences of subjective and manipulated appraisal, has a negative impact on the individual and organizational level (Baruch & Harel, 1993; Flint, 1999; Görün et al., 2018). There is a lack of investigations about MSF sensitivity to political organizational factors and their impact, being recommended further studies (Table 5).

The unification of the appraisals given by the different raters' groups, provides to the ratee and the organization an average of the ratings given (Baruch & Harel, 1993; Görün et al., 2018; Markham et al., 2014). On one hand, MSF offers complete and holistic data about one's performance, on the other hand, the collective appraisal result obtained through this process dilutes and removes the individual responsibility of the appraisals intentionally manipulated, that is, with malicious intentions (Russo et al., 2017). Markham et al.'s (2014) investigation, whose results pointed to the existence of subgroups with varied and non-congruent intentions, and differences in the feedback provided by the different raters' groups, confirm this statement. It was stressed that not all feedback groups should have their mean performance appraisal scores calculated (Markham et al., 2014). By making the raters not responsible by the ratings given, MSF can be used as an instrument of moral harassment and use as a mean of bullying and social exclusion in the organizational environment. This MSF facet has implications at the individual level, as it affects the employee performance and motivation, and at the organizational level, as it consequently leads to the organizational development stagnation or setback. The use of MSF as a mean of bullying affects its conceptualization as a tool capable of implanting more validity, reliability, fairness, and objectivity in performance appraisal since there is the possibility of being used with a subjective and evil purpose (Baruch & Harel, 1993; Flint, 1999; Görün et al., 2018; Table 5). Understanding the extent of this facet's negative impact and developing strategies to minimize it is crucial. Thus, future studies are recommended (Table 5).

The results found in this literature review are in line with the postulate (e.g., Atwater et al., 2002; Baruch & Harel, 1993; Görün et al., 2018; London & Smither, 1995). MSF and the tools used for its operationalization were considered valid, reliable, objective, fair and consistent over time, admitting small but not significant changes due to external factors. Moreover, several advantages and benefits inherent to the process were confirmed, for example, development benefits, recognition of the usefulness of

information, perceived increase of trust and communication with supervisors and cooperation, motivation and MSF structure and process' comprehension, cementing the knowledge that already exists. Still, the need for improvements was emphasized.

The link between the use of MSF and the current and future political, strategic, organizational and philosophical context has been established throughout this section and confirmed by the results. However, it has not been investigated (Table 5). Understanding how the various levels of leadership in organizations relate to the MSF and how to use MSF strategically is useful to make organizations more effective and to better understand the organizational phenomena. To do this, it is recommended to undertake research relating organizational strategy, policy, and philosophy to MSF inclusion in performance appraisal (Table 5).

### 1. Limitations

Despite our findings, our literature review contains some limitations. We only searched articles written in English, excluding in this way publications in different languages. The exclusion of several idioms can lead to a limitation of information regarding MSF in performance appraisal state of art. There are other publication databases beyond those that were used. Future studies should expand both the publication languages and the publication databases. Moreover, publication bias was not addressed so, other possible studies not published with different results from those here presented were not included.

### 2. Practical Implications

The inclusion of MSF in performance appraisal is desirable from a strategic, political and technical viewpoint, being its implementation and practice crucial within the organization. This fact, accompanied by all the advantages mentioned in the literature and confirmed by the results, enhances sustainable organizational development and allows the employee's skills enhancement. Managers and team leaders, as corroborated by Sharma and Kirkman (2015), are a vital element in this process. They should promote programs such as mentoring and appropriate training, while also maximizing positive attitudes and perceptions of employees with MSF.

The inclusion of MSF in performance appraisal is valid and reliable, requiring the use of multiple and diverse sources as much as possible. However, to ensure reliable assessments, a minimum number of two to 15 co-workers, three to nine peers, six to seven hierarchical superiors, and 11 to 25 external sources are needed. Moreover, setting size, proper training, clear roles and responsibilities are crucial keys for the correct implementation of evaluation and improvement processes.

## V. Conclusion

We tried to characterize MSF in performance appraisal, a concept that, despite having three decades of existence, and its recognized relevance in the current organizational, political, strategic and philosophical conjuncture, is still in the early stages of development in the organizational psychology field. The current context and its demands, as well as the advantages and features inherent to MSF, made its integration in human resources practices crucial. The incremental expansion at the geographical level and in all types of organizations is a reflection of this (Atwater et al., 2002; Ferraro et al., 2016; Sharma & Kirkman, 2015). MSF is linked to individual and organizational development as well as to the accomplishment of the objectives set for 2030, contributing in this way to the dignification of work (dos Santos, 2019; Ferraro et al., 2016).

The results confirm what exists in the literature (Edwards, 1996; Loredana & Mirabela, 2015; Sen, 2000). However, there are some aspects that need to be considered to further investigations. The sensitivity and adaptation of MSF to different cultures and organizational political factors have not been analyzed. Despite the full advantages and benefits, MSF potential negative side was not considered, overshadowing another possible facet of this concept. More empirical studies should be carried out given MSF importance to the organizational psychology field and the practical implications to the human resources management. The consolidation of existing knowledge and the creation and validation of new knowledge leads and allows a deeper understanding of this concept, this organizational phenomenon and the multiple political and strategic links that exist.

## VI. References

Alferes, V. (1997). *Investigação Científica em Psicologia: Teoria & Prática*. Almedina.

Anand, S., Hu, J., Vidyarthi, P., & Liden, R. C. (2018). Leader-member exchange as a linking pin in the idiosyncratic deals - Performance relationship in workgroups. *The Leadership Quarterly, 29*(6), 698–708. https://doi.org/10.1016/j.leaqua.2018.07.005

Araújo, S. & Taylor, S. (2012). The influence of emotional and social competencies on the performance of Peruvian refinery staff. *Cross Cultural Management, 19*(1), 19–29. https://doi.org/10.1108/13527601211195600

Atwater, L., Waldman, D., & Brett, J. (2002). Understanding and Optimizing Multisource Feedback. *Human Resource Management, 41*(2), 193-208. https://doi.org/10.1002/hrm.1003

Atwater, L. & Yammarino, F. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45*, 141-164. https://doi.org/10.1111/j.1744-6570.1992.tb00848.x

Bailey, C. & Fletcher, C. (2002). The impact of multiple source feedback on management development: findings from a longitudinal study. *Journal of Organizational Behavior, 23*, 853-867. https://doi.org/10.1002/job.167

Baruch, Y. & Harel, G. (1993). Multi-Source Performance Appraisal: an Empirical and Methodological Note. *Public Administration Quarterly*, *17*(1), 96–111.

Bindels, E., Boerebach, B., van der Meulen, M., Donkers, J., van den Goor, M., Scherpbier, A., Lombarts, K., & Heeneman, S. (2019). A New Multisource Feedback Tool for Evaluating the Performance of Specialty-Specific Physician Groups: Validity of the Group Monitor Instrument. *Journal of Continuing Education in the*

*Health Professions, 39*(3), 168–177. https://doi.org/10.1097/CEH.0000000000000262

Buccieri, K., Brown, R., Malta, S., & Ward, B. (2008). Evaluating the performance of the academic Coordinator/Director of clinical education in physical therapist education: Developing a tool to solicit input from center coordinators of clinical education and clinical instructors. *Journal of Physical Therapy Education*, *22*(3), 64-73. http://dx.doi.org/10.1097/00001416-200810000-00008

Craig, S. & Kaiser, R. (2003). Applying Item Response Theory to Multisource Performance Ratings: What Are the Consequences of Violating the Independent Observations Assumption? *Organizational Research Methods, 6*(1), 44-60. https://doi.org/10.1177/1094428102239425

Darr, W. & Catano, V. M. (2008). Multisource Assessments of Behavioral Competencies and Selection Interview Performance. *International Journal of Selection & Assessment, 16(1),* 68–72. https://doi.org/10.1111/j.1468-2389.2008.00410.x

deLeon, L. & Ewen, A. (1997). Multi-source performance appraisals. *Review of Public Personnel Administration, 17*(1), 22-36. https://doi.org/10.1177/0734371X9701700103

dos Santos, N. R. (2011). O desempenho e a sua avaliação: dicotomias na prática, desafios para a investigação. In A. D. Gomes (Eds.), *Psicologia das organizações, do trabalho e dos recursos humanos* (577-613). Imprensa da Universidade de Coimbra.

dos Santos, N. R. (2019). Decent work expressing universal values and respecting cultural diversity: propositions for intervention. *Psychologica, 62*(1), 233-250. https://doi.org/ 10.14195/1647-8606_62-1_12

dos Santos, N. R. & Pais, L. (2015). Structured actions of intentional development. In K. Kraiger, J. Passmore, N. R dos Santos, & S. Malvezzi (Eds), *The Wiley-Blackwell Handbook of the Psychology of Training, Development, and Performance Improvement* (398-418). Wiley Blackwell.

Dupee, J., Ernst, N., & Caslin, E. (2011). Recruitment & Retention Report. Does multisource feedback influence performance appraisal satisfaction? *Nursing Management, 42*(3), 12–16. https://doi.org/10.1097/01.NUMA.0000394060.41121.ce

Edwards, M. (1996). Improving performance with 360-degree feedback. *Career Development International, 1*(3), 5-8. https://doi.org/10.1108/13620439610118528

Entrekin, L. & Chung, Y. (2001). Attitudes towards different sources of executive appraisal: a comparison of Hong Kong Chinese and American managers in Hong Kong. *International Journal of Human Resource Management, 12*(6), 965–987. https://doi.org/10.1080/09585190110063183

Facteau, J. & Craig, S. (2001). Are Performance Appraisal Ratings From Different Rating Sources Comparable? *Journal of Applied Psychology, 86*(2), 215–227. https://doi.org/10.1037/0021-9010.86.2.215

Fang, J., Ko, Y., Chien, C., & Yu, K. (2013). Assessing medical students' performance in core competencies using multiple admission programs for colleges and universities: From the perspective of multi-source feedback. *Biomedical Journal, 36*(4), 188-197. https://doi.org/10.4103/2319-4170.113856

Ferraro, T., dos Santos, N. R., Pais, L., & Mónico, L. (2016). Historical landmarks of decent work. *European Journal of Applied Business Management, 2* (1), 77-96.

Ferraro, T., Pais, L., dos Santos, N. R., & Moreira, J. (2018). The Decent Work Questionnaire: Development and validation in two samples of knowledge workers. *International Labour Review, 157*(2), 243-265. https://doi.org/10.1111/ilr.12039

Ferris, G., Treadway, D., Perrewe, P., Brouer, R., Douglas, C., & Lux, S. (2007). Political skill in organizations. *Journal of Management, 33*, 290–320. https://doi.org/10.1177/0149206307300813

Fletcher, C. (2011). Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology*, *74*(4), 473–487. https://doi.org/10.1348/096317901167488

Flint, D. (1999). The role of organizational justice in multi-source performance appraisal: theory-based applications and directions for research. *Human Resource Management Review, 9*(1), 1–20. https://doi.org/10.1016/S1053-4822(99)00009-1

Gabriel, A., Acosta, J., & Grandey, A. (2015). The Value of a Smile: Does Emotional Performance Matter More in Familiar or Unfamiliar Exchanges? *Journal of Business & Psychology, 30*(1), 37–50. https://doi.org/10.1007/s10869-013-9329-2

Garg, S. (2018). An Empirical Study on Relevance of 360 Degree Performance Evaluation Practice with Special Reference to Delhi-NCR Private Banking Sector. *ITIHAS - The Journal of Indian Management*, *8*(3), 32–36.

Görün, M., Kayar, İ., & Varol, B. (2018). 360-Degree Performance Appraisal and Feedback System: A Study with Heads of Departments in Çanakkale Onsekiz Mart University. *Gaziantep University Journal of Social Sciences*, *17*(4), 1425–1437. https://doi.org/10.21547/jss.449154

Grandey, A., Houston III, L., & Avery, D. (2018). Fake It to Make It? Emotional Labor
Reduces the Racial Disparity in Service Performance Judgments. *Journal of Management, 20*(10), 1-30. https://doi.org/10.1177/0149206318757019

Gutermann, D., Lehmann-Willenbrock, W., Boer, D., Born, M., & Voelpel, S. (2017).
How Leaders Affect Followers' Work Engagement and Performance: Integrating
Leader−Member Exchange and Crossover Theory. *British Journal of Management, 28(2*), 299–314. https://doi.org/10.1111/1467-8551.12214

Harris, H. & Schaubroeck, J. (1988). A Meta-Analysis of Self-Supervisory, Self-Peer,
and Peer-Supervisor Ratings. *Personnel Psychology, 41*, 43-62.
https://doi.org/10.1111/j.1744-6570.1988.tb00631.x

Hosain, S. (2016). 360 Degree Feedback as a Technique of Performance Appraisal: Does
it Really Work? *Asian Business Review, 6*(13), 21-24.

Hoffman, B. & Woehr, D. (2009). Disentangling the Meaning of Multisource
Performance Rating Source and Dimension Factors. *Personnel Psychology, 62*, 735–
765. https://doi.org/10.1111/j.1744-6570.2009.01156.x

Hofstede, G. (1984). Cultural Dimensions in Management and Planning. *Asia Pacific Journal of Management, 1*(2), 81-99. https://doi.org/10.1007/BF01733682

Jong, M., Elliott, N., Nguyen, M., Goyke, T., Johnson, S., Cook, M., Lindauer, L., Best,
K., Gernerd, D., Morolla, L., Matuzsan, Z., & Kane, B. (2019). Assessment of
Emergency Medicine Resident Performance in an Adult Simulation Using a
Multisource Feedback Approach. *Western Journal of Emergency Medicine, 2*(1), 64–
70. https://doi.org/10.5811/westjem.2018.12.39844

Karabat, B., Calis, S., & Karabat, C. (2013). Performance Appraisal Systems and Implementations: A Literature Review and Research Agenda. *International Journal of Business and Management Studies, 5*(1), 103-112.

Kline, T. & Sulsky, L. (2009). Measurement and Assessment Issues in Performance Appraisal. *Canadian Psychology*, *50*(3), 161–171. https://doi.org/10.1037/a0015668

Lee, J. & Lee, Y. (2018). Job Crafting and Performance: Literature Review and Implications for Human Resource Development. *Human Resource Development Review, 17*(3), 277–313. https://doi.org/10.1177/1534484318788269

Lev, S. & Koslowsky, M. (2012). On-the-job embeddedness as a mediator between conscientiousness and school teachers' contextual performance. *European Journal of Work & Organizational Psychology, 21(1)*, 57–83. https://doi.org/10.1080/1359432X.2010.535656

Li, H. & Wong, N. (2008). Implementing Performance Indicators of early learning and teaching: A Chinese study. *International Journal of Early Years Education, 16*(2), 115–131. https://doi.org/10.1080/09669760701516942

London, M. & Smither, J. (1995). Can Multi-Source Feedback Change Perceptions of Goal Accomplishment, Self-Evaluations, and Performance-Related Outcomes? *Theory-Based Applications and Directions for Research. Personnel Psychology, 48*(4), 803–839. https://doi.org/10.1111/j.1744-6570.1995.tb01782.x

Loredana, L. & Mirabela, F. (2015). Multi-Rater Feedback - in Promoting and Stimulating an Institution or Organization Employees. *Annals of the "Constantin Brâncuşi" University of Târgu Jiu, Economy Series,* 399–403.

Lyde, A., Grieshaber, D., & Byrns, G. (2016). Faculty Teaching Performance: Perceptions of a Multi-Source Method for Evaluation (MME). *Journal of the*

*Scholarship of Teaching and Learning, 16*(3), 82–94. https://doi.org/10.14434/josotl.v16i3.18145

Markham, S., Witt Smith, J., & Braekkan, K. (2014). A new approach to analyzing the Achilles' heel of multisource feedback programs: Can we really trust ratings of leaders at the group level of analysis? *The Leadership Quarterly, 25*, 1120-1142. http://dx.doi.org/10.1016/j.leaqua.2014.10.003

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L., Estarli, M., Barrera, E., Martínez-Rodríguez, R., Baladia, E., Agüero, S., Camacho, S., Buhring, K., Herrero-López, A., Gil-González, D., Altman, D., Booth, A.…Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1). https://doi.org/10.1186/2046-4053-4-1

Motowidlo, S. (2003). Job performance. In W. Borman, D. Ilgen, R. Klimoski (Eds.), *Handbook of Psychology, Volume 12: Industrial and Organizational Psychology* (39-53). John Wiley & Sons.

Narayanan, A., Farmer, E., & Greco, M. (2018). Multisource feedback as part of the medical board of Australia's professional performance framework: Outcomes from a preliminary study. *BMC Medical Education, 18*, 323-334. https://doi.org/10.1186/s12909-018-1432-7

Noonan, C., Monagle, J., & Castanelli, D. (2011). Development of a multi-source feedback tool for consultant anaesthetist performance. *Australian Health Review, 35*(2), 141–145. https://doi.org/10.1071/AH10928

Nuryanti, A., Nursalam, N., & Triharini, M. (2017). Development of Performance Assessment Instrument for Nurses Based on Web in Inpatient Unit. *Jurnal Ners, 12*(1), 108-112. https://doi.org/10.20473/jn.v12i1.2253

Overeem, K., Wollersheimh, H., Arah, O., Cruijsberg, J., Grol, R., & Lombarts, K. (2012a). Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Medical Education, 12*(52). https://doi.org/10.1186/1472-6920-12-52

Overeem K., Wollersheim H., Arah O., Cruijsberg J., Grol R., & Lombarts K. (2012b). Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. *BMC Health Services Research, 12*(80). https://doi.org/10.1186/1472-6963-12-80

Pereira, P., dos Santos, N. R., & Pais, L. (2019). Empirical Research on Decent Work: A Literature Review. *Scandinavian Journal of Work and Organizational Psychology, 4*(1), 1-15. https://doi.org/10.16993/sjwop.53

Reinhardt, C. (1985). The State of Performance Appraisal: A Literature Review. *Human Resource Planning*, *8*(2), 105-110.

Rohlfer, S. (2018). Employee involvement and participation in SMEs: a synthesis of extant research. *Journal of Evolutionary Studies in Business*, *3*(1), 112-136. https://doi.org/10.1344/jesb2018.1.j039

Russo, S., Miraglia, M., & Borgogni, L. (2017). Reducing Organizational Politics in Performance Appraisal: The Role of Coaching Leaders for Age-diverse Employers. *Human Resource Management, 56*(5), 769-783. https://doi.org/10.1002/hrm.21799

Rynes, S., Gerhart, B., & Parks, L. (2005). Personnel Psychology: Performance Evaluation and Pay for Performance. *Annual Review of Psychology*, *56*(1), 571–600. https://doi.org/10.1146/annurev.psych.56.091103.070254

Sargeant, J., Macleod, T., Sinclair, D., & Power, M. (2011). How do physicians assess their family physician colleagues' performance? Creating a rubric to inform assessment and feedback. *Journal of Continuing Education in the Health Professions, 31*(2), 87–94. https://doi.org/10.1002/chp.20111

Schuh, S., Zhang, X., Morgeson, F., Tian, P., & van Dick, R. (2018). Are you really doing good things in your boss's eyes? Interactive effects of employee innovative work behavior and leader–member exchange on supervisory performance ratings. *Human Resource Management, 57*(1), 397–409. https://doi.org/10.1002/hrm.21851

Seaburg, L., Wang, A., West, C., Reed, D., Halvorsen, A. J., Engstler, G., Oxenteko, A., & Beckman, T. (2016). Associations between resident physicians' publications and clinical performance during residency training. *BMC Medical Education, 16* (22). https://doi.org/10.1186/s12909-016-0543-2

Selvarajan, T. & Cloninger, P. (2012). Can performance appraisals motivate employees to improve performance? A Mexican study. *International Journal of Human Resource Management, 23*(15), 3063-3084. https://doi.org/10.1080/09585192.2011.63706

Sen, A. (2000). Work and rights. *International Labour Review*, *139*(2), 119-128. https://doi.org /10.1111/j.1564-913X.2000.tb00406.x

Shahzileh, Z. & Aghajan, A. (2015). Performance appraisal: A review and case study. *International Journal of Business Excellence*, *8*(5), 584–600. https://doi.org/10.1504/IJBEX.2015.071279

Sharma, P. & Kirkman, B. (2015). Leveraging Leaders: A Literature Review and Future Lines of Inquiry for Empowering Leadership Research. *Group and Organization Management*, *40*(2), 193–237. https://doi.org/10.1177/1059601115574906

Smither, J., London, M., Vasilopoulos, M., Reilly, R., Millsap, R., & Salvemini, N. (1995). An Examination of the Effects of an Upward Feedback Program over Time. *Personnel Psychology, 48*(1), 1–34. https://doi.org/10.1111/j.1744-6570.1995.tb01744.x

Srinivas, S. & Ashok, K. (2018). The Role of Job Crafting in Feedback-Performance Relationship: A Study of Supervisors and Managers of the Indian Railways. *IUP Journal of Organizational Behavior*, *17*(4), 58–76.

Stratton, K. (1988). Performance appraisal and the need for an organizational grievance procedure: A review of the literature and recommendations for future research. *Employee Responsibilities and Rights Journal*, *1*(3), 167–179. https://doi.org/10.1007/BF01384979

Sung, S. & Choi, J. (2018). Effects of training and development on employee outcomes and firm innovative performance: Moderating roles of voluntary participation and evaluation. *Human Resource Management, 57*(6), 1339–1353. https://doi.org/10.1002/hrm.21909

Treadway, D., Breland, J., Williams, L., Cho, J., Yang, J., & Ferris, G. (2013). Social Influence and Interpersonal Power in Organizations: Roles of Performance and Political Skill in Two Studies. *Journal of Management, 39*(6), 1529–1553. https://doi.org/10.1177/0149206311410887

Tureta, C., Rosa, A., & Ávila, S. (2006). Da Teoria Sistêmica ao Conceito de Redes Interorganizacionais: Um Estudo Exploratório da Teoria das Organizações. *Revista de Administração da UNIMEP, 4*(1), 1-16.

United Nations. (2000). United Nations Millennium Declaration: Resolution 55/2 adopted by the United Nations General Assembly on 8 Sep. 2000 Declaration. United Nations website: http://www.un.org/ mil-lennium/declaration/ares552e.pdf

van der Meulen, M., Boerebach, B., Smirnova, A., Heeneman, S., oude Egbrink, M., Van der Vleuten, C., Arah, O., & Lombarts, K. (2017). Validation of the INCEPT: A Multisource Feedback Tool for Capturing Different Perspectives on Physicians' Professional Performance. *Journal of Continuing Education in the Health Professions, 37*(1), 9–18. https://doi.org/10.1097/CEH.0000000000000143

van Hooft, E., van der Flier, H., & Minne, M. (2006). Construct Validity of Multi-Source Performance Ratings: An Examination of the Relationship of Self-, Supervisor-, and Peer-Ratings with Cognitive and Personality Measures. *International Journal of Selection & Assessment, 14*(1), 67–81. https://doi.org/10.1111/j.1468-2389.2006.00334.x

van Veelen, R. & Ufkes, E. G. (2019). Teaming Up or Down? A Multisource Study on the Role of Team Identification and Learning in the Team Diversity–Performance Link. *Group & Organization Management, 44*(1), 38–71. https://doi.org/10.1177/1059601117750532

Violato, C., Lockyer, J., & Fidler, H. (2008). Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Medical Education, 42*, 1007–1013. https://doi.org/10.1111/j.13652923.2008.03127.x

Weekley, J. & Gier, J. (1989). Ceiling in the Reliability and Validity of Performance Ratings: The Case of Expert Raters. *Academy of Management Journal, 32,* 213-222. https://doi.org/10.5465/256428

Weigl, M., Heinrich, M., Keil, J., Wermelt, J. Z., Bergmann, F., Hubertus, J., & Hoffmann, F. (2019). Team performance during postsurgical patient handovers in paediatric care. *European Journal of Pediatrics*, 1-10. https://doi.org/10.1007/s00431-019-03547-w

Whitaker, B. (2012). The Influence of Covariates on Latent Job Performance Ratings from Differing Sources: A Mimic Modeling Approach. *International Journal of Organization Theory & Behavior, 15*(2), 176–198. https://doi.org/10.1108/IJOTB-15-02-2012-B002

Williams, B., Byrne, P, Williams, N., & Williams, M. (2017). Dreyfus and Dreyfus and Indicators of Behavioral Performance:A Study of Measurement Convergence. *Journal of Continuing Education in the Health Professions, 37*(1), 50–54. https://doi.org/10.1097/CEH.0000000000000138

Wyland, R., Lester, S., Ehrhardt, K., & Standifer, R. (2016). An Examination of the Relationship Between the Work-School Interface, Job Satisfaction, and Job Performance. *Journal of Business & Psychology, 31*(2), 187–203. https://doi.org/10.1007/s10869-015-9415-8

Yammarino, F. & Atwater, L. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management, 32*, 231-247. https://doi.org/10.1002/hrm.3930320204