CIPD

Championing better work and working lives

PEOPLE PERFORMANCE DEFINITIONS AND MEASURES

An evidence review

Scientific summary

June 2022

The CIPD is the professional body for HR and people development. The registered charity champions better work and working lives and has been setting the benchmark for excellence in people and organisation development for more than 100 years. It has more than 160,000 members across the world, provides thought leadership through independent research on the world of work, and offers professional training and accreditation for those working in HR and learning and development.

People performance definitions and measures: an evidence review

Scientific summary

Contents

Acknowledgements

This report was written by Eric Barends, Denise Rousseau and Emilia Wietrak of the Center for Evidence-Based Management (CEBMa), and Jonny Gifford and Jake Young of the CIPD.

Publication information

Please cite this report as: Barends, E., Rousseau, D., Wietrak, E., Gifford, J. and Young, J. (2022) *People performance definitions and measures: an evidence review.* Scientific Summary. London: Chartered Institute of Personnel and Development.

This report and the accompanying practice summary are available at cipd.co.uk/evidence-people-performance

1 Introduction

Rationale for this review

Improving people performance is a key issue for HR professionals, both in terms of their own decision-making, and where they need to show evidence of the value of investing in people to their colleagues and leaders. Performance is the most common outcome considered in management and organisations. To establish how to improve performance, employers need to have a solid understanding of what they mean by performance and how to measure it.

For this reason, the Chartered Institute of Personnel and Development (CIPD) approached the Center for Evidence-Based Management (CEBMa) to undertake a review of the research literature to learn more about the evidence on measuring people performance at both individual and team levels (that is, excluding organisational-level effectiveness or performance). This review presents an overview of the findings.

What is a rapid evidence assessment?

Evidence reviews come in many forms. One of the best known is the conventional literature review, which provides an overview of the relevant scientific literature published on a topic. However, a conventional literature review's trustworthiness is often low: clear criteria for inclusion are typically lacking and studies are selected based on the researcher's personal preferences. As a result, conventional literature reviews are prone to severe bias. For this reason, 'rapid evidence assessments' (REAs) are used. This type of review uses a specific research methodology to identify the most relevant studies on a specific topic as comprehensively as possible, and to select appropriate studies based on explicit criteria. In addition, the methodological quality of the studies included is assessed by independent reviewers using explicit criteria. In contrast to a conventional literature review, a REA is transparent, verifiable and reproducible, and, as a result, the likelihood of bias is considerably smaller.

Main question: What will the review answer?

What is known in the scientific literature about the measurement of people performance?

Sub-questions that form the basis of the review are:

1 What is meant by people performance?
2 How can performance be measured in a valid and reliable way?

2 Methods

Search strategy: How was the evidence sought?

To answer the review questions, the following databases were used: ABI/INFORM Global from ProQuest, Business Source Premier from EBSCO, PsycINFO from Ovid, APA PsycTests and Google Scholar. For our search, the following general search filters were applied:

1 scholarly journals, peer-reviewed
2 published in the period 2000–2021
3 articles in English.

We conducted nine different search queries, which yielded 1,200+ papers. An overview of all search terms and queries is provided in **Appendix 1**.

Selection: How were studies selected?

Study selection took place in two phases. First, titles and abstracts of the 1,200+ studies identified were screened for relevance. In case of doubt or lack of information, the study was included. Duplicate publications were removed. This first phase yielded nine meta-analyses and 47 primary studies. Second, studies were selected based on the full text of the article using these inclusion criteria:

1 **type of studies**: focusing on quantitative, empirical studies
2 **measurement**: only studies in which people performance was quantitatively measured
3 **context**: only studies related to workplace settings
4 **level of trustworthiness**: only studies that were graded Level C or above (see below).

Further, the following exclusion criteria were applied for primary studies:

- studies that focus on other types of performance measurement, such as financial performance, process optimalisation, logistics or supply chain performance
- studies that focus on the measurement of organisational performance
- studies in which performance measurement is used to test a behavioural model or theoretical framework
- studies that use performance measures developed to evaluate the outcomes of occupational therapy or medical interventions
- studies that provide insufficient information regarding the psychometric qualities of the measures, scales or questionnaires used.

This second phase yielded a total number of eight meta-analyses and 36 primary studies. An overview of the selection process is provided in **Appendix 2**.

Data extraction: What data was extracted?

From each study, information relevant to the review question, such as year of publication, research design, sample size, population (for example, industry, type of employees), type of practice, possible moderators or mediators, main findings, and effect sizes were extracted. An overview of all data extracted is provided in **Appendix 3**.

Critical appraisal: How was the quality of the included studies judged?

Methodological appropriateness

The classification system of Shadish et al (2002), and Petticrew and Roberts (2006) was used to determine the methodological appropriateness of the research design of the studies included. Any discrepancies were resolved through discussion or by consulting a third party where necessary. The following levels of appropriateness were used for the classification:

Purpose	Example	Study design				
		RCT	CBA	C/BA	Cross	Qual
Effect, impact	Does A have an effect/impact on B? What are the critical success factors for A? What are the factors that affect B?	A	B	C	D	n/a

Association	Is A related to B? Does A often occur with B? Do A and B co-vary?	A	A	A	A	n/a
Frequency	How often does A occur? How many people prefer A?	n/a	n/a	n/a	A	n/a
Difference	Is there a difference between A and B?	n/a	n/a	A	A	n/a
Attitude, opinion	What is people's attitude toward A? Are people satisfied with A? Do people agree with A?	n/a	n/a	n/a	A	C
Experience, perceptions, feelings, needs	What are people's experience with A? What are people's feelings about A? What are people's perceptions about A?	n/a	n/a	n/a	B	A
Exploration, Theory-building	Why does A occur? Why is A different from B? In what context does A occur?	n/a	n/a	n/a	B	A

RCT = randomised controlled trial; CBA = non-randomised controlled before–after study; C = controlled study; BA = before–after study; Cross = cross-sectional study; Qual = qualitative study; n/a = not appropriate.

Methodological quality

To determine methodological quality, all the studies included were systematically assessed, based on explicit quality criteria, such as the PRISMA (Moher et al 2009) and the CONSORT statement (Moher et al 2001), the CASP checklists (Critical Appraisal Skills Programme, n.d.) and the critical appraisal criteria developed by the Center for Evidence-Based Management. Based on a tally of the number of weaknesses, the trustworthiness was downgraded. The final level was determined as follows: downgrade one level if two weaknesses were identified, downgrade two levels if four weaknesses were identified, and so on.

Effect sizes

To determine the magnitude of an effect, Cohen's rule of thumb (Cohen 1988) was applied. According to Cohen (1988) a 'small' effect is one that is only visible through careful examination. A 'medium' effect, however, is one that is 'visible to the naked eye of the careful observer'. Finally, a 'large' effect is an effect that anyone can easily see because it is substantial.

3 Main findings

Outcome of the appraisal: What is the quality of the studies included?

The final number of studies included in this review was 44 (8 meta-analyses and 36 primary studies), indicating that the measurement of people performance is a well-established topic and is based on a large body of research. In addition, three of the meta-analyses included controlled and/or longitudinal studies, and most of the performance measurement tools that were used in the primary studies showed good reliability and validity, indicating that the quality of the empirical evidence is moderate to high. An overview of all study characteristics and the data extracted can be found in **Appendix 3**.

Question 1: What is meant by people performance?

In management and business, 'performance' is probably the most widely used outcome measure to assess whether a person (such as an employee) or a group of persons (such as a team) have achieved their goals. Likewise, people performance may well be the single most studied and used dependent variable in organisational and personnel psychology. A high performance indicates that people do an excellent job, exceed accepted standards and work to the best of their ability, whereas a low performance suggests that people can do better. Looking more closely, however, it is not always clear what people performance is and how it can be measured. In fact, the term is used for a wide range of different constructs, and can refer to both the outcome of an activity as well as the way that outcome was achieved. For clarity, we provide an overview of the three most widely used dimensions of people performance.

Task performance

Related terms: in-role performance, work role performance, job performance

Task performance refers to the execution and outcome of job-specific activities that are part of one's formal job description. It concerns the core job responsibilities of an employee and is often tied to specific quantitative and/or qualitative work outcomes, as well as the way these outcomes are delivered (Borman 2017; Sonnentag et al 2008). For some occupations and functions, indicators of task performance are relatively straightforward. For example, task performance indicators for a firefighter may include responding to fire alarms, extinguishing fires, performing rescue operations and mitigating chemical spills. For other occupations and functions, however, it may be quite difficult to find valid and reliable indicators. For example, knowledge workers seldom have one single, standard outcome. In addition, the outcome of their tasks is often hard to quantify and contingent on a wide range of contextual and situational factors outside an employee's control (Ramirez and Steudel 2008).

Contextual performance

Related terms: extra-role performance, organizational citizenship behavior (OCB)

Contextual performance refers to activities that go beyond the formal job description. It concerns voluntary behaviour that contributes to the organisation's social and psychological climate: this acts in support of employee task activities, or benefits the organisation as a whole (Borman 2001; Harper 2015; Podsakoff et al 2009). Examples of contextual performance are: helping co-workers finish a project, coaching junior co-workers, and organising or participating in the company's social events. Often the way employees interact with co-workers and whether their behaviours and actions reflect the company's values are also considered contextual performance; hence the term 'organizational citizenship behavior' is widely used.

Adaptive performance

Related terms: flexible work behaviour, agile performance, individual creative/innovative performance (CIP)

Adaptive performance refers to employees' ability to adapt and adjust to unforeseen changes and demands in the workplace. It concerns an employee's capability to efficiently deal with new, uncertain or unpredictable work situations (Harari et al 2016; Jundt et al 2015; Pulakos et al 2000; Sonnentag et al 2008). Examples of adaptive performance are: handling crisis situations, solving problems creatively, coping with work stress, learning new tasks and procedures, proposing new, creative and innovative ways of working, and participating in change initiatives.

Individual versus team performance

In addition to these three performance dimensions, a distinction can be made between individual performance and team performance. In most cases, team performance is simply the sum of team

members' individual performance. However, in some functions, key tasks are performed in collaboration with others, in particular where complex tasks require the input and expertise of multiple employees – in those cases, outcome measures at the team level should be used.

Outcome versus process performance

Finally, performance can refer to both the outcome of an activity and the way that outcome was achieved. For this reason, sometimes a distinction is made between outcome performance and process or behavioural performance. Some scholars argue that task performance is more outcome-focused, whereas contextual and adaptive performance are more process- or behaviour-focused.

Question 2: How can performance be measured in a valid and reliable way?

Objective versus subjective measures of performance

When measuring employees' performance, often a distinction is made between objective and subjective measures. Objective measures typically concern measures of countable behaviours or outcomes, whereas subjective measures consist of a supervisor's or co-workers' ratings of an employee's performance. Another widely used subjective measure of performance is the self-report measure, where employees rate their own performance. In the past decades, many studies have been published on the topic of objective versus subjective measures of performance. The most relevant findings are provided below.

Finding 1: Both objective and subjective measures of performance have validity and reliability issues (Level A)

It is often assumed that 'hard' quantifiable outcome measures are the most objective; that is, the most valid and reliable indicators of employee task performance. A wide range of hard outcome measures are used, often based on the specific output of an employee's task. For example, a hard outcome measure for an orthopaedic surgeon may include the number of patients treated in the past month, the number of surgical procedures performed, or the number of patients re-admitted due to medical complications.

However, a large number of studies have consistently demonstrated that hard outcome measures are a less valid and reliable indicator of employees' task performance than often expected (Bommer et al 1995; Rich et al 1999; Roth et al 2012; Sturman et al 2005). For example, the surgeon mentioned above may also teach and supervise junior doctors and, as a result, may treat fewer patients than other surgeons. In addition, the surgeon may conduct more complex surgical procedures with a higher chance of medical complications (and thus re-admissions) than less experienced colleagues.

Subjective measures of performance, however, have similar issues. For example, a large number of studies have demonstrated that subjective measures, such as supervisor and peer ratings of contextual performance, may be negatively biased by an employee's ethnicity, gender, age or sexual orientation (Bowen et al 2000; Kraiger and Ford 1985) or affected by the quality of their relationship with the employee (Elicker et al 2006; Sutton et al 2013). In addition, self-report ratings of performance may be influenced by personality traits such as self-esteem or confidence.

Finally, it was found that the purpose of the performance measurement affects its validity and reliability. For example, supervisor and peer ratings obtained for administrative purposes (for example, decisions on promotion and compensation) tend to be higher than those obtained for employee development purposes (Jawahar and Williams 1997; Salgado and Moscoso 2019).

Finding 2: The correlation between objective and subjective performance measures is low, indicating that these measures are not interchangeable (Level A)

Several meta-analyses have found that, in general, the relationship between objective and subjective performance is low (Bommer et al 1995; Rich et al 1999; Roth et al 2012; Sturman et al 2005). Further, this relationship was not affected by contextual factors, such as job type or gender. The relationship is somewhat stronger, however, when subjective measures like supervisory ratings are based on a comparison with a standard rather than relative to the performance of other employees (Heneman 1986). This weak relationship between objective and subjective measures of performance indicates that the measures are not interchangeable and cannot be used as a proxy for one another.

Finding 3: A combination of objective and subjective performance measures can lead to a more accurate measure of an employee's true task performance (Level A)

Subjective measures of performance may be biased, and, for this reason, objective 'hard outcome' measures are often used. However, hard outcome measures are susceptible to contextual factors that are outside an employee's control. Supervisors and co-workers are often aware of these factors and can take them into account when evaluating an employee's performance (Heneman 1986; Rich et al 1999; Sturman et al 2005). A combination of objective and subjective performance measures can therefore lead to a more accurate measure of an employee's true task performance, provided that they account for possible bias.

Measuring task, contextual or adaptive performance

Although sometimes 'overall' or 'general' measures of performance are used, in most cases task, contextual and adaptive performance are measured separately. In past decades, numerous primary studies and meta-analyses have measured task, contextual and adaptive performance as their main outcome variables. Some relevant findings are provided below.

Finding 4: Measures of task, contextual and adaptive performance assess different things (Level A)

Task, contextual and adaptive performance are related but are empirically different dimensions of people performance (see above). This means that measures of task, contextual and adaptive performance measure different things (Borman and Motowidlo 1997; Harari et al 2016; Rich et al 1999; Salgado and Moscoso 2019). Consequently, they cannot be used as a proxy for one another. For example, when it is determined that the number of sales is the most important indicator of a sales agent's task performance, their reward should not be based on a supervisory rating that includes contextual performance. Conversely, if all dimensions of performance are deemed equally important, it is inappropriate to reward the sales agent solely on gross sales.

Finding 5: Ratings of task, contextual and adaptive performance mutually influence each other when rated by the same person (Level A)

Although task and contextual performance are different dimensions, several meta-analyses found that they are correlated (Hoffman et al 2007). One explanation for this finding is that ratings of task, contextual and adaptive performance mutually influence each other, in particular when rated by the same person. Indeed, a meta-analysis of 81 controlled studies has shown that employee contextual performance influences supervisor rating of task performance (Podsakoff et al 2013). Similarly, a meta-analysis found that individual creative and innovative performance, a subtype of adaptive performance, was positively related to both task performance and contextual performance when rated by the same person (Harari et al 2016). An explanation for this finding is what is

referred to as common method variance (CMV), that is, *'biasing effects that measuring two or more constructs with the same method have on estimates of the relationships between them'* (Podsakoff et al 2012). Put differently, when an employee's task, contextual and adaptive performance are assessed by the same person (such as a supervisor), these performance ratings may be biased. For example, supervisors may interpret adaptive and contextual performance as behavioural manifestations of commitment and/or loyalty (Allen and Rush 2001), which may positively influence their performance ratings. It was found that common rater effects (CMV) may lead to ratings that are 60–90% higher compared with ratings from different raters (Podsakoff et al 2013).

Finding 6: Employee performance remains stable over time (Level A)

Although it is apparent that the performance of employees changes as they learn and develop on the job, a meta-analysis of 23 longitudinal studies that measured people's performance over three or more time periods found that individual performance tends to be stable over time (Sturman et al 2005). Although most studies found that performance *ratings* tend to change over time, there is strong evidence that this variation is due to a lack of stability and test–retest reliability of the performance measures used. True performance actually tends to remain stable. Objective measures of task performance specifically are associated with lower test–retest reliability, particularly for highly complex jobs.

Performance measurement scales

When subjective measures of performance are used, many organisations use survey questionnaires in which respondents are asked to indicate the degree to which they agree or disagree with statements. An example is: 'In the past three months, I was able to carry out my work well with minimal time and effort,' followed by a five-point Likert scale ranging from 'seldom' to 'always'. In the past decades, a wide range of survey questionnaires – in academia referred to as 'scales' – were developed that measure (dimensions of) people performance. This review identified 36 different scales (an overview is provided in **Appendix 3**).

Scales measuring task performance

Most of the scales identified in this review measure (elements of) task performance. As explained, what constitutes task performance depends on the specific activities that are part of someone's formal job description. For this reason, numerous scales are available for different occupations and functions. For example, there are measurement scales for the task performance of nurses (Karayurt et al 2009), sales agents (Amyx et al 2009), account managers (Liu et al 2018), university lecturers (Molefe 2010), physicians (Wright et al 2012), and police officers (Tarescavage et al 2015). In addition, there are scales that focus only on a specific element of task performance, such as service performance (Ali et al 2017) or safety performance (Valenzuela and Burke 2020). Only a limited number of scales measure 'general' task performance, independent of the employee's function or occupation. It should be noted, however, that although many scales are available, the underlying research to establish their reliability and validity is rather limited. A widely used 'generic' scale is the Individual Work Performance Questionnaire (IWPQ), an 18-item self-report scale shown to have acceptable psychometric qualities (Koopmans et al 2014, 2016).

Scales measuring contextual performance

Although scales exist to measure contextual performance in a specific function or occupation (Carlos and Rodrigues 2016; Greenslade and Jimmieson 2017), most contextual performance scales are generic. Widely used scales are the IWPQ (see above) and the Organizational Citizenship Behavior (OCB) scale developed by Podsakoff et al (1990).

Scales measuring adaptive performance

Adaptive performance refers to an employee's capability to deal with novel, uncertain or unpredictable work situations. Some scales assessing task performance also measure elements of adaptive performance, such as responsiveness (Amyx et al 2009), behavioural flexibility (Darr et al 2017) or learning ability (Lo and Li 2005). Scales that solely measure adaptive performance, however, are often generic. A widely used generic scale is the Job Adaptability Inventory (JAI), a self-report scale developed by Pulakos et al (2000) that measures eight dimensions of adaptive behaviour.

4 Conclusion

The findings of this review show that, for several reasons, measuring people's performance is not easy.

First, performance types and dimensions cannot be treated as substitutes for one another, as they measure different things. It is therefore important that organisations clearly define what constitutes performance and what dimension(s) they value most.

Second, one type of performance measure is not necessarily more valid and reliable than another. Regardless of whether it concerns task, contextual, process, subjective or objective performance, all types of performance measures can be (in)valid and (un)reliable, depending on their purpose and application. For this reason, a combination of performance measures should be used, preferably from multiple sources or raters.

Third, although many scales exist to measure performance, the underlying research to establish their reliability and validity is rather limited.

Limitations

This REA aims to provide a balanced assessment of what is known in scientific literature about the measurement of performance by using the systematic review method to search and critically appraise empirical studies. However, to be 'rapid', concessions were made in relation to the breadth and depth of the search process, such as the exclusion of unpublished studies, the use of a limited number of databases and a focus on empirical research published in the period 2000–2021. As a consequence, some relevant studies may have been missed.

A second limitation concerns the psychometric qualities of scales and questionnaires. Studies in peer-reviewed journals tend to meet basic psychometric standards, but generalisability to other settings cannot be assumed, and reliability and validity should be established in the setting in which used.

Given these limitations, care must be taken not to present the findings presented in this REA as conclusive.

5 References

Abbas, M. and Raja, U. (2015) Impact of psychological capital on innovative performance and job stress. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*. Vol 32, No 2. pp128–38.

Ali, F., Hussain, K. and Ryu, K. (2017) Resort hotel service performance (RESERVE) – an instrument to measure tourists' perceived service performance of resort hotels. *Journal of Travel and Tourism Marketing*. Vol 34, No 4. pp556–69.

Allen, T.D. and Rush, M.C. (2001) The influence of ratee gender on ratings of organizational citizenship behavior. *Journal of Applied Social Psychology*. Vol 31, No 12. pp2561–87.

Amyx, D. and Bhuian, S. (2009) Salesperf: the salesperson service performance scale. *The Journal of Personal Selling and Sales Management*. Vol 29, No 4.

Barends, E., Rousseau, D.M. and Briner, R.B. (eds) (2017) *CEBMa guideline for rapid evidence assessments in management and organisations*. Version 1.0. Amsterdam: Center for Evidence Based Management.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M. and Mackenzie, S.B. (1995) On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personnel Psychology*. Vol 48, No 3.

Borman, W.C. and Motowidlo, S.J. (1997) Task performance and contextual performance: the meaning for personnel selection research. *Human Performance*. Vol 10, No 2. pp99–109.

Bowen, C-C., Swim, J.K. and Jacobs, R.R. (2000) Evaluating gender biases on actual job performance of real people: a meta-analysis. *Journal of Applied Social Psychology*. Vol 30, No 10. pp2194–2215.

Carlos, V.S. and Rodrigues, R.G. (2016) Development and validation of a self-reported measure of job performance. *Social Indicators Research*. Vol 126, No 1. pp279–307.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Darr, W., Borman, W.C., St-Pierre, L., Kubisiak, C. and Grossman, M. (2017) An applied examination of the computerized adaptive rating scale for assessing performance. *International Journal of Selection and Assessment*. Vol 25, No 2. pp149–53.

Elicker, J.D., Levy, P.E. and Hall, R.J. (2006) The role of leader-member exchange in the performance appraisal process. *Journal of Management*. Vol 32, No 4. pp531–51.

Greenslade, J.H. and Jimmieson, N.L. (2007) Distinguishing between task and contextual performance for nurses: development of a job performance scale. *Journal of Advanced Nursing*. Vol 58, No 6. pp602–11.

Harari, M.B., Reaves, A.C. and Viswesvaran, C. (2016) Creative and innovative performance: a meta-analysis of relationships with task, citizenship, and counterproductive job performance dimensions. *European Journal of Work and Organisational Psychology*. Vol 25, No 4. pp495–511.

Harper, P.J. (2015) Exploring forms of organisational citizenship behaviors (OCB): antecedents and outcomes. *Journal of Management and Marketing Research*. Vol 18, No 1.

Heneman, R.L. (1986) The relationship between supervisory ratings and results-oriented measures of performance: a meta-analysis. *Personnel Psychology*. Vol 39, No 4. p811.

Hoffman, B.J., Blair, C.A., Meriac, J.P. and Woehr, D.J. (2007) Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology*. Vol 92. pp555–66.

Jawahar, I.M. and Williams, C.R. (1997) Where all the children are above average: the performance appraisal purpose effect. *Personnel Psychology*. Vol 50, No 4. pp905–25.

Jundt, D.K., Shoss, M.K. and Huang, J.L. (2015) Individual adaptive performance in organisations: a review. *Journal of Organisational Behavior*. Vol 36, No S1. ppS53–S71.

Karayurt, Ö., Mert, H. and Beser, A. (2009) A study on development of a scale to assess nursing students' performance in clinical settings. *Journal of Clinical Nursing*. Vol 18, No 8. pp1123–30.

Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., van Buuren, S., van der Beek, A.J. and de Vet H.C.W. (2014) Improving the Individual Work Performance Questionnaire using Rasch analysis. *Journal of Applied Measurement*. Vol 15, No 2. pp160–75.

Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., de Vet, H.C.W. and van der Beek, A.J. (2014) Construct validity of the Individual Work Performance Questionnaire. *Journal of Occupational and Environmental Medicine*. Vol 56, No 3. pp331–37.

Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., Lerner, D., de Vet, H.C.W. and van der Beek, A.J. (2016) Cross-cultural adaptation of the Individual Work Performance Questionnaire. *Work*. Vol 53, No 3. pp609–19.

Kraiger, K. and Ford, J.. (1985) A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*. Vol 70, No 1. pp56–65.

Král, M. (2021) 20-year history of performance measurement in the local public sector: a systematic review. *International Journal of Public Administration*. pp1–15.

Lefkowitz, J. (2000) The role of interpersonal affective regard in supervisory performance ratings: a literature review and proposed causal model. *Journal of Occupational and Organisational Psychology*. Vol 73. pp67–85.

Liu, Y., Huang, Y. and Fan, H. (2018) Influence tactics, relational conditions, and key account managers' performance. *Industrial Marketing Management*. Vol 73. pp220–31.

Lo, K.K.Y. and Li, E.P.Y. (2005) Content validation on the Work Performance Rating Scale for sheltered workshop workers. *Work*. Vol 25, No 4. pp341–46.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and Prisma Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*. Vol 6, No 7. e1000097.

Moher, D., Schulz, K.F., Altman, D. and CONSORT Group. (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association*. Vol 285, No 15. pp1987–91.

Molefe, G.N. (2010) Performance measurement dimensions for lecturers at selected universities: an international perspective. *South African Journal of Human Resource Management*. Vol 8, No 1. pp1–13.

Petticrew, M. and Roberts, H. (2006) *Systematic reviews in the social sciences: a practical guide*. New York: John Wiley & Sons, pp125–63.

Podsakoff, P.M., MacKenzie, S.B., Moorman, R.H. and Fetter, R. (1990) Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organisational citizenship behaviors. *The Leadership Quarterly*. Vol 1, No 2. pp107–42.

Podsakoff, N.P., Whiting, S.W., Podsakoff, P.M. and Blume, B.D. (2009) Individual- and organisational-level consequences of organisational citizenship behaviors: a meta-analysis. *Journal of Applied Psychology*. Vol 94. pp122–41.

Podsakoff, P.M., MacKenzie, S.B. and Podsakoff, N.P. (2012) Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*. Vol 63. pp539–69.

Podsakoff, N.P., Whiting, S.W., Welsh, D.T. and Mai, K.M. (2013) Surveying for 'artifacts': the susceptibility of the OCB–performance evaluation relationship to common rater, item, and measurement context effects. *Journal of Applied Psychology*. Vol 98, No 5. pp863–74.

Pulakos, E.D., Arad, S., Donovan, M.A. and Plamondon, K.E. (2000) Adaptability in the workplace: development of a taxonomy of adaptive performance. *Journal of Applied Psychology*. Vol 85, No 4. p612.

Ramirez, Y.W. and Steudel, H.J. (2008) Measuring knowledge work: the knowledge work quantification framework. *Journal of Intellectual Capital*. Vol 9, No 4. pp564–84.

Rich, G.A., Bommer, W.H., MacKenzie, S.B., Podsakoff, P.M. and Johnson, J.L. (1999) Methods in sales research: Apples and apples or apples and oranges? A meta-analysis of objective and subjective measures of salesperson performance. *Journal of Personal Selling and Sales Management*. Vol 19, No 4. pp41–52

Roth, P.L., Purvis, K.L. and Bobko, P. (2012) A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management*. Vol 38, No 2. pp719–39.

Salgado, J.F. and Moscoso, S. (2019) Meta-analysis of interrater reliability of supervisory performance ratings: effects of appraisal purpose, scale type, and range restriction. *Frontiers in Psychology*. Vol 10.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

Shaughnessy, J.J. and Zechmeister, E.B. (1985) *Research methods in psychology*. New York: Alfred A. Knopf.

Sonnentag, S., Volmer, J. and Spychala, A. (2008) Job performance. In: Barling, J. and Cooper, C. (eds), *The Sage handbook of organizational behavior*. Volume 1. London: Sage, pp427–47.

Sturman, M.C., Cheramie, R.A. and Cashen, L.H. (2005) The impact of job complexity and performance measurement on the temporal consistency, stability, and test–retest reliability of employee job performance ratings. *Journal of Applied Psychology*. Vol 90, No 2. pp269–83.

Sutton, A.W., Baldwin, S.P., Wood, L. and Hoffman, B.J. (2013) A meta-analysis of the relationship between rater liking and performance ratings. *Human Performance*. Vol 26, No 5. pp409–29.

Tarescavage, A.M., Corey, D.M. and Ben-Porath, Y.S. (2015) Minnesota Multiphasic Personality Inventory–2–Restructured Form (MMPI-2-RF) predictors of police officer problem behavior. *Assessment*. Vol 22, No 1. pp116–32.

Valenzuela, L.S. and Burke, M.J. (2020) Toward a greater understanding of Colombian professional truck drivers' safety performance. *Transportation Research Part F: Traffic Psychology and Behaviour*. Vol 73. pp188–204.

Wright, C., Richards, S.H., Hill, J.J., Roberts, M.J., Norman, G.R., Greco, M., et al (2012) Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council Patient and Colleague Questionnaires. *Academic Medicine*. Vol 87, No 12. pp1668–78.

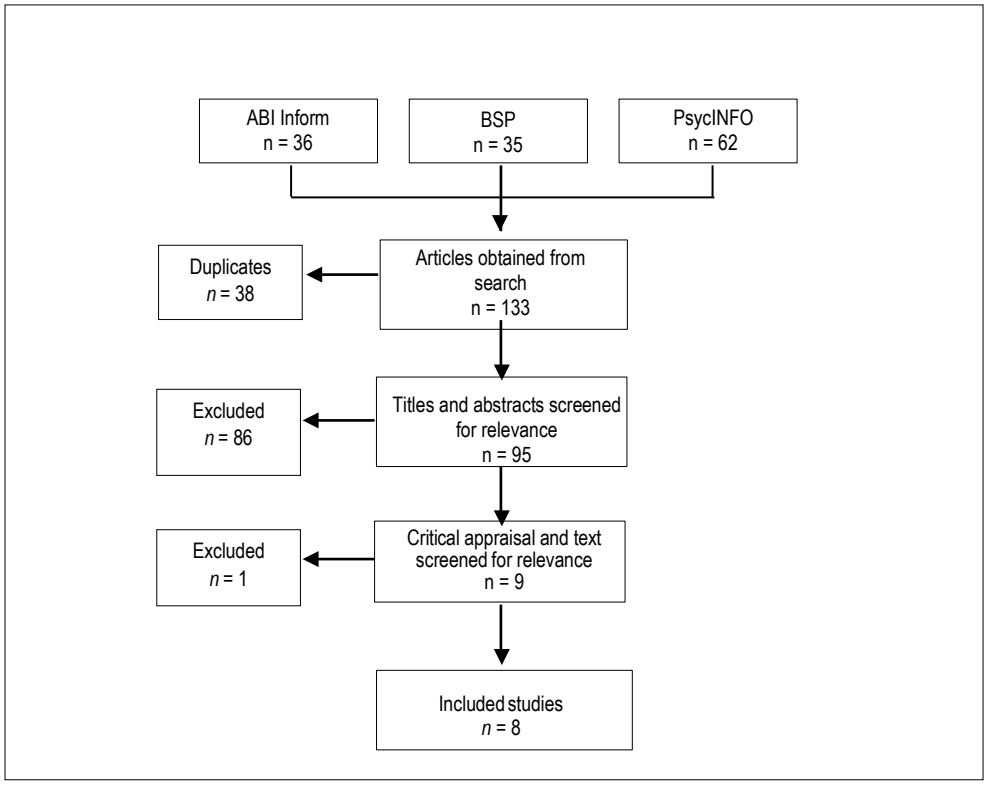
Appendix 1: Search terms and results

ABI/Inform Global, Business Source Elite, PsycINFO, peer-reviewed, scholarly journals, April 2021			
Search terms	**ABI**	**BSP**	**PSY**
S1: ti(perform*)	64,610	91,683	67,645
S2: ti(measur*) OR ti(scale*) OR ti(questionnaire*) OR ti(indicator*)	58,390	83,021	126,526
S3: S1 AND S2	5,262	6,412	3,623
S4: S1 AND S2 filter meta-analyses and/or systematic reviews	**36**	**35**	**62**
S5: ab('reliability') OR ab('validity') OR ab('consistency') OR ab('psychometric')	60,948	87,706	185,630
S6: S3 AND S5 NOT S4, limit > 2000	**330**	**322**	**373**

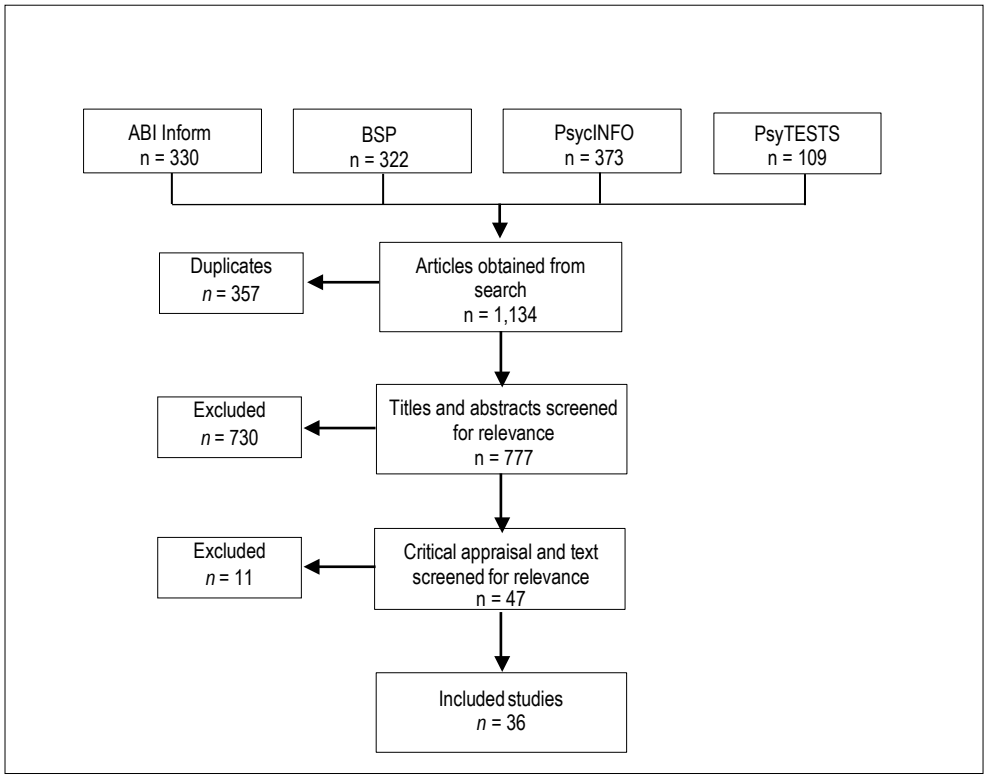
PsycTests, July 2021	
Search terms	**PSY-T**
S1: ti('performance'), Limiters – age group: Adulthood (18 yrs and older); Instrument type: Rating scale; APA PsycTests Classification: 7000 Organisational, Occupational, and Career Development	18
S2: ti('performance'), Limiters – age group: Adulthood (18 yrs and older); Instrument type: Inventory/questionnaire; Language: English; APA PsycTests Classification: 7000 Organisational, Occupational, and Career Development	56
S3: zw('job performance') or zw('job performance measure') or zw('work performance') or zw('employee performance') or zw('employee performance evaluation') or zw('performance appraisal') or zw('performance assessment') or zw('performance assessment tool') or zw('performance assessment tool') or zw('performance assessments') or zw'(performance evaluation') or zw('performance feedback') or zw('performance measure') or zw('performance measurement') or zw('performance measures') or zw('performance feedback'). Limiters – Age group: Adulthood (18 yrs and older)	47
S4: S1 OR S2 OR S3, Limiters – Release/Update Date: 20000101–20211231; Language: English	**109**

Appendix 2: Study selection

Meta-analyses and systematic reviews



Primary studies



Appendix 3a: Data extraction – meta analyses

1st author and year	Design and sample size	Sector/ population	Main findings	Effect sizes	Limitations	Level
1 Bommer (1995)	meta-analysis, k=40, s=50	currently employed workers being evaluated by their supervisor	1. The overall correlation between objective and subjective performance measures is low, indicating that the measures are not interchangeable. 2. The relationship between subjective and objective performance measures is stronger when the objective measure assessed concerns (a) performance quantity, rather than (b) performance quality. 3. Job type does not moderate the objective/subjective performance measure relationship.	1. r=.39 95% CI [.27 to .37] 2a. perf. quantity r=.38 95% CI [.26 to .37] 2b. perf. quality r=.24 95% CI [.13 to .26] 3=ns (sales r=.41, non-sales r=.34)	limited search design of the included studies not reported subgroup analyses (H3-H4) contained only three samples	A (ass, diff)
2 Harari (2016)	meta-analysis, k=40, s=39	not specified	1. Overall, individual creative and innovative performance (CIP) was positively related to (a) task performance and (b) OCB and negatively related to (c) CWB (ρ=−.23). 2. CIP measurement, rating source, OCB target, or CWB type did not moderate these relationships.	1a. task performance: ρ=.55 1b. OCB ρ=.56 1c. CWB ρ=−.23	design of included studies not reported	A (ass)
3 Henemann (1986)	meta-analysis k=23,	not specified	1. The relationship between supervisory ratings and results measures of employee performance is weak. 2. The relationship between supervisory ratings and results measures of employee performance is stronger when relative (versus absolute) ratings used. 3. The relationship between supervisory ratings and results measures of employee performance is stronger when composite (versus overall) ratings are used. Absolute comparison = employee compared with standards Relative comparison = employee compared with employee Overall = performance rated on a one-item scale Composite = performance rated on a multi-item scale (scores averaged to a final rating)	1. r=.27 2. relative r=.66 absolute r=.21 3. overall r=.19 composite r =.37	very limited search design of the included studies not reported unclear what constitutes 'results measures'	B (ass, diff)

4 Podsakoff (2013)	Meta-analysis of controlled studies (lab experiments excluded) k=81 s=173 n=31,146	not specified	Study examines the extent to which common rater, item, and measurement context characteristics bias the relationships between organisational citizenship behaviours and performance evaluations (= common method variance, CMV). 1. The correlation between employees' OCB and their performance evaluation is moderate. 2. CMV substantially biases the correlations. When taken together with other study-level predictors, CMV accounted for over half of the between-study variance in the focal correlations.	1. r=.47 2. Sources of CMV led to estimates that were between 60% and 96% larger when comparing measures obtained from a common rater (versus different raters)	no serious limitations	A
5 Rich (1999)	meta-analysis k=21 n=4,092	salespersons	1. The correlation between salesperson's objective and subjective measures of sales performance is moderate. 2. The relationship between subjective and objective measures is stronger when composite (versus overall) ratings are used. Thus, the findings suggest that objective and subjective measures of salesperson performance are not interchangeable	1. r=.45 2. overall r=.36 composite r=.48	limited search design of the included studies not reported	A (diff, ass)
6 Roth (2012)	meta-analysis, includes RCTs k=61 n=45,733	studies involving students were excluded	1. Gender differences on measures of (a) subjective and (b) objective job performance are very small.	1a. d=−.14 1b. d=−.02	no serious limitations	AA (eff)
7 Salgado (2019)	meta-analysis s=224		Examines the variance of the interrater reliability coefficients of supervisory ratings of overall-, task-, contextual-, and positive job performance. 1. Interrater reliability is larger for ratings collected for research purposes than for administrative purposes. 2. Interrater reliability is greater for multi-item (multi-scale) measures than for single-item scales.	1a. Overall job perf adm r=.45 res r=.61 1b. Task perf adm r=.38 res r=.52 1c. Cont perf adm r=.36 res r=.56 2. Partially supported, but the differences are small	design of the included studies not reported	C
8 Sturman (2005)	meta-analysis of longitudinal studies (=three or more time periods) k=22 n=4,294		Examines what portion of performance dynamism is attributable to a lack of stability in individual job performance versus test–retest unreliability. 1. As the time lag between performance measures increases, the correlation between those measures decreases (note: this decrease is non-linear). 2. Objective measures of performance are associated with lower test–retest reliability. 3. The test–retest reliability of individual job performance is lower in more complex jobs.	1. r=.57 95% CI [.53 to .61] 2. subj + low com r =.83 subj + hi com r=.72 obj + low com r=.61	calculations of the test–retest coefficients somewhat unclear	A

				obj + hi com r=.50 (all 95% CIs sufficiently narrow)		

Excluded studies

1st author and year	Reason for exclusion
1 Kral (2021)	Concerns the performance measurement of public sector organisations.

Appendix 3b: Data extraction – primary studies*

1st author and year	Setting/Population	Description measurement tool	Construct or outcome measure	Reliability	Validity	Comments
1 Ali (2017)	hospitality industry / staff/tourists	Resort Hotel Service Performance (RESERVE) – 3 dimensions, 23 items – third party rating	3 dimensions of service performance: setting, audience, and actors	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant)	
2 Allen (2020)	medical (maternity) / obstetricians and gynaecologists	performance indicators developed by the Royal College of Obstetricians and Gynaecologists, – 14 items – direct report	Maternity service performance /quality of medical care	not reported	criterion validity	performance indicators did not correlate with inspection rating score
3 Amyx (2009)	newspaper publishing industry / sales agents	SALESPERF: scale measuring the service performance of sales representatives, adopted from SERVPERF scale, – 14 items – third party rating	Salesperson's service performance (includes reliability, responsiveness, assurance, empathy, and tangibles)	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant) criterion validity (concurrent, predictive)	
4 Barker (2011)	healthcare / nurses	evaluates nurses' perceptions of their performance – 9 items – self-report	Aspects of mental, physical and general performance during nurse work tasks	internal consistency (Cronbach's alpha)	content validity (expert panel)	

* All reliability and validity scores are available upon request

5 Brady (2002)	service, healthcare, entertainment, and fast food industry / customers	SERVPERF scale, (adopted from the SERVQUAL scale) – 4 dimensions, 28 items – third party rating	Consumer perceptions > service performance and expectations, service quality, satisfaction and purchase intentions	not reported	construct validity (convergent, discriminant)	performance-based measures of service quality (SERVPERF) represent a better operationalisation of the service quality construct than SERVQUAL
6 Carlos (2016)	higher education / lecturers	job performance measure – 29 items, – self-report	Task performance (knowledge, organisational skills, efficiency) contextual performance (persistent effort, relational skills, co-operation, conscientiousness)	internal consistency (Cronbach's alpha, composite)	content validity (expert panel) construct validity	
7 Cherniko va (2016)	retail (supermarket) / employees	supervisors' perception of employees' performance – 4 items – third party rating	Job performance (quality and quantity)	internal consistency (Cronbach's alpha)	not reported	
8 Darr (2017)	Canadian army / officers	computerised adaptive rating scales (CARS) – third party rating	Performance (based on five competencies: action orientation and initiative, behavioural flexibility/change management, teamwork, developing self and others, communication)	inter-rater reliability	criterion validity (measurement precision data were compared between BARS and CARS, with CARS being better)	CARS are similar to a behaviourally anchored rating scale (BARS) in that it contains specific performance-relevant behaviours of varying levels of effectiveness
9 DeArmo nd (2011)	construction industry / construction workers	safety performance measure, – 10 items, – self-report	Individual safety performance (safety participation, safety compliance)	internal consistency (Cronbach's alpha)	construct validity (convergent) criterion validity (concurrent)	
10 Dhammi ka (2012)	public sector / employees	performance measurement tool, (adopted from the Minnesota Satisfaction Questionnaire and O'Reilly and Chatman's organisational commitment measure) – 5 dimensions, 20 items, – self-report	Performance (job, career, innovation, team, organisation)	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant)	specific sample, limited generalisability
11 Greensla de (2017)	healthcare / nurses	job performance measure, – 8 dimensions, 41 items – self-report	Job performance (includes task and contextual performance)	internal consistency (Cronbach's alpha)	construct validity (convergent) criterion validity (concurrent)	

12 Hanif (2004)	higher education / lecturers	Teachers Perceived Job Performance Scale (TPJP) – 43 items – self-report	Perceived job performance, (includes task performance, contextual performance, and adaptive performance)	internal consistency (Cronbach's alpha)	content validity (expert panel) construct validity (convergent)	no data to support the convergent validity statement, unclear what measures were used besides the newly developed scale
13 Hatton (2009)	community-based housing services / managers, staff, service users, family members	job performance measure – 23 items (managers) – 26 items (staff) – 17 items (service users) – 24 items (family) – self-report and third party rating	Job performance	internal consistency (Cronbach's alpha, inter-item correlation) test–retest reliability	content validity (expert panel) construct validity (convergent, discriminant)	
14 Imel (2013)	healthcare / psychotherapists and patients	measure of patient–therapist alliance, – 3 items – third party rating	Patient–therapist alliance (=agreement on tasks, goals for treatment, and bond) as a measure for therapist performance	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant) criterion validity (concurrent)	
15 Kaplan (2009)	healthcare / physicians and patients	measure of physician performance, (adopted from existing scales NCQA/ADA DPRP) – 9 items	Physician performance, (=quality of care for diabetes)	inter-rater reliability internal consistency (Cronbach's alpha)	not reported	
16 Karayurt (2009)	healthcare / nurse students	measure of nursing performance, – 3 dimensions, 26 items – third party rating	Nursing performance, (included 'nursing process', 'professionalism' and 'ethical principles')	internal consistency (Cronbach's alpha, item-to-scale correlation)	content validity (expert panel) construct validity	concerns a student sample
17 Kinicki (2013)	international companies / managers	Performance Management Behaviour Questionnaire (PMBQ), – 6 dimensions, 27 items – third party rating and direct report	Performance management behaviour (=goal-setting, communication, feedback, coaching, and establishing/monitoring performance expectations)	internal consistency (Cronbach's alpha)	• content validity (expert panel) construct validity (convergent, discriminant) criterion validity (incremental validity)	Limitations: 1. student sample in Phase 3 analyses; 2. relatively high intercorrelations between PMBQ dimensions
18 Koopmans (2014-I)	mixed / mixed	Individual Work Performance Questionnaire (IWPQ), – 3 dimensions, 27 items – self-report	Individual work performance (task performance, contextual performance, and counterproductive work behaviour)	internal consistency (Cronbach's alpha, item-to-scale correlation)	not reported	Limitations: The IWPQ is not recommended for use in individual evaluations, assessments, and/or feedback
19 Koopmans (2014-II)	mixed / mixed	Individual Work Performance Questionnaire (IWPQ) – 3 dimensions, 18 items – self-report	Individual work performance (task performance, contextual performance, and counterproductive work behaviour)	not reported	construct validity (convergent, discriminant)	this study expands research on the IWPQ by examining its construct validity

20 Koopmans (2016)	healthcare / employees	Individual Work Performance Questionnaire (IWPQ) – 3 dimensions, 18 items – self-report	Individual work performance (task performance, contextual performance, and counterproductive work behaviour)	internal consistency (Cronbach's alpha, item-to-scale correlation)	content validity (expert panel, forward/back translation)	
21 Liu (2018)	transnational corporations / account managers	Key Account Manager Performance – various (?) items – self-report and third party rating	Key account manager performance, including sales performance; threats; promises; recommendations; information exchange; ligratiation; inspirational appeals; communication frequency; industry relational norms	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant)	
22 Lo (2005)	social work / sheltered workshop (social) workers	Work Performance Rating Scale (WPRS), – 14 items – n/a	Work performance: work accuracy, work speed, operational skills, initiative, work tolerance, co-operation, punctuality, appearance, social skills, emotional control, learning ability, attendance, work overtime	not reported	content validity (expert panel)	specific sample, limited generalisability
23 Lynch (1999)	retail / employees	Employee Performance Questionnaire, items were derived from previous studies and scales – 2 dimensions, multi-item measurement scale – third party rating	In-role and extra-role employee performance	internal consistency (Cronbach's alpha)	content and construct validity was demonstrated in previous studies; after factor analysis a two-factor solution was supported	
24 Molefe (2010)	higher education / university lecturers	Performance Measurement Dimension Questionnaire – 7 dimensions - self report	Lecturer performance: • knowledge (subject knowledge) • testing (assessment) procedures • student–teacher relations • organisational skills • communication skills • subject relevance	internal consistency (Cronbach's alpha)	not reported	Limitation: questionnaire measures lecturers' *perceived* performance

			• utility of assignments			
25 Na-Nan (2018-I)	SMEs in high-growth and high-impact sectors / entrepreneurs	Performance Management (PM) scale, – 5 dimensions, 33 items – self-report	Performance management: • prerequisites (understanding of organisation's vision, mission, strategy, goals) • performance planning • performance evaluation (evaluation and assessment based on standards/criteria during a set period of time) • performance review • performance application	internal consistency (Cronbach's alpha, composite)	content validity (expert panel) construct validity	specific sample, limited generalisability
26 Na-Nan (2018-II)	automotive industry / workers	Employee Job Performance (EJP) scale – 3 dimensions, 13 items – self-report	Job performance (job time, job quality, job quantity)	internal consistency (Cronbach's alpha, composite)	construct validity	specific sample, limited generalisability.
27 Onweze n (2014)	non-profit service industry / service workers	Job Performance Measure, adapted from the Task-based Job Performance Scale (Goodman and Svyantek 1999) – 9 items – self-report	Job performance	internal consistency (Cronbach's alpha)	not reported	
28 Poropat (2009)	higher education / students, non-academic supervisors and staff	Unifactorial Citizenship Performance (UCP) scale (adopted from Borman, 2001) – 3 dimensions, 12 items – self-report and third party rating	Citizenship performance (=personal support > the amount of help and co-operation provided to colleagues; organisational support > the degree to which people comply with rules and show loyalty to the organisation; conscientious initiative > the level of persistence and initiative demonstrated)	internal consistency (Cronbach's alpha) inter-rater reliability	construct validity (convergent, discriminant, nomological)	limitations: student sample in Study 1

29 Pulakos (2000)	military, federal government, state government, and private sector / employees from 24 different types of job	Job Adaptability Inventory (JAI) – 8 dimensions – 132 total items (15–18 items per dimension) – self-report	Adaptive performance	internal consistency (Cronbach's alpha)	construct validity (convergent, discriminant)	
30 Ramos (2019)	mixed / employees	Individual Work Performance Questionnaire (IWPQ, Koopmans) – 3 dimensions, 18 items, – self-report	Job performance (=task performance, contextual performance, and counterproductive behaviour)	internal consistency (Cronbach's alpha)	construct validity (convergent, nomological)	
31 Tsai (2015)	tourism and hospitality industry / employees	Creativity Scale – 13 items – self-report	Creative performance	internal consistency (Cronbach's alpha)	unclear, validity may be assessed in previous studies (Zhou and George 2001)	
32 Valenzuela (2020)	transportation industry / truck drivers	Safety Performance Measure – 6 subscales, 20 items – self-report	Safety performance	internal consistency (Cronbach's alpha)	content validity (expert panel) construct validity	
33 Van Hooft (2006)	public sector / employees	multi-source instrument to measure performance – 4 dimensions, 14 items – self-report and third party rating	Employee performance, including the following dimensions: • administrative skills • human skills • technical skills • other (eg customer orientation, stress tolerance, initiative, communication)	inter-rater reliability	construct validity (convergent, nomological) criterion validity (incremental)	
34 Wang (2004)	advertising industry / sales executives	creative performance scale – 7 items – self-report	Salesperson creative performance, conceptualised as the amount of new ideas generated or behaviours exhibited by the salesperson in performing their job activities	internal consistency (Cronbach's alpha)	construct validity (discriminant, nomological)	
35 Weng (2010)	healthcare / physicians (internal medicine)	ABIM Diabetes PIM – self-report	Physician performance = indicators of quality of care for diabetes	internal consistency (composite, intra-class) test–retest reliability	construct validity (convergent, nomological)	

36 Wright (2012)	healthcare / physicians	General Medical Council Patient Questionnaire (PQ) and Colleague Questionnaire (CQ), – 9 items and 18 items – third party rating	Physician performance (a mix of task performance and contextual performance)	internal consistency (Cronbach's alpha) inter-rater reliability test–retest reliability	construct validity (convergent)	

Excluded studies

1st author and year	Reason for exclusion
1 Ali-Rahmat (2010)	Tool measuring project performance in ISO-certified contractors, in terms of time and cost variances, level of complaints, clients' satisfaction, functionality, and health and safety. The only data is from a survey asking people for their satisfaction with measuring these components of project performance. No psychometric properties are reported.
2 Arns (2001)	Toolkit to assess performance in psychological rehabilitation programmes.
3 Bagnoli (2011)	Discusses a framework for assessing performance in social enterprises. No psychometric properties are presented; it's a discussion of one case study.
4 Bar-On (2018)	The tool assesses predictors of performance, not performance itself.
5 Bennet (2000)	Concerns the development of a tool to measure workplace deviance (defined as behaviour that violates organisational norms and, in so doing, threatens the wellbeing of the organisation or its members).
6 Bican (2020)	Tool focused on assessing R&D performance at company level.
7 Chan (2004)	KPIs to assess construction project success.
8 Chen (2005)	Tool to evaluate company performance in terms of knowledge management.
9 Cochenour (2000)	Editorial, not an empirical study.
10 Deadrick (2008)	Performance (typical and maximal) is measured through number of pieces produced by sewing machine operators.
11 Forth (2008)	Describes measures of organisational performance (productivity, profitability at company level).
12 Goyal (2019)	Did not measure performance, but the degree of relevance of three metrics in assessing channel partners' performance (also, not individual performance).
13 Hallowell (2020)	Deals with a metric of safety performance at company level (total recordable incident rate).
14 Hansen (2002)	Questionnaire measuring performance of (businesses participating at) trade shows.
15 Holmboe (2010)	Examines the reliability and validity of two types of composite medical performance scores – one for each specific disease condition (eg, diabetes, osteoarthritis) and a more comprehensive composite created by aggregating conditions by care type (ie, acute care, chronic care, and preventive services).
16 Khan (2012)	Insufficient information regarding test reliability and validity.
17 Kock (2017)	The objective of the study was to compare two ways of measuring job performance: self-perceptions and official supervisor evaluation – covered by Salgado (2019).
18 Lazzarotti (2011)	The authors propose a model for R&D performance management; however, this model has not been validated (not sufficient data).

19 Lee (2012)	Limited generalisability – the outcome variable is very specific (job performance of Australian expats in China).
20 Leppanen (2019)	The study aims to analyse the appropriateness of self-report as a measure of hand preference and unimanual object-based task performance; not related to job performance. Student sample.
21 Mielke (2019)	The study focuses on high-performance work system (HPWS), which seems a set of high-performance preconditions (skills, rewards, information, teamwork, workplace, appraisal, quality, job security, survey, candidate), rather than performance itself.
22 Neuenfeldt (2015)	Focuses on organisation performance (sectorial development of franchises in Brazil); the authors provide a sophisticated mathematical analysis, applying analytic hierarchy process (AHP) methodology, but it's not clear what data they are using.
23 Nowack (2007)	Insufficient information regarding test reliability and validity.
24 O'Connor (2015)	The study focuses on investigating the factors that influence the subjective performance measurement decision – covered by REA on performance appraisal.
25 O'Neill (2014)	Insufficient information regarding test reliability and validity.
26 Pan (2010)	The study's aim is to develop an index (measure based on a set of KPIs), not a questionnaire. No reliability or validity coefficients are provided.
27 Salgado (2015)	The focus of the study is on comparing coefficients: alpha, test–retest, and inter-rater correlations with coefficient of equivalence and stability (CES) rather than on the scale.
28 Sharma (2016)	The purpose of this paper is to explore the operationalisation of the construct 'employee perception of performance management system (PMS) effectiveness'.
29 Stewart (2007)	The study focuses on intra-individual variations in objective performance measure of professional football players (it's a sport context rather than work context).
30 Takala (2006)	The paper proposes a framework (process) to measure white-collar workforce performance, not a specific scale.
31 Tarescavage (2015)	Insufficient information regarding test reliability and validity.
32 Wang (2016)	Insufficient information regarding test reliability and validity.
33 Waterman (2014)	This paper discusses reliability of performance measures in general; no specific scale to measure performance was analysed.

CIPD

Chartered Institute of Personnel and Development
151 The Broadway London SW19 1JQ United Kingdom
T +44 (0)20 8612 6200 **F** +44 (0)20 8612 6201
E cipd@cipd.co.uk **W** cipd.co.uk

Incorporated by Royal Charter
Registered as a charity in England and Wales (1079797)
and Scotland (SC045154)

Issued: June 2022 Reference: 8239

CEBMa
center for
Evidence-Based Management