Measuring Individual Work Performance

Measuring Individual Work Performance

Linda Koopmans

Linda Koopmans

Body@Work

TNO
EMGO+ Institute for Health and Care Research
BODY@WORK

BODY@WORK

Measuring

Individual Work Performance

Linda Koopmans

The study presented in this thesis was performed at Body@Work, Research Center on Physical Activity, Work and Health. This research center is a joint initiative of the Netherlands Organisation of Applied Scientific Research TNO, Leiden, The Netherlands, and the VU University Medical Center (Department of Public and Occupational Health, EMGO+ Institute for Health and Care Research) and VU University, Amsterdam, The Netherlands.

The study was funded by the Netherlands Organisation of Applied Scientific Research TNO, Leiden, The Netherlands.

Financial support for the printing of this thesis has kindly been provided by Body@Work, Research Center on Physical Activity, Work and Health.

English title:	Measuring Individual Work Performance
Nederlandse titel:	Meten van Individuele Werkprestatie
ISBN:	978-94-6203-559-1
Cover picture by:	Jarrik Muller
Printed by:	CPI Koninklijke Wöhrmann, Zutphen

VRIJE UNIVERSITEIT

Measuring Individual Work Performance

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op woensdag 7 mei 2014 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

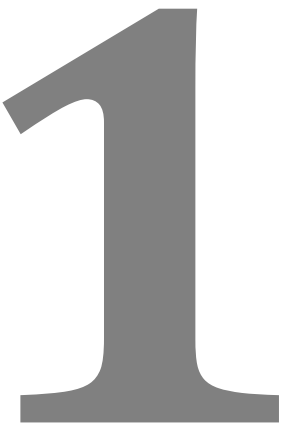
door
Linda Koopmans
geboren te Hieslum, gemeente Wonseradeel

promotoren:	prof.dr. A.J. van der Beek
	prof.dr.ir. H.C.W. de Vet
copromotoren:	dr. C.M. Bernaards
	dr. V.H. Hildebrandt

Contents

Chapter 1

General Introduction



Individual work performance is a hot issue. It plays an important role in our day to day workplace, in popular media, and in multiple fields within the scientific world, such as occupational health, work and organizational psychology, and management and economics. The following news headings and scientific articles give an indication of how individual work performance can be mentioned:

Popular media:

> *"Challenging individual work performance goals more effective than financial bonus", nu.nl news article* [1].
>
> *"Productivity of a smoker decreases the moment he or she wants to smoke a sigaret", television program De Rekenkamer* [2].
>
> *"Does 24/7 working on mobile increase overall productivity?", The Guardian* [3].

Scientific articles:

> *"When a happy worker is really a productive worker: A review and further refinement of the happy-productive worker thesis"* [4].
>
> *"Absenteeism or Presenteeism? Not at work or not working well"* [5].
>
> *"Freedom to surf: the positive effects of workplace Internet leisure browsing on employee productivity"* [6].

Trends in work

Why is individual work performance such a popular and interesting topic? Several trends in the labor market are responsible. First of all, individual work performance is important because of ongoing **globalization** of the economy. Economic globalization is the increasing economic integration and interdependence of national economies across the world through a rapid increase in cross-border movement of goods, service, technology, and capital [7]. As a result, competition between companies from all over the world increases. Therefore, it is essential for companies to maintain or improve their competitive ability. Individual work performance is one

of the key indicators for team and company performance, and consequently, it contributes to the productivity and competitive ability of companies.

Second, individual work performance is important in the light of the current **economic recession**. Recession refers broadly to "a period of reduced economic activity" [8]. It is characterized by a greater supply than demand in products, a drop in international trade, debts, bankruptcies, high unemployment rates, and low consumer confidence. Companies have to cut costs to stay afloat in these times of economic hardship, for example by employee reductions and/or outsourcing work to cheaper markets. Also, company employees will be encouraged to increase their individual work performance levels, in order to boost company performance and productivity. In addition, employees will tend to increase their individual work performance levels, in order to increase chances to survive possible reorganizations with reductions of personnel.

Third, individual work performance is important considering **sustainable employability**. The 'grey wave' in Europe means that there is an accelerated growth of the older working population and a decline in numbers of the younger working population. As a result, a scarcity of workers in the labor force threatens to arise [9]. Consequently, a shrinking number of workers will have to do the same - or an even larger - amount of work. Also, they will have to pay for the increasing costs associated with a growing elderly population, such as retirement pensions, social security, and health care [10]. It is therefore essential to improve the individual work performance of workers in the labor force. Also, the retirement age of workers will be increased in coming years, in order to prolong the stay of older workers in the work force [11]. This means that employees have to work at the same – or even higher – level of individual work performance, despite possible limitations caused by an older age, such as reduced health.

Maintaining, improving, and optimizing individual work performance

Due to the above trends in work, it becomes increasingly important to maintain, improve, and optimize the individual work performance of employees. In order to do so, various approaches and solutions have been proposed by different scientific fields.

The field of **occupational health** has focused primarily on the relation between health complaints and losses in individual work performance due to sickness absenteeism or presenteeism, and how to prevent productivity loss due to a certain disease or health impairment. Within this field, numerous studies have

been done aiming to maintain or improve an employee's performance levels by targeting their health. These interventions studies address, for example, working conditions [e.g., 12-14], ergonomics at work [e.g., 15-17], safety at work [e.g., 18-20], and physical activity and/or healthy nutrition [e.g., 21-23]. In addition, pre-employment examinations [e.g., 24-26], and health risk appraisals [e.g., 27, 28] are performed in order to prevent occupational injuries, diseases and sickness absence.

The field of **work and organizational psychology** has traditionally been interested in determinants, such as employee ability, motivation, and resources, in order to understand, predict, and improve individual work performance [29]. Work and organizational psychologists have long been involved in hiring and recruiting personnel, using for example interviews, collection of biographical data, and knowledge and personality tests, in order to select the most appropriate candidate for the job [30]. Assessment procedures were developed to evaluate success on the job and identify improvements needed to optimize individual work performance [31]. Also, training and development programs are designed to teach knowledge, skills, and abilities needed to improve individual work performance [e.g., 32].

The field of **management and economics** has primarily been occupied with optimizing the individual work performance of employees, with the ultimate goal of optimizing the company's productivity and competitive ability. Contrary to work and organizational psychology, which generally focuses on the individual, the field of management and economics has a larger focus on the entire work system, including factors such as work processes, technological constraints, and organizational structure [33]. Tools and strategies for optimizing individual work performance include for example the Balanced Scorecard [34], total quality management [33], high performance work systems [35], and pay for performance [36].

Defining and measuring individual work performance

Within the field of **occupational health**, there is a focus on instruments that measure losses in individual work performance due to health complaints [e.g., 37]. These include for example the Stanford Presenteeism Scale [38], Work Productivity and Activity Impairment questionnaire [39], and the Health and Performance Questionnaire [40]. A loss in individual work performance due to health complaints is usually expressed as absenteeism (absence from work due to health complaints) or presenteeism (being present at work but ill). Absenteeism or presenteeism measures are then taken as proxies for losses in individual work performance. The question is, however, whether the equation of a loss in individual work performance

with absenteeism or presenteeism is just. As Johns [41] pointed out in his review, a loss in individual work performance is a result of being absent, or of being present at work while ill, rather than the same thing. Also, within the field of occupational health, the terms of individual work performance and productivity are often used interchangeably. This is perhaps driven by the goal to relate losses in performance or productivity to costs, as productivity usually refers to objective output. The unclear definitions and content of the terms individual work performance, productivity, absenteeism, and presenteeism – and their interchangeable use – have resulted in a multitude of instruments with heterogeneous content. Also, many of these instruments are specifically developed for workers with health complaints. When applied to workers without health complaints, a pronounced ceiling effects in these scales is created. Therefore, these measures are unsuitable for a general, mostly healthy, working population.

Within the field of **work and organizational psychology**, defining the construct of individual work performance, and attempting to understand its underlying structure, has received much attention [42]. Traditionally, the focus has been on task performance, which can be defined as the proficiency with which individuals perform the core substantive or technical tasks central to his or her job [43]. More recently, there has been an increasing interest in discretionary, positive work behaviors that indirectly contribute to the goals of the organization. Various labels have been used for this type of behavior, such as organizational citizenship behavior, extra-role behavior, and contextual performance [e.g., 44]. Also, counterproductive work behaviors that harm the well-being of the organization have received attention [e.g., 45]. Numerous scales have also been developed to measure task performance [e.g., 46], contextual performance [e.g., 47], or counterproductive work behavior [e.g., 48]. The multitude of scales in this research field is perhaps best illustrated by LePine, Erez and Johnson [44], who identified more than 40 different measures of contextual performance. However, heterogeneous content between scales measuring the same dimension, and overlapping content between scales measuring a different dimension, can be observed. Also, the scales are often developed based on a specific type of occupation, making these scales less suitable for use in a generic working population.

Within the field of **management and economics**, performance measurement mostly focuses on the company level, using key performance indicators such as employee turnover, customer satisfaction, and financial performance. Especially in the United States of America, performance measurement

has taken off since President Clinton signed the National Performance Review Act in 1993. Within companies, human resource management has developed performance measurement systems to determine individual work performance. Thus, these performance measurement systems are often specific to the job or company. Also, human resource managers have drawn on research from work and organizational psychology to measure individual work performance [49].

A lack of consensus

Despite its importance and popularity, there is little consensus on how to define and measure individual work performance. It follows naturally that a multitude of instruments exists to measure individual work performance or one of its related constructs. When considering the research on individual work performance from the different research fields, it seems evident that a clear definition and conceptual framework of individual work performance is lacking. This seems to be true not only between research fields, but also within research fields. Various terms are used to refer to individual work performance, such as productivity, absenteeism, and presenteeism. Often, their exact definition and content is unclear, and the terms are used interchangeably. This raises the question of how individual work performance should be defined. And what exactly constitutes individual work performance? Is it a single, unidimensional construct, or is it made up of multiple components, or dimensions? The lack of consensus on how to define and conceptualize individual work performance is undesirable, because a clear definition and conceptual framework of individual work performance is a prerequisite for its valid measurement.

In accordance with the lack of consensus of the definition and conceptualization of individual work performance, a multitude of instruments exists to measure individual work performance or one of its related constructs. Three concerns arise when considering the numerous instruments to measure individual work performance developed by the different research fields. First, do the measures capture the complete range of individual work performance? Second, do they include the right content? And third, are they suitable in a generic, mostly healthy, working population? The lack of consensus on how to measure individual work performance is undesirable, because valid measurement is a prerequisite for accurately establishing the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize individual work performance.

Objective of this thesis

The lack of consensus on how to define, conceptualize and measure individual work performance impedes valid measurement of the construct. As a result, it remains difficult to establish the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize individual work performance. Research would benefit greatly from a comprehensive, generic, and short instrument to measure individual work performance. As Lord Kelvin (1883) said: "*Measurement is knowledge*". Therefore, the main objective of this thesis was to develop and validate a comprehensive, generic, and short questionnaire to measure individual work performance.

Outline of this thesis

Part I of this thesis describes the developmental phase of the Individual Work Performance Questionnaire. First, Chapter 2 presents a systematic review of the literature on conceptual frameworks of individual work performance. In this chapter, a definition of individual work performance is given, and an integrated, conceptual framework is proposed, which is used as the starting point for the development of the Individual Work Performance Questionnaire (IWPQ). As existing knowledge was insufficient to operationalize the dimensions in this conceptual framework, in Chapter 3, numerous indicators used to measure individual work performance are identified via the scientific literature, existing questionnaires, and expert interviews. Subsequently, the most relevant indicators per dimension are selected by experts from different professional backgrounds.

Part II of this thesis describes the field-testing phase of the Individual Work Performance Questionnaire. In Chapter 4, the development, pilot-testing, and field-testing of the Individual Work Performance Questionnaire is described. For each dimension of individual work performance, a generic, short scale was constructed. In Chapter 5, the improvements that were made to obtain optimal targeting of the IWPQ are described. Here, the final version of the questionnaire – the IWPQ 1.0 – is presented. In this chapter, the calculation of sumscores and the interpretability of scores are also discussed.

Part III of this thesis describes the validation of the IWPQ 1.0. In Chapter 6, two types of construct validity of the IWPQ are described. First, the relations of the IWPQ with presenteeism and work engagement are examined (convergent validity). Second, it is examined whether workers low and high in job satisfaction, and workers low and high overall health, can be discriminated on IWPQ scores (discriminative

validity). In Chapter 7, the responsiveness of the IWPQ is examined in a randomized controlled trial (RCT) on physical activity and relaxation in the workplace. In Chapter 8, the cross-cultural validation of the IWPQ to the American-English language is presented. The thesis concludes with a general discussion in Chapter 9.

References

1. Nu.nl. Challenging performance goals more effective than financial bonus. www.nu.nl. 2010 Nov 19. Available from: http://www.nu.nl/werk-en-prive/2382835/uitdagende-prestatiedoelen-effectiever-dan-bonus.html.
2. KRO, De Rekenkamer. Wat kost een rokende werknemer? Hilversum: De Rekenkamer; 2011.
3. Robbins B. Does 24/7 working on mobile increase overall productivity? The Guardian. 2013 Jan 30.
4. Cropanzano R, Wright TA. When a "happy" worker is really a "productive" worker: A review and further refinement of the happy-productive worker thesis. Consulting Psychology Journal: Practice and Research. 2001;53(3):182-99.
5. Johns G. Absenteeism and presenteeism: Not at work or not working well. In: Cooper CL, Barling J, editors. The Sage Handbook of Organizational Behavior. 1st ed. London, UK: Sage; 2008. p. 160-77.
6. Coker BLS. Freedom to surf: The positive effects of workplace internet leisure browsing. New Technology, Work and Employment. 2011;26(3):238-47.
7. Joshi RM. International business. Oxford University Press, New Delhi and New York; 2009.
8. Merriam-Webster collegiate dictionary online. "Headword "recession". Available from: http://www.merriam-webster.com/dictionary/recession (accessed 18-11-2013).
9. United Nations. The Madrid international plan of action on ageing: Guiding framework and toolkit for practitioners and policy makers. New York: Department of Economic & Social Affairs Division for Social Policy & Development; 2008.
10. World Health Organization. The European health report 2009: Health and health systems. Copenhagen, Denmark: WHO Regional Office for Europe; 2009.
11. European Commission. EUROPE 2020: A strategy for smart, sustainable and inclusive growth. Brussels: European Commission; 2010.
12. Holden L, Scuffham PA, Hilton MF, Vecchi NN, Whiteford HA. Work performance decrements are associated with australian working conditions, particularly the demand to work longer hours. Journal of Occupational and Environmental Medicine. 2010;52(3):281-90.

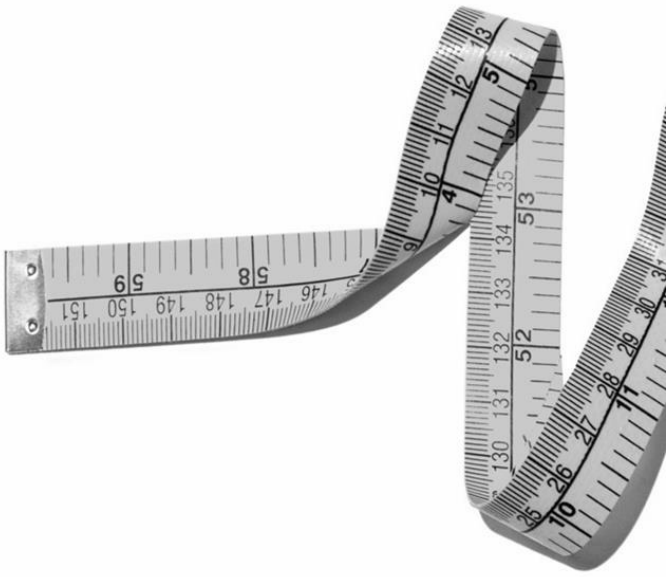
13. Shimizu M, Wada K, Wang G, Kawashima M, Yoshino Y, Sakaguchi H, et al. Factors of working conditions and prolonged fatigue among teachers at public elementary and junior high schools. Industrial health. 2011;49(4):434-42.
14. Srinivasan D, Mathiassen SE. Motor variability in occupational health and performance. Clinical Biomechanics. 2012;27(10):979-93.
15. Leyshon R, Chalova K, Gerson L, Savtchenko A, Zakrzewski R, Howie A, et al. Ergonomic interventions for office workers with musculoskeletal disorders: A systematic review. Work: Journal of Prevention, Assessment & Rehabilitation. 2010;35(3):335-48.
16. Tompa E, Dolinschi R, de Oliveira C, Amick BC3, Irvin E. A systematic review of workplace ergonomic interventions with economic analyses. Journal of Occupational Rehabilitation. 2010;20(2):220-34.
17. Westgaard RH, Winkel J. Occupational musculoskeletal and mental health: Significance of rationalization and opportunities to create sustainable production systems - A systematic review. Applied Ergonomics. 2011;42(2):261-96.
18. Zacharatos A, Barling J, Iverson RD. High-performance work systems and occupational safety. Journal of Applied Psychology. 2005;90(1):77-93.
19. Robson LS, Stephenson CM, Schulte PA, Amick BC3, Irvin EL, Eggerth DE, et al. A systematic review of the effectiveness of occupational health and safety training. Scandinavian Journal of Work, Environment and Health. 2012;38(3):193-208.
20. Christian MS, Bradley JC, Wallace JC, Burke MJ. Workplace safety: A meta-analysis of the roles of person and situation factors. Journal of Applied Psychology. 2009;94(5):1103-27.
21. Rongen A, Robroek SJ, van Lenthe FJ, Burdorf A. Workplace health promotion: A meta-analysis of effectiveness. American Journal of Preventive Medicine. 2013;44(4):406-15.
22. Jensen JD. Can worksite nutritional interventions improve productivity and firm profitability? A literature review. Perspectives in Public Health. 2011;131(4):184-92.
23. Maes L, Van Cauwenberghe E, Van Lippevelde W, Spittaels H, De Pauw E, Oppert JM, et al. Effectiveness of workplace interventions in europe promoting healthy eating: A systematic review. European journal of public health. 2012;22(5):677-83.

24. Pachman J. Evidence base for pre-employment medical screening.. Bulletin of the World Health Organization. 2009;87(7):529-34.
25. Mahmud N, Schonstein E, Schaafsma F, Lehtola MM, Fassier JB, Reneman MF, et al. Pre-employment examinations for preventing occupational injury and disease in workers. The Cochrane Database of Systematic Reviews. 2010;12.
26. Fenner P. The pre-employment medical - nuisance or great opportunity?. Australian Family Physician. 2011;40(7):541-4.
27. Anderson DR, Staufacker MJ. The impact of worksite-based health risk appraisal on health-related outcomes: A review of the literature. American Journal of Health Promotion. 1996;10(6):499-508.
28. Maron DJ, Forbes BL, Groves JR, Dietrich MS, Sells P, DiGenio AG. Health-risk appraisal with or without disease management for worksite cardiovascular risk reduction. Journal of Cardiovascular Nursing. 2008;23(6):513-8.
29. Waldman DA, Spangler WD. Putting together the pieces: A closer look at the determinants of job performance. Human Performance. 1989;2(1):29-59.
30. Schmitt N. The oxford handbook of personnel assessment and selection. 1st ed. Schmitt N, editor. New York, NY, US: Oxford University Press; 2012.
31. Guion RM. Assessment, measurement, and prediction for personnel decisions. 2nd ed. Guion RM, editor. New York, NY, US: Routledge; 2011.
32. Aswathappa K. Human resource and personnel management. 4th ed. Aswathappa K, editor. New Delhi: Tata McGraw-Hill; 2005.
33. Waldman DA. The contributions of total quality management to a theory of work performance. Academy of Management Review. 1994;19(3):510-36.
34. Kaplan RS, Norton DP. The balanced scorecard: Measures that drive performance. Harvard Business Review. 1992;Jan - Feb:71-80.
35. Lawler ED, Morhrman SA, Ledford GE. Creating high performance organizations. Lawler ED, Morhrman SA, Ledford GE, editors. San Francisco: Jossey-Bass; 1995.
36. Durham CC, Bartol KM. Pay for performance. In: Locke EA, editor. Handbook of Principles of Organizational Behavior. 2nd ed. West Sussex, UK: John Wiley & Sons, Ltd; 2009. p. 217-38.
37. Mattke S, Balakrishnan A, Bergamo G, Newberry SJ. A review of methods to measure health-related productivity loss. The American Journal of Managed Care. 2007;13(4):211-7.

38. Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: Health status and employee productivity. Journal of Occupational and Environmental Medicine. 2002;44:14-20.
39. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproductibility of a work productivity and activity impairment instrument. Pharmacoeconomics. 1993;4:353-65.
40. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The world health organization health and work performance questionnaire (HPQ). Journal of Occupational and Environmental Medicine. 2003;45:156-74.
41. Johns G. Presenteeism in the workplace: A review and research agenda. Journal of Organizational Behavior. 2010;31(4):519-42.
42. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol. 2005;90:1241-55.
43. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, eds. Handbook of industrial and organizational psychology. Palo Alto, CA, US: Consulting Psychologists Press. 1990:687-732.
44. LePine JA, Erez A, Johnson DE. The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. J Appl Psychol. 2002;87(1):52-65.
45. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. J Appl Psychol. 2002;87(1):66-80.
46. Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. Journal of Management. 1991;17(3):601-17.
47. Podsakoff PM, MacKenzie SB. A second generation measure of organizational citizenship behavior. Indiana University, Bloomington; 1989.
48. Bennett RJ, Robinson SL. Development of a measure of workplace deviance. Journal of Applied Psychology. 2000;85(3):349-60.
49. Den Hartog DN, Boselie P, Paauwe J. Performance management: A model and research agenda. Applied Psychology. 2004;53(4):556-69.

Part I

Developmental Phase



Chapter 2

Conceptual Frameworks of Individual Work Performance - A Systematic Review

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt,
Wilmar B. Schaufeli, Henrica C.W. de Vet, Allard J. van der Beek

Journal of Occupational and Environmental Medicine. 2011; 53(8):856-866

DOI: 10.1097/JOM.0b013e318226a763

2

Abstract

Objective: Individual work performance is differently conceptualized and operationalized in different disciplines. Aim of the current review was twofold: 1) identifying conceptual frameworks of individual work performance, 2) integrating these in order to reach a heuristic conceptual framework.

Methods: A systematic review was conducted in medical, psychological and management databases. Studies were selected independently by two researchers, and included when they presented a conceptual framework of individual work performance.

Results: 17 Generic frameworks (applying across occupations) and 19 job-specific frameworks (applying to specific occupations) were identified. Dimensions frequently used to describe individual work performance were task performance, contextual performance, counterproductive work behavior, and adaptive performance.

Conclusion: Based on the literature, a heuristic conceptual framework of individual work performance was proposed. This framework can serve as a theoretical basis for future research and practice

Introduction

Individual work performance is an issue that not only has grasped companies all over the world, but also has fueled a great deal of research in fields of management, occupational health and work and organizational psychology [1-4]. Numerous studies on individual work performance have been conducted. However, different approaches of studying individual work performance circulate in today's literature. Whereas the field of management has primarily occupied itself with how one can make an employee as productive as possible, the field of occupational health has focused on how to prevent productivity loss due to a certain disease or health impairment [5;6]. Work and organizational psychologists, on the other hand, have an interest in the influence of determinants, such as work engagement, satisfaction, and personality, on individual work performance [7-9].

In all of the above mentioned research fields, individual work performance is a relevant outcome measure of studies in the occupational setting. However, despite its importance, no comprehensive conceptual framework of individual work performance exists. A solid theoretical framework is a prerequisite for optimal measurement of the construct [4]. It has typically been assumed that what constitutes individual work performance differs from job to job. As a result, countless measures of work performance have been used [10]. So far, the assessment of individual work performance has primarily focused either on objective measures of work productivity (such as number of days absent, counts of specified acts, or output maintained in organizational records) or on subjective judgments of quantity and quality of work from the employee him- or herself, peers or supervisors [11;12]. While these methods may provide valuable information, it can be argued that none of them capture the complexity and full range of behaviors that constitute an employee's performance at work [4;13].

This raises the question of what exactly constitutes individual work performance. Work performance is an abstract, latent construct that cannot be pointed to or measured directly [14]. It is made up of multiple components, or *dimensions*. These dimensions, in turn, are made up of indicators that can be measured directly. In order to conceptualize and operationalize individual work performance, we should explicate the construct domain of work performance and identify its dimensions and indicators [4;14;15]. Whereas the dimensions may generalize across jobs, the exact indicators can differ between jobs [14]. In the field of psychology, the conceptualization of work performance has received relatively

much attention. A widely endorsed definition of work performance is that of Campbell: “behaviors or actions that are relevant to the goals of the organization” [4]. Three notions accompany this definition: 1) work performance should be defined in terms of behavior rather than results, 2) work performance includes only those behaviors that are relevant to the organization’s goals, and 3) work performance is multidimensional. As distinguishing between behavior and results can be difficult, others have included results in their definition of work performance. For example, Viswesvaran and Ones [11]. defined work performance as: “scalable actions, behavior and outcomes that employees engage in or bring about that are linked with and contribute to organizational goals.”

Work performance should be distinguished from work productivity, two concepts that often seem to be used interchangeably in the literature. Work productivity is defined as input divided by output [12]. Thus, work productivity is a narrower concept than work performance. It is also important to distinguish between causal variables and indicators of work performance. Causal variables determine or predict one’s level of work performance, whereas indicators are reflections of work performance [16]. For example, job satisfaction is considered a determinant of work performance [8], whereas work quality is an indicator of work performance [4]. The current review only focuses on indicators of work performance and not on its determinants.

Thus, until now, no clear consensus exists on what exactly constitutes individual work performance. Aim of the current review was twofold: 1) identifying conceptual frameworks of individual work performance, and 2) integrating the conceptual frameworks in order to reach a heuristic conceptual framework of individual work performance.

Methods

Search strategy

A systematic search was conducted to identify frameworks describing the construct of individual work performance. The primary searches were conducted April/May 2010 in two medical databases (PubMed and Embase.com), one psychological (PsycINFO) and one management (ABI Inform) database. The search was restricted to literature written in English or Dutch. No restrictions were placed on year of publication or publication type. All search strategies were developed with the aid of experienced search specialists. Search strategies are presented in Table 1. Additional

studies were identified by scanning the reference lists of suitable studies and in personal collections.

Study selection

The first reviewer determined the eligibility of studies on the basis of title and abstract. Studies that presented a framework describing the construct of individual work performance were included. Exclusion criteria were: 1) not on work performance, 2) not at the individual level, or 3) not on a framework describing the construct of individual work performance. A second reviewer independently determined the eligibility of the studies that the first reviewer found suitable or doubtful. In addition, the second reviewer determined the eligibility of 100 random studies per database, thus, a total of 400 random studies, to get an indication of whether the first reviewer had missed relevant studies. Differences in judgment were resolved through a consensus procedure. Finally, the first reviewer determined eligibility of all suitable or doubtful studies based on full text.

Table 1. Search strategies

Database	*Search strategy*
PubMed	("employee performance appraisal"[Mesh] OR "task performance and analysis"[Mesh] OR "efficiency"[Mesh] OR "absenteeism"[Mesh] OR "sick leave"[Mesh] OR "performance"[Title/Abstract] OR "productivity"[Title/Abstract] OR "absenteeism"[Title/Abstract] OR "presenteeism"[Title/Abstract]) AND ("work"[Mesh] OR "workplace"[Mesh] OR "employment"[Mesh] OR "occupations"[Mesh])AND ("Models, Nursing"[Mesh] OR "Models, Theoretical"[Mesh] OR "Models, Economic"[Mesh] OR "Models, Psychological"[Mesh] OR "Models, Organizational"[Mesh] OR "model*"[Title/Abstract] OR "theor*"[Title/Abstract]) AND (Humans[Mesh])
Embase.com	('job performance':cl,ab,ti OR 'task performance':cl,ab,ti OR 'productivity':cl,ab,ti OR 'absenteeism':cl,ab,ti OR 'medical leave':cl,ab,ti OR 'presenteeism':ab,ti) AND ('work':cl,ab,ti OR 'occupation':cl,ab,ti OR 'employee':cl,ab,ti OR 'job':cl,ab,ti) AND ('model':cl,ab,ti OR 'theory':cl,ab,ti OR 'conceptual framework':cl,ab,ti) AND [humans]/lim
PsycINFO	(("job performance" OR "employee productivity" OR "occupational success" OR "employee absenteeism" OR “presenteeism” OR "sick leave") AND ("models" OR "theories" OR "model" OR "theory")).ti,ab,id.
ABI Inform	LSU(job performance) OR LSU(performance appraisal) OR LSU(organizational behavior) OR LSU(employee attitude) OR

	LSU(performance management) OR LSU(performance evaluation) AND TI OR ABS("job performance" OR "performance appraisal" OR "employee performance" OR "work productivity" OR "absenteeism" OR "presenteeism") AND (LSU(models) OR LSU(theory) OR TI("model") OR TI("theory") OR ABS("model") OR ABS ("theory")) AND LSU(individual) OR ABS("individual") OR TI("individual")

Results

Study selection

The searches in PubMed, Embase.com, PsycINFO and ABI Inform resulted in 760, 553, 1328, and 478 hits, respectively. From this total of 3119 hits, 252 duplicates were removed, resulting in 2867 studies to be screened. Of these, 107 studies were found eligible based on title and abstract. Consensus rates between the two reviewers were 75% for PubMed, 79% for Embase.com, 84% for PsycInfo, and 68% for ABI Inform. Differences in judgment were resolved through a consensus procedure, resulting in full agreement. The full papers of the 107 eligible studies were screened. These included 65 articles, 11 book chapters and 31 dissertations. Of these 107 results, 24 dissertations could not be retrieved in full text. They could therefore not be judged on their eligibility and were excluded from the review. Out of the 83 remaining studies, 49 were included in the review, based on full text. Seven of the included studies were identified in PubMed, one in Embase.com, 33 in PsycInfo and 8 in ABI Inform. With 9 articles found in additional searches of reference lists and of TNO and VU University medical centre literature databases, a total of 58 studies were included in this review. Forty of these were articles, 12 were book chapters and 6 were dissertations. Figure 1 presents a flowchart of the study selection process. In short, 921 of the studies were excluded because they did not report on work performance, 119 because they did not report on work performance at the individual level, and 1754 because they did not present a framework describing the construct of individual work performance.

General description of the studies

Fifty-eight studies were identified that presented a conceptual framework of individual work performance. In 35 of the 58 studies an original conceptual framework was presented. The remaining 23 studies did not present a new conceptual framework, but referred to one of the other original conceptual frameworks. Seventeen generic frameworks (i.e. applying to work performance

across all occupations) and 18 job-specific frameworks (i.e. applying to work performance in a specific occupation) were identified. Table 2 presents an overview of identified conceptual frameworks. A description of some often presented frameworks follows.

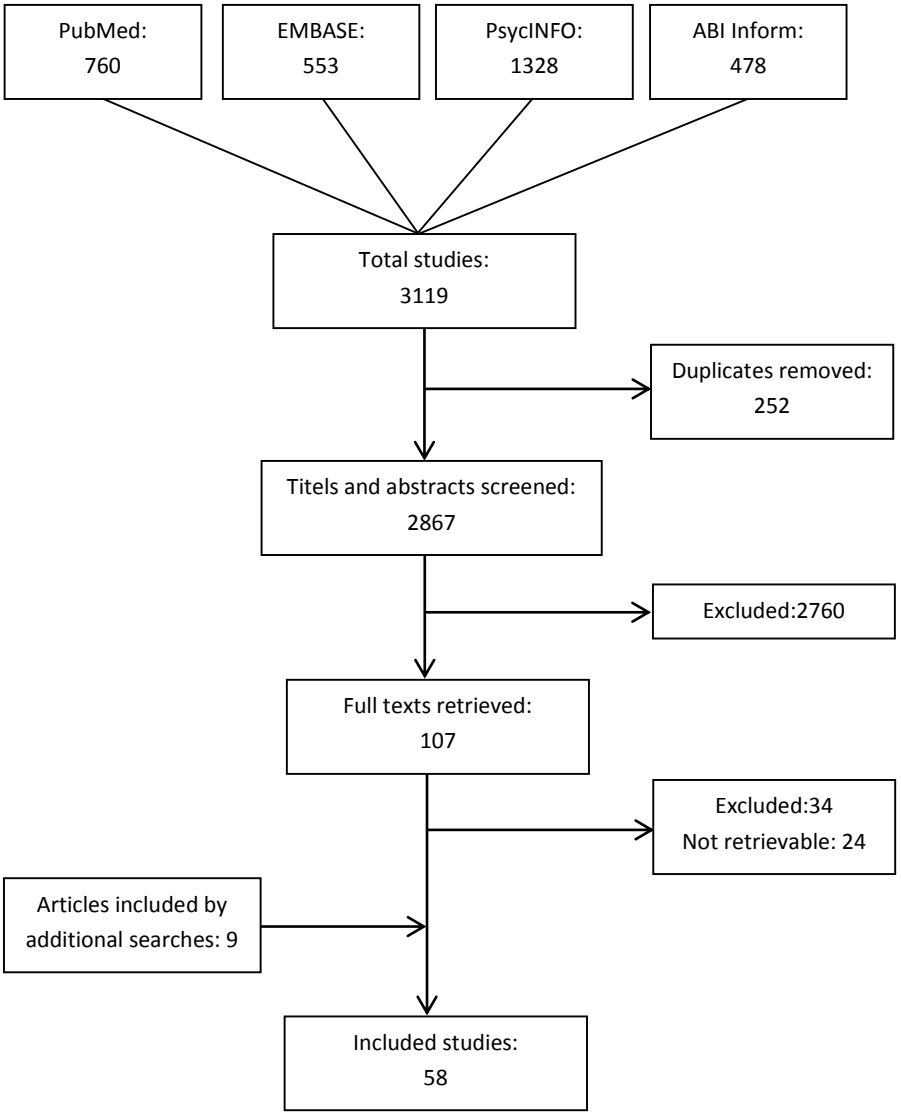


Figure 1. Flowchart of study selection process

Generic frameworks

Murphy [17] and Campbell [4] were among the first to define the domain of individual work performance by specifying the major dimensions of generic work performance. According to Murphy, the work performance domain could be modeled using the following four dimensions: 1) *task behaviors*, 2) *interpersonal behaviors* (communicating and cooperating with others), 3) *down-time behaviors* (work-avoidance behaviors) and 4) *destructive/hazardous behaviors* (behaviors that lead to a clear risk of productivity losses, damage, or other setbacks). Campbell's work performance framework proposed eight work performance dimensions: 1) *job-specific task proficiency*, 2) *non-job-specific task proficiency*, 3) *written and oral communication*, 4) *demonstrating effort*, 5) *maintaining personal discipline*, 6) *facilitating peer and team performance*, 7) *supervision*, and 8) *management and administration*. According to Campbell, these eight dimensions are sufficient to describe the latent structure of performance at a general level. However, he also noted that the eight factors can have different patterns of subdimensions, and their content and salience can vary across jobs.

On the basis of the conceptual grouping of 486 measures of work performance found in the literature, Viswesvaran developed 10 dimensions of individual work performance [18]. Besides a general factor of overall job performance, he distinguished the dimensions of *productivity, quality of work, job knowledge, communication competence, effort, leadership, administrative competence, interpersonal competence,* and *compliance with/acceptance of authority*.

Borman & Motowidlo argued that the entire work performance domain could be encompassed by the comprehensive dimensions of *task performance* and *contextual performance* [19]. They describe task performance as behaviors that directly or indirectly contribute to the organization's technical core, and contextual performance as behaviors that support the organizational, social and psychological environment in which the technical core must function. Examples of contextual activities are volunteering, persisting, helping, cooperating and following rules [19]. Task activities usually vary between different jobs, whereas contextual activities are common to many or all jobs.

In the early 2000s, Viswesvaran and Ones [11] and Rotundo and Sackett [20] conducted two narrative reviews on frameworks of individual work performance. Both reviews concluded that three broad dimensions of work performance could be distinguished: task performance, organizational citizenship behavior, and

counterproductive work behavior. The term 'organizational citizenship behavior' was first introduced by Organ [21], and is currently defined as individual behavior that contributes to the maintenance and enhancement of the social and psychological context that supports task performance [22]. Although originally there were some definitional differences between organizational citizenship behavior and contextual performance, Organ's definition of organizational citizenship behavior has evolved to greatly overlap with Borman & Motowidlo's definition of contextual performance [22]. In the current review, the term contextual performance will be used to refer to behaviors that support the organizational, social or psychological environment in which the technical core functions. The third dimension, counterproductive work behavior, was defined as behavior that harms the well-being of the organization [20]. It includes behaviors such as absenteeism, off-task behavior, theft, and substance abuse.

Job-specific frameworks

Frameworks developed for specific jobs were mainly targeted at professions in the army, managers, or sales and service industry. In 1990, Campbell, McHenry and Wise [23] developed a framework in which work performance in the army was described by five dimensions: 1) *core technical proficiency*, 2) *general soldiering proficiency*, 3) *effort and leadership*, 4) *personal discipline*, and 5) *physical fitness and military bearing*. The last referred to the degree to which individuals stay in good physical condition, maintain appropriate military appearance, and carry or conduct oneself appropriately. Campbell's more comprehensive 8-dimensional framework [4] is largely based on this framework. Borman and Brush [24] developed a framework, based on critical incidents analysis, in which managerial work performance was described by 1) *technical activities and mechanisms of management*, 2) *interpersonal dealings and communication*, 3) *leadership and supervision*, and 4) *useful personal behavior and skills* (e.g., persistence, handling crises and stress, organisational commitment). This framework was developed independent of Borman & Motowidlo's 2-dimensional framework [19]. Maxham et al. [25] described performance of retail employees as 1) *in-role performance*, 2) *extra-role performance towards customers*, and 3) *extra-role performance towards the organisation*.

Similarities between frameworks

Generic frameworks used more broad dimensions to describe work performance, whereas job-specific frameworks used more narrow dimensions to describe elements of work performance. Despite these different levels of specificity, similarities were observed between dimensions of individual work performance described in the frameworks. Based on conceptual grouping of individual work performance dimensions found in the literature, three broad dimensions could be distinguished: task performance, contextual performance, and counterproductive work behavior. Finally, some frameworks described dimensions which they did not classify in one of these three categories, such as proactive, creative and adaptive performance. Table 2 shows the classification of dimensions from each framework. The original dimensions of the frameworks were classified in the scheme of Table 2 based on the definitions of the dimensions as provided by the developers of the framework. When the original dimensions were not defined, they were classified based on the authors' own insight.

Task performance

Almost all frameworks mentioned task performance as an important dimension of individual work performance. Task performance can be defined as the proficiency (i.e. competency) with which one performs central job tasks [4]. Other labels sometimes used for task performance are *job-specific task proficiency* [13;26-28], *technical proficiency* [23;29;30], or *in-role performance* [25;31]. It includes for example work quantity, work quality, and job knowledge [4].

In Murphy's (1989) framework, the first dimension, labeled *task behaviors*, could be considered task performance. Campbell (1990) himself stated that his first two dimensions, *job-specific task proficiency* (core job tasks) and *non-job-specific task proficiency* (tasks not specific to a given job, but expected of all employees), represent task performance [11]. Viswesvaran's [18] first three dimensions, *productivity, quality* and *job knowledge*, could be considered task performance. Later developed individual work performance frameworks all included one dimension to describe task performance [11;13;19;20;31-34]. The only exception was Renn and Fedor's framework, in which task performance was split into *work quantity* and *quality* [35].

Of course, what constitutes core job tasks can differ from job to job. In contrast to generic frameworks, job-specific frameworks often used multiple, specific dimensions to describe task performance. For example, Arvey and Mussio

[36] described task performance of clerical workers using the dimensions of *working accurately*, *showing concern for time* and *detail and planning*. Jiambalvo [37] described task performance for public accountants as *understanding*, *planning* and *revising work*. Engelbrecht and Fischer [38] divided task performance for managers into *action orientation* (e.g., getting things done, decisiveness), *task structuring* (e.g., leadership, planning), and *probing, synthesis and judgment* (problem resolution). Furthermore, Tett et al. [39] divided task performance for managers into *traditional functions* (e.g., decision making, planning), and *occupational acumen and concerns* (e.g., job knowledge, concern for quantity and quality).

Contextual performance

Although task performance has been the traditional focus of research, researchers have come to believe that individual work performance is more than meeting prescribed work goals [11;19]. In both generic and job-specific frameworks, one or more dimensions of contextual performance have been included. Contextual performance can be defined as individual behaviors that support the organizational, social and psychological environment in which the technical core must function [19]. Several labels exist for this dimension, such as *non-job-specific task proficiency* [26;28], *extra-role performance* [25;31], *organizational citizenship behavior* [11;20;34] or *interpersonal relations* [17]. All concepts, however, refer to behaviors that go beyond the formally prescribed work goals, such as taking on extra tasks, showing initiative, or coaching newcomers on the job.

Seven of the generic frameworks used one broad dimension to describe contextual performance. Four generic frameworks used multiple dimensions to describe contextual performance. For example, in Campbell's framework [4], 6 of the 8 dimensions (*written and oral communication, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision and leadership,* and *management and administration*) could be regarded contextual performance. Also, 6 of Viswesvaran's dimensions [18] [*communication competence, effort, leadership, administrative competence, interpersonal competence,* and *compliance with/acceptance of authority*) could be regarded contextual performance.

Job-specific frameworks often used multiple, more specific dimensions to describe contextual performance. For example, Arvey and Mussio [36] described contextual performance of clerical workers using the dimensions of *cooperating and taking on extra load, showing responsibility and initiative*, *dealing with others in the*

organization, and *dealing with public*. Campbell, McHenry and Wise [23], distinguished *general soldiering proficiency, effort*, *leadership*, *personal discipline,* and *physical fitness and military bearing* as dimensions of work performance in the army. Borman and Brush [24] distinguished *leadership and supervision*, *interpersonal dealings and communication*, and *useful personal behavior and skills* as dimensions of managerial work performance. Altogether, dimensions frequently named under contextual performance are communication, effort, discipline, interpersonal behavior, and leading and developing others. Less frequently named dimensions are planning, solving problems, administration, and showing responsibility.

Counterproductive work behavior

Attention for counterproductive work behavior (CWB), defined as behavior that harms the well-being of the organization, has increased in recent years [20]. It includes behaviors such as absenteeism, being late for work, engaging in off-task behavior, theft, and substance abuse.

Almost half of the generic individual work performance frameworks incorporated one or more dimensions of counterproductive work behavior. Murphy used the dimensions of *destructive/hazardous behaviors* (behaviors leading to a clear risk of productivity losses, damage, or other setbacks) and *down-time behaviors* (work-avoidance behaviors) to describe behaviors that harm the organization [17]. Hunt's framework incorporated the four dimensions of *off-task behavior*, *unruliness*, *theft,* and *drug misuse* [40]. Viswesvaran and Ones [11] as well as Rotundo and Sackett [20], concluded in their reviews that counterproductive work behavior should be distinguished as a third broad dimension of individual work performance (in addition to task performance and organizational citizenship behavior). Finally, some individual work performance frameworks that only focus on counterproductive work behavior were identified. Burton et al. [41], Allen [42], and Escorpizo [43] approached the study of work performance from an occupational health perspective, and divided the work performance domain into *absenteeism* (not attending work) and *presenteeism* (attending work while ill). Both absenteeism and presenteeism could be regarded as counterproductive work behaviors, as they are behaviors that harm the well-being of the organization.

Sinclair and Tucker's framework was the only job-specific framework to incorporate counterproductive work behavior as a separate dimension of individual work performance [44].

Other dimensions

To examine the impact of fun at work on work performance, Fluegge divided the domain of individual work performance into task performance, organizational citizenship behavior, and creative performance [34]. Creative performance was defined as behavioral manifestations of creativity, which refer to the generation of ideas, procedures, and products that are both novel and useful.

Allworth and Hesketh [45], Pulakos et al. [33] and Griffin et al. [13] focused on the growing interdependency and uncertainty of work systems and the corresponding change in the nature of individual work performance. All three argued that adaptive performance should be a separate dimension of individual work performance. Adaptive performance is defined as the extent to which an individual adapts to changes in a work system or work roles [13]. It includes, for example, solving problems creatively, dealing with uncertain or unpredictable work situations, learning new tasks, technologies and procedures, and adapting to other individuals, cultures or physical surroundings. Griffin et al. [13] further argued for *task proactivity* as a separate dimension of work performance. Individual *task proactivity* reflected the extent to which individuals engage in self-starting, future-oriented behavior to change their work situations, their work roles, or themselves.

Sinclair and Tucker's job-specific framework [44] also regarded *adaptive performance* as a separate dimension of individual work performance, in addition to task performance, contextual performance, and counterproductive work behavior. In several other frameworks, adaptive performance was not included as a separate dimension, but rather as a part of contextual performance. For example, Hunt's dimension of *schedule flexibility* [40], Rollins and Fruge's dimension of *adaptability* [27], and Hedge et al.'s dimension of *leading change* [46] all reflected an employee's ability to adapt to new job conditions or requirements.

Table 2. Overview of identified conceptual frameworks of individual work performance and classification of their dimensions

	Dimension			
Generic framework	*Task performance*	*Contextual performance*	*Counterproductive work behavior*	*Other*
Murphy, 1989 [17] Also presented in [14;20;47;48]	Task behaviors	Interpersonal behaviors	Downtime behaviors Destructive/hazardous behaviors	
Campbell, 1990 [4] Also presented in [10;11;13;14;18;20;28;29;32;44;48-57]	Job-specific task proficiency Non-job-specific task proficiency	Written and oral communication Demonstrating effort Maintaining personal discipline Facilitating peer and team performance Supervision and leadership Management and administration		
Borman & Motowidlo, 1993 [19] Also presented in [7;10;11;13;14;18;20;25;32;44;48;50-53;55;56;58-63]	Task performance	Contextual performance		

Table 2. Continued

Generic framework	Dimension			
	Task performance	*Contextual performance*	*Counterproductive work behavior*	*Other*
Viswesvaran, 1993 [18] Also presented in [10;11;14;48;53;57;60]	Productivity Quality Job knowledge	Communication competence Effort Leadership Administrative competence Interpersonal competence Compliance with and acceptance of authority		Overall work performance
Hunt, 1996 [40] Also presented in [11;14;20;32;48;57]		Adherence to rules Industriousness Thoroughness Schedule flexibility Attendance	Off-task behavior Unruliness Theft Drug misuse	
Allworth & Hesketh, 1999 [45]	Task performance	Contextual performance		Adaptive performance

Table 2. Continued

	Dimension			
Generic framework	*Task performance*	*Contextual performance*	*Counterproductive work behavior*	*Other*
Viswesvaran & Ones, 2000 [11] Also presented in [60]	Task performance	Organizational citizenship behavior	Counterproductive behavior	
Michel, 2000 [32]	Task performance	Interpersonal performance Civic performance		
Pulakos et al., 2000 [33] Also presented in [13;48;55;64]	Task performance	Contextual performance		Adaptive performance
Renn & Fedor, 2001 [35]	Work quantity Work quality			
Rotundo & Sackett, 2002 [20] Also presented in [15;65;66]	Task performance	Organizational citizenship behavior	Counterproductive behavior	
Bakker et al., 2004 [31]	In-role performance	Extra-role performance		
Burton et al., 2004 [41]			Absenteeism Presenteeism	

Table 2. Continued

Generic framework	Dimension			
	Task performance	*Contextual performance*	*Counterproductive work behavior*	*Other*
Griffin et al., 2007 [13] Also presented in [15]	Task proficiency			Adaptability Proactivity
Allen, 2008 [42]			Absenteeism Presenteeism	
Escorpizo, 2008 [43]			Absenteeism Presenteeism	
Fluegge, 2008 [34]	Task performance	Organizational citizenship behavior		Creative performance

Table 2. Continued

Job-specific framework	Dimension			
	Task performance	*Contextual performance*	*Counterproductive work behavior*	*Other*
Arvey & Mussio, 1973 [36]	Working accurately Showing concern for time Detail and planning	Cooperating and extra time Dealing with others in organization Dealing with public Showing responsibility and initiative		
Jiambalvo, 1979 [37]	Understanding Planning Revising	Promoting Providing training Recognizing problems Suggesting solutions Reviewing work Cooperation Respect Special competence		
C. Campbell et al., 1990 [26]	Job-specific proficiency	Non-job-specific proficiency		
J. Campbell et al., 1990 [23] Also presented in [30]	Core technical proficiency	General soldiering proficiency Effort and leadership Personal discipline Physical fitness and military bearing		

Table 2. Continued

Job-specific framework	Dimension			
	Task performance	*Contextual performance*	*Counterproductive work behavior*	*Other*
Lance et al., 1992 [29]	Technical proficiency	Interpersonal proficiency		
Rollins & Fruge, 1992 [27]	Task proficiency	Action Teamwork Creativity Communication Decision making Leadership Accountability Adaptability Development		
Borman & Brush, 1993 [24] Also presented in [20;48;53]	Technical activities and mechanics of management	Interpersonal dealings and communication Leadership and supervision Useful personal behavior and skills		
Engelbrecht & Fischer, 1995 [38]	Action orientation Task structuring Probing, synthesis and judgment	Empathy Development Managing information		

Table 2. Continued

	Dimension			
Job-specific framework	*Task performance*	*Contextual performance*	*Counterproductive work behavior*	*Other*
Tett et al., 2000 [39]	Traditional functions Occupational acumen and concerns	Task orientation Person orientation Dependability Open mindedness Emotional control Communication Developing self and others		
Van Dyne et al., 2002 [54]	Sales performance	Creativity		
Hedge et al., 2004 [46]	Resource stewardship	Coaching and mentoring Professionalism and integrity Communication skills Leading change Leading people Organizational savvy Personal and professional development		

Table 2. Continued

	Dimension			
Job-specific framework	*Task performance*	*Contextual performance*	*Counterproductive work behavior*	*Other*
Chan, 2006 [67]		Communication skills Interpersonal skills Customer service Analytical skills		
Sinclair & Tucker, 2006 [44]	Task performance	Contextual performance	Counterproductive behavior	Adaptive performance
Greenslade & Jimmison, 2007 [62]	Task performance	Contextual performance		
Wisecarver et al., 2007 [28]	Job-specific task proficiency Interpersonal job-specific task proficiency	Non-job-specific task proficiency Management Peer-team interaction Discipline Effort		

Table 2. Continued

	Dimension			
Job-specific framework	*Task performance*	*Contextual performance*	*Counterproductive work behavior*	*Other*
Luo et al., 2008 [68]	Military training Task accomplishment Work capability	Helping others Love of learning Promoting organizational benefit Self-discipline		
Maxham et al., 2008 [25]	In-role performance	Extra-role performance toward customers Extra-role performance toward organization		
Mael et al., 2010 [69]	Providing clinical services Clinical support	Employee citizenship behavior Managerial behavior		

Heuristic conceptual framework of individual work performance

The second aim of the current review was to integrate existing conceptual frameworks in order to formulate a heuristic conceptual framework of individual work performance. We propose a heuristic framework, presented in Figure 2, which may serve as a guide towards understanding the construct of individual work performance. At the highest level appears the latent, general factor of individual work performance. Research has shown the existence of a general factor, which accounts for substantial variation in job performance ratings [60]. At the second level, four dimensions of individual work performance are located. At the third level, the individual measures corresponding to each dimension are located. The importance of these dimensions, and the exact indicators associated with each dimension, may differ depending on the context involved.

The first dimension, task performance, refers to the proficiency with which central job tasks are performed [4]. The second dimension, contextual performance, refers to behaviors that support the organizational, social and psychological environment in which the technical core must function [19]. As a third dimension, adaptive performance is included in the heuristic framework. Three reasons support the inclusion of adaptive performance, referring to an employee's ability to adapt to changes in a work system or work roles [13], as a separate dimension. First, because of the technological changes occurring in today's society, being able to adapt to a changing work environment is increasingly important. Second, conceptually, adaptive performance does not fit neatly under task performance, contextual performance or counterproductive work behavior. Whereas contextual performance comprises behaviors that positively influence the work environment, adaptive performance comprises behaviors in reaction to the changing work environment. Third, empirical support for adaptive performance as a separate dimension was provided by Allworth and Hesketh [45]. They found that adaptive performance had differential predictors than task or contextual performance. The fourth dimension, counterproductive work behavior, refers to behavior that harms the well-being of the organization [20].

Other dimensions that have been suggested as separate dimensions are proactive and creative performance [13;34]. Although proactive and creative performance can be a part of task performance in some jobs, we considered these part of contextual performance, as both contribute to a positive organizational, social and psychological work environment. Each of the four dimensions are latent, meaning that they cannot be measured directly [14]. Example indicators of each

dimension that were gathered from the identified frameworks are presented in the square boxes in Figure 2.

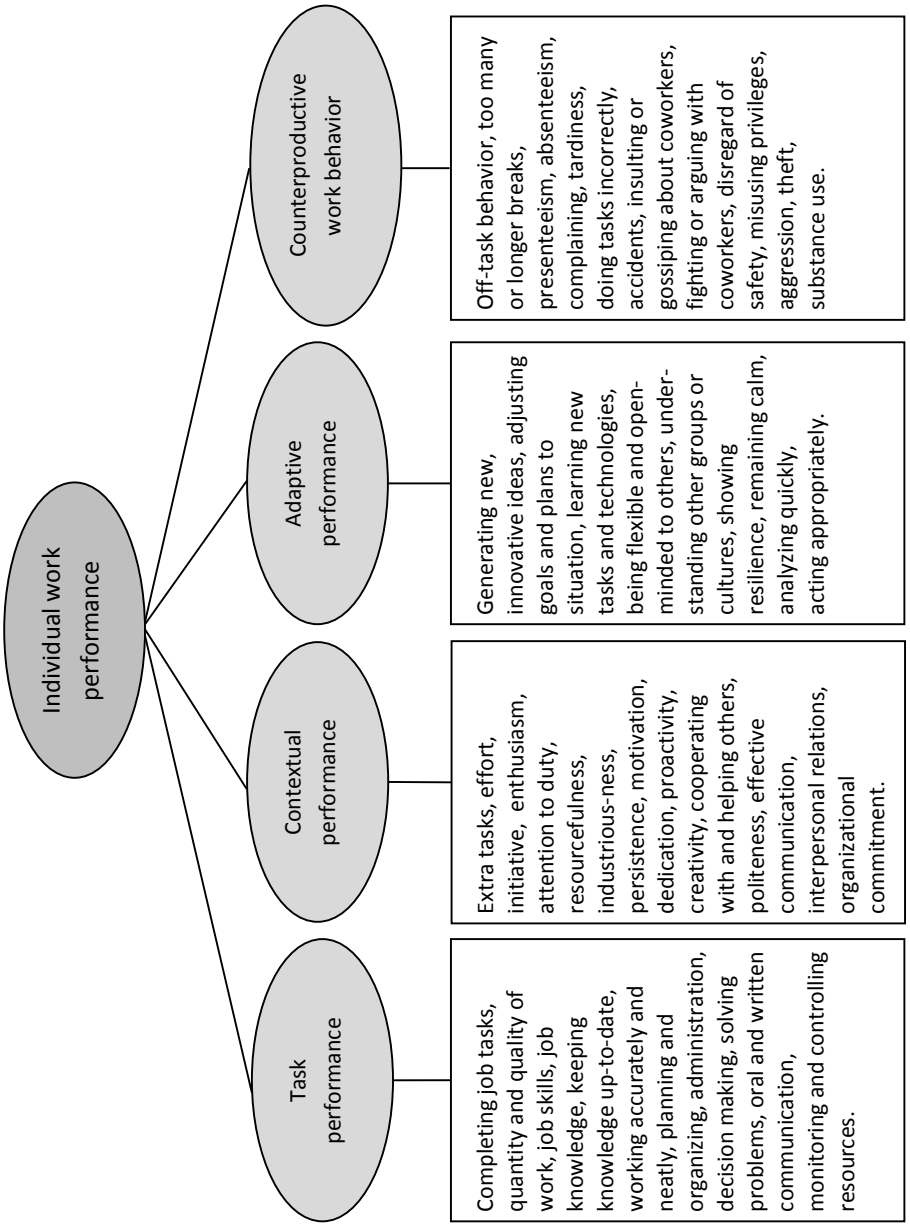


Figure 2. Heuristic framework of individual work performance

Relation between dimensions

Not only are the separate dimensions related to the general factor of work performance, they are also related to each other [60]. Interesting is the question of how the separate dimensions interrelate. Task performance is distinct, albeit strongly positively related, with contextual performance [57;70]. Both types of behavior independently contribute to overall performance, but through different means [19;52]. Due to the changing nature of today's work, the distinction between task and contextual performance may become more blurred [53]. Increasingly, contextual behaviors are implicitly or explicitly required as task behaviors. Also, some behaviors can be seen as task behaviors in some jobs, while they may be seen as contextual behaviors in other jobs. Findings on the relation between task performance and counterproductive work behavior are inconclusive and have been found to be either moderately or strongly negative [71]. The inconclusive findings could be caused by differences in definition and measurement of task performance. When task performance is defined as what a person generally 'will do,' it is more strongly related to counterproductive work behavior than when task performance is defined as what a personal maximally 'can do.' This is because typical work performance is usually assessed over a longer time period, in which counterproductive work behaviors are more likely to occur. In addition, typical task performance is often less closely monitored than maximal task performance, making counterproductive work behaviors more likely to occur [71].

Intuitively, one would expect a negative relation between contextual behavior and counterproductive work behavior. Someone who often engages in behavior that helps the organization, will not often engage in behavior that harms the organization, and vice versa. Although a strong negative correlation has been found previously [71], meta-analysis demonstrated that the true relation between contextual performance and counterproductive work behavior is modestly negative [72]. Three methodological artifacts may have caused the strong negative relation between contextual performance and counterproductive work behavior in previous research. First, the relation was found to be more strongly negative when the behaviors were rated by supervisors rather than by the employees themselves. This is because supervisors often cannot accurately observe an employee's counterproductive work behavior, and make their judgment based on general impressions of the employee. Second, the relation was more strongly negative when contextual behavior inventories included dysfunctional behaviors (e.g., "not adhering to organizational rules") or when counterproductive work behavior

inventories included functional behaviors (e.g., “adhering to organizational rules”). Dalal [72] termed these overlapping items *antithetical items*. Third, asking respondents to indicate the extent to which they agreed or disagreed with statements about their behavior resulted in a stronger negative relation between contextual performance and counterproductive work behavior than when they were asked to indicate the frequency of their behavior. In conclusion, the relation between contextual performance and counterproductive work behavior is modestly negative. Thus, employees who engage in helping behavior tend not to engage in harming behavior (or vice versa), but both types of behavior can occur together, at least to some extent [73].

Although Pulakos [74] stated that adaptive performance does not occur completely independent of task and contextual performance, to our knowledge, no research has been published that examines the relation between adaptive performance and the other individual work performance dimensions. However, as adaptive performance is regarded behavior that positively influences individual work performance, one can expect a positive relation with task and contextual performance, and a negative relation with counterproductive work behavior.

Discussion

The aim of the current review was to identify conceptual frameworks of individual work performance from different fields, in order to formulate a heuristic conceptual framework. In total, seventeen generic frameworks were identified that addressed individual work performance across occupations. Eighteen job-specific frameworks were identified that addressed work performance of either professionals in the army, managers or employees in the service and sales industry. Although job-specific frameworks often used multiple, more specific dimensions than generic frameworks to describe the construct of individual work performance, clear similarities were observed in the dimensions of these frameworks. A heuristic framework of individual work performance was proposed in which individual work performance consists of four dimensions, namely task performance, contextual performance, adaptive performance, and counterproductive work behavior. These four types of behavior can be considered to capture the full range of behaviors that constitute individual work performance in virtually any job. The importance of the four dimensions, and their exact indicators, may however differ based on the specific context.

In accordance with two previous narrative reviews [11;20], the heuristic framework incorporates task performance, contextual performance and counterproductive work behavior as important dimensions of individual work performance. However, the current review presents an updated conceptual framework in which adaptive performance is added to the domain of individual work performance. Although some frameworks have included adaptive performance as a part of contextual performance, we presented societal, conceptual, and empirical reasons for distinguishing adaptive performance as a separate dimension.

Surprisingly, none of the identified frameworks included all four dimensions proposed in the heuristic framework, except for the recent Sinclair and Tucker framework for work performance of soldiers [44]. While most of the identified individual work performance frameworks circulate in the field of management or in the field of work and organizational psychology, they have been almost absent in the field of occupational health. In this field, only three studies were identified, describing individual work performance as existing of absenteeism and presenteeism [41-43]. Hopefully, the present review will facilitate information exchange between the different areas of research.

Strengths and limitations

The present review has several strengths. First of all, this is the first study to examine the construct of individual work performance from different research fields, namely occupational health, psychology, and management. We concluded that there are considerable similarities between the different fields. Second, this is the first study to conduct a comprehensive, systematic literature search in order to identify frameworks describing the construct of individual work performance. Both earlier reviews [11;20] were narrative reviews describing a limited amount of conceptual frameworks. Third, both generic and job-specific frameworks were included in the present review. Overall, the present review provides a strong basis for the proposed heuristic framework.

The present review has some limitations as well. We aimed to describe all individual work performance frameworks as comprehensively as possible. However, it cannot be ruled out that some frameworks were missed. Although no restrictions in year or type of publication were made, only Dutch and English literature was searched. Furthermore, all literature was searched for the search terms in their title or abstract. This may have excluded studies that did present a conceptual framework of individual work performance, but did not mention this in their title or abstract.

We tried to minimize these limitations by searching four databases, using broad search terms and checking reference lists from identified studies. Unfortunately, 24 dissertations that appeared eligible based on title and abstract had to be excluded from this review, because, after thorough searches, we were unable to retrieve them in full text.

The original dimensions of the identified frameworks were classified into the heuristic framework based on the definitions of the dimensions as provided by the developers of the framework. However, depending on the specific context, the importance and the place of the original dimensions may differ. For example, in some jobs, adaptive performance may not be important at all, and could therefore be excluded from the heuristic framework. In some jobs, communication competence may be an aspect of contextual performance, while in others jobs it may be an aspect of task performance. Thus, the dimensions, and subsequently the indicators used for its measurement, may differ depending on the context. In addition, the heuristic framework, or any job performance framework for that matter, is influenced by the Zeitgeist. Contextual performance (and its variants) have gained currency in the 1960s, whereas adaptive performance has gained currency only in recent years. In one or two decades, other dimensions may become important (for example, environmental sustainability) and some existing dimensions may become oblivious.

Recommendations for future research

The heuristic framework of individual work performance that is presented in the current review can serve as a guide for future research and practice. An important next step is to identify existing measures or to develop a new measurement instrument that can adequately measure the individual work performance dimensions proposed in the heuristic framework. This will involve determining the exact indicators of each dimension more comprehensively. Empirical data gathered with the measurement instrument provides information on whether the four-dimensional structure of the proposed framework is supported, whether the indicators belong to the expected dimension, and on the exact relation between the different dimensions.

Furthermore, future research is needed to examine whether the proposed framework is generalizable across all types of jobs. Considering the similarity in dimensions observed between generic and job-specific frameworks, we believe that the broad dimensions of the proposed framework will likely generalize across all job types. The importance of the dimensions, and their exact indicators, may however

be job-specific. An important next step for future research is therefore to determine whether or not the four dimensions proposed in the heuristic framework are generic, and whether the indicators per dimensions are job-specific or generic. Future research is also needed to determine whether adding adaptive performance as a separate dimension of individual work performance is justified. In this sense, the proposed framework may be considered a theory-driven 'working' framework that can be adapted based on future empirical research.

Practical use of the framework

An important use of the heuristic framework is in shaping the design of workplace interventions and assessing the effects of that intervention on individual work performance. Think, for example, of intervention studies that target managerial style or employee lifestyle and health in order to increase an individual's work performance. Furthermore, the heuristic framework can be used in companies for employee selection, evaluation, training and development.

Future research may expand the heuristic framework to include causes and consequences of individual work performance. Individual work performance is inextricably linked to team and organizational performance, although the exact nature of this relationship is yet unknown. A final extension would be to expand the heuristic framework to the team and organizational levels, and possibly, to relate it to company costs.

Conclusion

Concluding, the dimensions of task performance, contextual performance, adaptive performance and counterproductive work behavior comprise the heuristic framework of individual work performance. Future research will have to determine empirical support for and practical relevance of this framework. Hopefully, this review provides a step towards reaching consensus on the conceptualization and operationalization of individual work performance. A better understanding of this construct will improve theory, research and practice in all fields occupied with individual work performance.

References

1. Waldman DA. The contributions of total quality management to a theory of work performance. The Academy of Management Review 1994;19(3):510-36.
2. Evans CJ. Health and work productivity assessment: state of the art or state of flux? J Occup Environ Med 2004;46(6):S3-S11.
3. Lerner D, Mosher Henke R. What does research tell us about depression, job performance, and work productivity? J Occup Environ Med 2008;50(4):401-10.
4. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, eds. Handbook of industrial and organizational psychology. Palo Alto, CA, US: Consulting Psychologists Press. 1990:687-732.
5. Beaton D, Bombardier C, Escorpizo R, et al. Measuring worker productivity: frameworks and measures. J Rheumatol 2009;36(9):2100-9.
6. Schultz AB, Chen CY, Edington DW. The cost and impact of health conditions on presenteeism to employers. Pharmacoeconomics 2009;27(5):365-78.
7. Halbeslebe JRB, Wheeler AR, Buckley MR. Clarifying the relationship between organizational commitment and job performance: Extending the conservation of resources model. In: Kiefer KH, ed. Applied psychology research trends. Hauppauge, NY, US: Nova Science Publishers; US. 2008:35-58.
8. Judge TA, Bono JE, Thoreson CJ, Patton GK. The job satisfaction-job performance relationship: a qualitative and quantitative review. Psychological Bulletin 2001;127(3):376-407.
9. Barrick MR, Mount MK, Judge TA. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? International Journal of Selection and Assessment 2001;9(1/2):9-30.
10. Tubre T, Arthur WJ, Bennett WJ. General Models of Job Performance: Theory and Practice. In: Bennett W Jr., Lance CE, Woehr DJ, eds. Performance measurement: Current perspectives and future challenges. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. 2006:175-203.
11. Viswesvaran C, Ones DS. Perspectives on models of job performance. International Journal of Selection and Assessment 2000 Dec;8(4):216-26.
12. Kemppila S, Lonnqvist A. Subjective productivity measurement. The Journal of American Academy of Business 2003;2(2):531-7.

13. Griffin MA, Neal A, Parker SK. A new model of work role performance: positive behavior in uncertain and interdependent contexts. Academy of Management Journal 2007;50(2):327-47.
14. Viswesvaran C. Assessment of individual job performance: A review of the past century and a look ahead. Anderson N, Ones DS, Sinangil HK, Viswesvaran C, eds. Handbook of industrial, work and organizational psychology, Volume 1: Personnel psychology. Thousand Oaks, CA: Sage Publications Ltd. 2002:110-126.
15. Fay D, Sonnentag S. A look back to move ahead: New directions for research on proactive performance and other discretionary work behaviours. Applied Psychology: An International Review 2010;59(1):1-20.
16. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. Journal of the Royal Statistical Society 2002;165(2):233-61.
17. Murphy KR. Dimensions of job performance. In: Dillon RF, Pellegrino JW, eds.Testing: Theoretical and Applied Perspectives. New York: Praeger. 1989:218-47.
18. Viswesvaran C. Modeling job performance: Is there a general factor? Ph.D. dissertation. United States - Iowa: The University of Iowa. 1993.
19. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, eds. Personnel selection in organizations. San Francisco, CA: Jossey Bass. 1993:71-98.
20. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: a policy-capturing approach. J Appl Psychol 2002;87(1):66-80.
21. Organ DW. Organizational citizenship behavior: the good soldier syndrome. Lexington, MA: Lexington Books. 1988.
22. LePine JA, Erez A, Johnson DE. The nature and dimensionality of organizational citizenship behavior: a critical review and meta-analysis. J Appl Psychol 2002;87(1):52-65.
23. Campbell JP, McHenry JJ, Wise LL. Modeling job performance in a population of jobs. Personnel Psychology 1990;43(2):313-33.
24. Borman WC, Brush DH. More progress toward a taxonomy of managerial performance requirements. Human Performance 1993;6(1):1-21.

25. Maxham JGI, Netemeyer RG, Lichtenstein DR. The retail value chain: Linking employee perceptions to employee performance, customer evaluations, and store performance. Marketing Science 2008;27(2):147-67.
26. Campbell CH, Ford P, Rumsey MG, et al. Development of multiple job performance measures in a representative sample of jobs. Personnel Psychology 1990;43(2):277-300.
27. Rollins T, Fruge M. Performance Dimensions: Competencies with a Twist. Training 1992;29(1):47-51.
28. Wisecarver MM, Carpenter TD, Kilcullen RN. Capturing interpersonal performance in a latent performance model. Military Psychology 2007;19(2):83-101.
29. Lance CE, Teachout MS, Donnelly TM. Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. J Appl Psychol 1992;77(4):437-52.
30. Campbell JP, Hanson MA, Oppler SH. Modeling performance in a population of jobs. In: Campbell JP, Knapp DJ, eds. Exploring the limits in personnel selection and classification. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. 2001;307-333.
31. Bakker AB, Demerouti E, Verbeke W. Using the Job Demands-Resources Model To Predict Burnout and Performance. Human Resource Management 2004;43(1):83-104.
32. Michel RP. A model of entry-level job performance. Dissertation Abstracts International: Section B: The Sciences and Engineering. Vol.61(6-B), Jan 2000:3313.
33. Pulakos ED, Arad S, Donovan MA, Plamondon KE. Adaptability in the workplace: development of a taxonomy of adaptive performance. J Appl Psychol 2000;85(4):612-24.
34. Fluegge ER. Who put the fun in functional? Fun at work and its effects on job performance. Dissertation Abstracts International Section A: Humanities and Social Sciences. Vol69(7-A),2009:2781.
35. Renn RW, Fedor DB. Development and field test of a feedback seeking, self-efficacy, and goal setting model of work performance. Journal of Management 2001;27(5):563.
36. Arvey RD, Mussio SJ. A test of expectancy theory in a field setting using female clerical employees. Journal of Vocational Behavior 1973;3(4):421-32.

37. Jiambalvo J. Performance Evaluation and Directed Job Effort: Model Development and Analysis in a CPA Firm Setting. Journal of Accounting Research 1979;17(2):436.
38. Engelbrecht AS, Fischer AH. The managerial performance implications of a developmental assessment center process. Human Relations 1995;48(4):387-404.
39. Tett RP, Guterman HA, Bleier A, Murphy PJ. Development and content validation of a "hyperdimensional" taxonomy of managerial competence. Human Performance 2000;13(3):205-51.
40. Hunt ST. Generic work behavior: an investigation into the dimensions of entry-level, hourly job performance. Personnel Psychology 1996;49(1):51-83.
41. Burton WN, Pransky G, Conti DJ, Chen CY, Edington DW. The association of medical conditions and presenteeism. J Occup Environ Med 2004;46(6 Suppl):S38-S45.
42. Allen H. Using routinely collected data to augment the management of health and productivity loss. J Occup Environ Med 2008;50(6):615-32.
43. Escorpizo R. Understanding work productivity and its application to work-related musculoskeletal disorders. International Journal of Industrial Ergonomics 2008;38(3-4):291-7.
44. Sinclair RR, Tucker JS. Stress-CARE: An Integrated Model of Individual Differences in Soldier Performance under Stress. In: Britt TW, Castro CA, Adler AB, eds. Military life: The psychology of serving in peace and combat (Vol. 1): Military performance. Westport, CT: Praeger Security International. 2006:202-231.
45. Allworth E, Hesketh B. Construct-oriented Biodata: Capturing Change-related and Contextually Relevant Future Performance. International Journal of Selection and Assessment 1999;7(2):97-111.
46. Hedge JW, Borman WC, Bruskiewicz KT, Bourne MJ. The development of an integrated performance category system for supervisory jobs in the US Navy. Military Psychology 2004;16(4):231-43.
47. Murphy KR. Job performance and productivity. In: Murphy KR, Saal FE, eds. Psychology in organizations: Integrating science and practice. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. 1990:157-76.
48. Levine PJ. Substandard job performance: Development of a model and the examination of poor organizational performers. Dissertation Abstracts

International: Section B: The Sciences and Engineering. Vol.67(9-B), 2007:5454.

49. Campbell JP. Alternative models of job performance and their implications for selection and classification. In: Rumsey MG, Walker CB, Harris JH, eds. Personnel selection and classification. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. 1994:33-51.
50. Conway JM. Additional construct validity evidence for the task/contextual performance distinction. Human Performance 1996;9(4):309-29.
51. Borman WC, Hanson MA, Hedge JW. Personnel selection. Annu Rev Psychol 1997;48:299-337.
52. Motowidlo SJ, Borman WC, Schmit MJ. A theory of individual differences in task and contextual performance. Human Performance 1997;10(2):71-83.
53. Arvey RD, Murphy KR. Performance evaluation in work settings. Annu Rev Psychol 1998;49:141-68.
54. Van Dyne L, Jehn KA, Cummings A. Differential effects of strain on two forms of work performance: Individual employee sales and creativity. Journal of Organizational Behavior 2002;23(1):57.
55. Borman WC, Hedge JW, Ferstl KL, Kaufman JD, Farmer WL, Bearden RM. Current directions and issues in personnel selection and classification. In: Martocchio JJ, Ferris GR, eds. Research in personnel and human resources management. Oxford, England: Elsevier Science Ltd. 2003:297-355.
56. Surface EA. An integration of the training evaluation and job performance modeling literatures: Confirming BE KNOW DO with United States army special forces training data. Dissertation Abstracts International: Section B: The Sciences and Engineering. Vol.65(1-B),2004:471.
57. Hoffman BJ, Blair CA, Meriac JP, Woehr DJ. Expanding the criterion domain? A quantitative review of the OCB literature. J Appl Psychol 2007;92(2):555-66.
58. Van Scotter JR, Motowidlo SJ. Interpersonal facilitation and job dedication as separate facets of contextual performance. J Appl Psychol 1996;81(5):525.
59. Borman WC, Penner LA, Allen TD, Motowidlo SJ. Personality predictors of citizenship performance. International Journal of Selection and Assessment 2001;9(1-2):52-69.
60. Viswesvaran C, Schmidt FL, Ones DS. Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences. J Appl Psychol 2005;90(1):108-31.

61. Ellington JK. Systematic sources of variance in supervisory job performance ratings: A multilevel analysis of between-rater and between-context variance. Dissertation Abstracts International: Section B: The Sciences and Engineering. Vol.67(12-B), 2007:7413.
62. Greenslade JH, Jimmieson NL. Distinguishing between task and contextual performance for nurses: Development of a job performance scale. J Adv Nurs 2007;58(6):602-11.
63. Wang H, Law K, Chen Z. Leader-member exchange, employee performance, and work outcomes: an empirical study in the Chinese context. The International Journal of Human Resource Management 2008;19(10):1809.
64. Pulakos ED, Schmitt N, Dorsey DW, Arad S, Hedge JW, Borman WC. Predicting adaptive performance: Further tests of a model of adaptability. Human Performance 2002;15(4):299-324.
65. Colquitt JA, Scott BA, LePine JA. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. J Appl Psychol 2007;92(4):909-27.
66. Mount MK, Oh IS, Burns M. Incremental validity of perceptual speed and accuracy over general mental ability. Personnel Psychology 2008;61(1):113-39.
67. Chan DC. Core competencies and performance management in Canadian public libraries. Library Management 2006;27(3):144.
68. Luo Z, Shi K, Li W, Miao D. Construct of job performance: Evidence from Chinese military soldiers. Asian Journal of Social Psychology 2008;11:222-31.
69. Mael FA, O'Shea PG, Smith MA, Burling AS, Carman KL, Haas A, et al. Development of a model and measure of process-oriented quality of care for substance abuse treatment. J Behav Health Serv Res 2010;37(1):4-24.
70. Conway JM. Distinguishing contextual performance from task performance for managerial jobs. J Appl Psychol 1999;84:3-13.
71. Sackett PR. The structure of counterproductive work behaviors: dimensionality and relationships with facets of job performance. International Journal of Selection and Assessment 2002;10(1/2):5-11.
72. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol 2005;90(6):1241-55.
73. Spector PE, Fox S. Theorizing about the deviant citizen: an attributional explanation of the interplay of organizational citizenship and

counterproductive work behavior. Human Resource Management Review 2010;20:132-43.

74. Pulakos ED, Dorsey DW, White SS. Adaptability in the workplace: selecting an adaptive workforce. Advances in Human Performance and Cognitive Engineering Research 2006;6:41-71.

Chapter 3

Measuring Individual Work Performance - Identifying and Selecting Indicators

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt,
Henrica C.W. de Vet, Allard J. van der Beek

WORK: A Journal of Prevention, Assessment & Rehabilitation. 2013; 45(3)

DOI 10.3233/WOR-131659

3

Abstract

Background: Theoretically, individual work performance (IWP) can be divided into four dimensions: task performance, contextual performance, adaptive performance, and counterproductive work behavior. However, there is no consensus on the indicators used to measure these dimensions.

Objective: This study was designed to 1) identify indicators for each dimension, 2) select the most relevant indicators, and 3) determine the relative weight of each dimension in ratings of work performance.

Methods: IWP indicators were identified from multiple research disciplines, via literature, existing questionnaires, and expert interviews. Subsequently, experts selected the most relevant indicators per dimension and scored the relative weight of each dimension in ratings of IWP.

Results: In total, 128 unique indicators were identified. Twenty-three of these indicators were selected by experts as most relevant for measuring IWP. Task performance determined 36% of the work performance rating, while the other three dimensions respectively determined 22%, 20% and 21% of the rating.

Conclusions: Notable consensus was found on relevant indicators of IWP, reducing the number from 128 to 23 relevant indicators. This provides an important step towards the development of a standardized, generic and short measurement instrument for assessing IWP.

Introduction

Although an individual's performance at work is one of the most important outcomes of studies in the occupational setting, recent research has shown that there is no consensus on the definition and measurement of individual work performance [1-3]. Various terms (often used interchangeably) are used to describe individual work performance (IWP), such as presenteeism, performance, or productivity. The definitions of these terms are often unclear. This is undesirable, because a clear definition and theoretical framework of IWP is a prerequisite for its valid measurement. Valid measurement, in turn, is necessary to accurately establish the causes and consequences of IWP.

Defining individual work performance

Considering the importance of IWP, it is not surprising that disciplines other than occupational medicine have concerned themselves with defining and measuring the concept. Within work and organizational psychology, defining the construct of IWP and attempting to understand its underlying structure has received much attention [4]. In the latter discipline, IWP is generally defined as "behaviors or actions that are relevant to the goals of the organization" [5]. Thus, IWP is defined in terms of behaviors or actions of employees, rather than the results of these actions. In addition, IWP consists of behaviors that are under the control of the individual, thus excluding behaviors that are constrained by the environment [6].

Recently, a heuristic framework of IWP was proposed in a multi-disciplinary, systematic literature review [7], in which IWP consisted of four broad and generic dimensions. The first dimension, *task performance*, refers to the proficiency with which an employee performs central job tasks [5]. The second dimension, *contextual performance*, refers to employee behaviors that support the organizational, social, and psychological environment in which the central job tasks are performed [8]. The third dimension, *adaptive performance*, refers to an employee's proficiency in adapting to changes in work roles or environment [9]. The fourth dimension, *counterproductive work behavior*, refers to behavior that is harmful to the well-being of the organization [6].

Measuring individual work performance

While four generic dimensions of IWP can be distinguished, there is still little consensus on how to measure the concept. Within occupational medicine, various

questionnaires exist to measure IWP or similar constructs, such as the Stanford Presenteeism Scale [10], Work Productivity and Activity Impairment [11], and Health and Performance Questionnaire [12]. Within work and organizational psychology, numerous scales have also been developed to measure task performance [e.g., 13], contextual performance [e.g., 14], or counterproductive work behavior [e.g., 15]. The multitude of scales in this discipline is perhaps best illustrated by LePine, Erez and Johnson [16], who identified more than 40 different measures of contextual performance.

The heterogeneous content of IWP measures is likely related to the use of different definitions of IWP, or a lack thereof, and by the use of different developmental or target populations. What is noticeable in the measures developed in occupational medicine is that a clear definition and theoretical model of IWP is often lacking [17-19]. Although the measures developed in work and organizational psychology do use definitions of IWP dimensions, none of them captured the complete range of individual behaviors at work. Moreover, measures from occupational medicine are often designed for individuals with physical or mental health problems [20]. This makes these measures less suitable for assessing IWP in healthy workers. In addition, the measures developed in work and organizational psychology, although intended to be generic, were often developed and refined based on a specific occupation.

Goal of the current study

The current lack of consensus on how to measure IWP impedes valid measurement of the construct Research on IWP would benefit greatly from a standardized, generic, short instrument. The four-dimensional framework of IWP [7] provides a theoretical starting point for developing such an instrument. Next, it is essential to gain consensus on the indicators (important employee behaviors or actions) for each dimension of IWP. Therefore, we designed a study to: 1) identify indicators for each IWP dimension, 2) select the most relevant indicators, and 3) determine the relative weight of each dimension in ratings of IWP. Secondary aims of this study were to determine: 4) whether there were differences between the views of experts from different professional backgrounds (researchers, managers, human resource managers, and occupational health professionals) on the aforementioned research questions, and 5) whether the experts preferred a generic or job-specific questionnaire.

Methods

Identification of indicators

First, indicators were identified from a literature review on conceptual frameworks of IWP [7]. Second, indicators were identified from existing IWP related questionnaires. For this purpose, a systematic search was conducted to identify questionnaires measuring individual work performance or work productivity. As the terms work performance and work productivity are often used synonymously, we incorporated both terms in our search strategies. Search strategies were developed with the aid of experienced search specialists. Search terms included *work performance, work productivity, job performance, employee performance*, or *employee productivity* and *questionnaire, scale*, or *index*. Searches were conducted in two medical databases (PubMed and Embase.com) and one psychological (PsycINFO) database in September 2010. Additional questionnaires were identified by scanning the authors' personal collection of IWP literature. Third, interviews with 16 key-experts were held in November 2010. Key-experts were national and international researchers with over 20 years of experience in the areas of occupational health, psychology, or management.

A review of the literature, questionnaires and data from the experts interviewed resulted in an initial list of IWP indicators. From this list, the first and second author independently removed causal variables (e.g., motivation). In addition, overlapping indicators between dimensions (e.g., 'concentrating' in contextual performance, 'not concentrating' in counterproductive work behavior) were removed from the dimension where they were least well suited. Differences in judgment were resolved through a consensus procedure. Based on conceptual overlap, the first and second author independently reduced the remaining list of indicators, and using the definitions categorized each indicator into the dimension where it best fitted. Differences in judgment were again resolved through a consensus procedure.

Selection of indicators

Participants

A sample of 695 experts from different professional backgrounds (researchers, managers, human resource managers (HRM), and occupational health professionals (OHP)), including the key-experts, were invited to select the most relevant indicators

of IWP. This sample consisted of participants from one national and one international occupational health conference that took place in The Netherlands. Each participant was invited by e-mail to participate in the study. A link to the online questionnaire was included. Depending on their country of residence, participants completed the questionnaire in Dutch or English. Participants had three weeks to complete the questionnaire. After two weeks, non-responders received a reminder via e-mail.

Questionnaire

After a brief introduction to the study, participants were shown a list of all indicators per IWP dimension (task performance, contextual performance, adaptive performance, and counterproductive work behavior). Within dimensions, indicators were presented in random order. Per dimension, participants were asked to select the 6 most important indicators, keeping in mind work performance on a generic level. This means that they had to select indicators that were important in all types of professions, not just in their own profession. For practical reasons, contextual performance was split into two sub-dimensions (interpersonally directed and organizationally directed) because of the large number of indicators related to this dimension. In total, each participant selected 30 (5x6) indicators from the full list of indicators, which he or she believed to be most relevant. After every category of indicators, space was provided for comments and/or suggestions for additional indicators. In the second part of the questionnaire, participants were asked to divide 100 points between the four dimensions, according to the relative weight they would assign to each dimensions when rating IWP. They were also asked whether they believed it possible to develop a generic questionnaire, or whether a combination of a generic and job-specific questionnaire, or a job-specific questionnaire, was more likely. Finally, participants were asked to fill out their gender, age, highest educational level completed, profession, number of years experience in this profession, number of people they manage, and branch of industry. Participants were asked to leave their e-mail address if they wanted to receive the results of the study and be eligible for one of the five gift vouchers to be allotted.

Data analysis

To rank the indicators in order of importance, the percentage of participants who selected an indicator was calculated. Indicators were regarded relevant when they were selected by 40% or more of the participants. Chi-square tests were performed

to examine whether there were significant differences in scores between subgroups (gender, age, educational level, profession, number of years experience in this profession, number of people managed, and branch of industry). Although it was not possible to compute a statistical agreement score (e.g., Cohen's kappa), subgroups showing few statistically significant differences were considered to be in agreement.

To determine the relative weight of each dimension in IWP ratings, the mean number of points assigned to each dimension was calculated. Independent samples t-tests were performed to examine differences in weights between subgroups with two levels (gender). One-way analyses of variance were performed to examine differences in weights between subgroups with more than two levels (age, educational level, profession, number of years experience in this profession, number of people managed, and branch of industry). Post-hoc tests with Bonferroni correction were performed to determine which subgroups differed from one another. If assumptions of homogeneity of variance were not met, then Tamhane's T2 tests were performed to determine which subgroups differed [21].

Finally, the percentage of participants that believed in the development of a generic questionnaire, a combination of a generic and job-specific questionnaire, or a job-specific questionnaire, was calculated. SPSS version 17 was used for the analyses.

Results

Identification of indicators

In the literature review on conceptual frameworks, 54 IWP indicators were identified. In the systematic questionnaire search, 77 questionnaires were identified that aimed to measure the construct of individual work performance or productivity. Of these, full texts of 14 questionnaires could not be retrieved. Another 11 questionnaires were excluded based on full text, because they either did not measure work performance at the individual level, or did not measure work performance at all. Additional scanning of personal collections added 29 questionnaires. The 81 questionnaires yielded 231 IWP indicators which were not already identified in the literature review. None of the questionnaires were found to measure all dimensions of IWP. Only one fifth of the questionnaires explicitly measured one or two of the dimensions of IWP (task performance, contextual performance, adaptive performance, or counterproductive work behavior [e.g., 13,22,15]). More than one third of all questionnaires contained causal variables in

combination with indicator variables [e.g., 10,23]. In addition, a quarter of the questionnaires were developed for individuals with a mental or physical health problem [e.g., 24,25]. More than two third of the questionnaires were developed for generic purposes [e.g., 26,27], whereas around one third were developed for a specific job [e.g., 28,29]. The interviews with 16 key-experts yielded 32 additional IWP indicators, which were not identified in the literature or questionnaires.

In total, the literature, questionnaires, and expert interviews resulted in a list of 317 IWP indicators. The first and second author removed causal variables and indicators overlapping between dimensions. Based on conceptual overlap, the first and second author reduced the remaining list to 128 unique IWP indicators and categorized each indicator into one of the IWP dimensions. Task performance consisted of 26 indicators. Both sub-dimensions of contextual performance consisted of 30 indicators. Adaptive performance consisted of 18 indicators. Counterproductive work behavior consisted of 24 indicators.

Selection of indicators

In total, 253 participants (response rate of 36.4%) participated in the study, including 14 out of 16 key-experts. See Table 1 for participant characteristics. Table 2 presents a list of the indicators that were selected as most relevant for each dimension. The indicators are ranked based on the percentage of votes they received from the total group. Data on the indicators that did not make the final selection is available from the authors upon request. No additional indicators were suggested by the participants during the selection process.

There was high agreement between all subgroups on the importance of the indicators. For 20 items we found statistically significant differences between experts from different professional backgrounds. Hence, subgroups based on profession attributed similar importance to 84% (108 of 128) of the indicators. Subgroups based on gender agreed on 95% of the indicators. Subgroups based on age agreed on 93% of the indicators. Subgroups based on educational level agreed on 87% of the indicators. Subgroups based on number of years work experience agreed on 97% of the indicators. Subgroups based on the number of people one manages agreed on 93% of the indicators. Subgroups based on branch of industry agreed on 92% of the indicators.

Table 1. Participant characteristics

		Profession			
	Total	Researchers	Managers	HRM	OHP
N	253	113	48	54	38
Gender (% female)	47%	59%	33%	44%	29%
Age					
30 years or younger	11%	18%	0%	9%	5%
31 – 50 years	51%	51%	58%	52%	42%
51 years or older	38%	31%	42%	39%	53%
Completed educational level					
Middle-level applied – apprenticeship or certificate	1%	0%	2%	0%	5%
Higher vocational, - Bachelor's degree	26%	0%	35%	67%	34%
Higher academic, - Master's degree	38%	37%	40%	31%	50%
Postgraduate academic, PhD	35%	63%	23%	2%	11%
Work experience					
0 – 5 years	23%	26%	25%	17%	21%
6 – 10 years	21%	21%	19%	28%	16%
10 or more years	56%	53%	56%	56%	63%
Employees managed					
None	48%	47%	19%	59%	71%
1 – 5 employees	23%	28%	15%	22%	18%
6 – 20 employees	18%	17%	33%	15%	10%
21 – 50 employees	6%	4%	21%	2%	0%
50 or more employees	5%	4%	13%	2%	0%
Branch of industry					
Scientific	42%	88%	5%	2%	9%
Policy	5%	1%	2%	16%	3%
Commercial	5%	2%	12%	10%	0%
Service	26%	5%	50%	37%	43%
Trade & industry	4%	0%	10%	4%	6%
Other	18%	4%	21%	31%	40%

Task performance

The task performance dimension originally included 26 indicators. From these, 5 relevant task performance indicators were identified: work quality (69%), planning and organizing work (56%), being result-oriented (46%), prioritizing (45%), and working efficiently (44%).

There was a difference between experts from different professional backgrounds on one of these indicators. On average, researchers judged being result-oriented to be significantly less important (29%) than managers (64%) and human resources managers (66%).

Contextual performance

Initially, there were 60 indicators included in contextual performance. For practical reasons, these were split into two sub-dimensions (30 indicators at the interpersonal level and 30 indicators at the organizational level). Four relevant indicators at the interpersonal level were identified: taking initiative (51%), accepting and learning from feedback (48%), cooperating with others (45%), and communicating effectively (45%). Four relevant indicators at the organizational level were also identified: showing responsibility (67%), being customer-oriented (42%), being creative (41%), and taking on challenging work tasks (40%).

There were differences between experts from different professional backgrounds on two of these indicators. On average, managers found taking initiative significantly more important (75%) than researchers (41%), human resources managers (53%), and occupational health professionals (50%). Researchers found being customer-oriented significantly less important (21%) than managers (57%), human resources managers (69%), and occupational health professionals (47%).

Adaptive performance

The adaptive performance dimension originally included 18 indicators. Six relevant adaptive performance indicators were identified: showing resiliency (coping with stress, difficult situations and adversities; 70%), coming up with creative solutions to novel, difficult problems (66%), keeping job knowledge up-to-date (57%), keeping job skills up-to-date (52%), dealing with uncertain and unpredictable work situations (48%), and adjusting work goals when necessary (43%).

Experts from different professional backgrounds differed on two of these indicators. Managers found coming up with creative solutions to novel, difficult

problems significantly more important (82%) than human resources managers (57%) and occupational health professionals (58%). Researchers found keeping job skills up-to-date significantly more important (64%) than managers (41%), human resources managers (43%), and occupational health professionals (42%).

Counterproductive work behavior

Initially, there were 24 indicators included in counterproductive work behavior. Four relevant indicators were identified: displaying excessive negativity (62%), doing things that harm your organization (54%), doing things that harm your co-workers or supervisor (52%), and purposely making mistakes (48%).

Experts from different professional backgrounds differed on three of these indicators. Occupational health professionals found displaying excessive negativity significantly more important (86%) than researchers (57%), managers (55%), and human resources managers (61%). Managers (73%) and occupational health professionals (66%) found doing things that harm your organization significantly more important than researchers (44%). Last, human resources managers found purposely making mistakes significantly less important (31%) than researchers and occupational health professionals (both 54%).

Weight of dimensions in IWP ratings

Table 3 presents the relative weight that experts assigned to each of the IWP dimensions. On average, task performance received the heaviest weight when rating an employee's work performance (36 points). Contextual performance (22 points), adaptive performance (20 points), and counterproductive work behavior (21 points) received almost equal weightings. Experts from different professional backgrounds differed significantly on the mean weight they assigned to task performance ($F(3,225) = 3.318$; $p < 0.05$). Researchers (39 points) assigned a marginally significant ($p = 0.058$) greater weight to task performance than managers (33 points).

Generic versus job-specific questionnaire

Forty-four percent of the experts believed it possible to develop a generic questionnaire of IWP, while 23% of the experts believed that a combination of a generic and job-specific questionnaire was more likely to be effective, whereas 33% of the experts believed that a job-specific questionnaire was more likely appropriate.

Table 2. Individual work performance indicators that were selected as most relevant for each dimension and the percentage (%) of votes they received from the total group and per profession

			Profession			
		Total (n=253)	Resear-chers (n=113)	Mana-gers (n=48)	HRM (n=54)	OHP (n=38)
Indicator		%	%	%	%	%
Task performance						
1	Work quality	69	70	67	70	66
2	Planning and organising work	56	55	56	58	58
3	Being result-oriented *	46	29 [ab]	64 [a]	66 [b]	45
4	Prioritising	45	39	42	57	47
5	Working efficiently	44	47	38	43	42
Contextual performance - interpersonal						
1	Taking initiative *	51	41 [a]	75 [abc]	53 [b]	50 [c]
2	Accepting and learning from feedback	48	50	34	55	50
3	Cooperating with others	45	41	48	47	50
4	Communicating effectively (e.g., adequately expressing ideas and intentions)	45	43	48	45	50
Contextual performance - organizational						
1	Showing responsibility	67	67	70	35	67
2	Being customer-oriented *	42	21 [abc]	57 [a]	69 [bd]	47 [cd]
3	Being creative	41	44	41	39	33
4	Taking on challenging work tasks	40	45	41	25	42
Adaptive performance						
1	Showing resiliency (coping with stress, difficult situations and adversities)	70	71	70	73	67
2	Coming up with creative solutions to novel, difficult problems *	66	67	82 [ab]	57 [a]	58 [b]
3	Keeping job knowledge up-to-date	57	59	50	57	56

Table 2. Continued

Adaptive performance						
4	Keeping job skills up-to-date *	52	64 [abc]	41 [a]	43 [b]	42 [c]
5	Dealing with uncertain and unpredictable work situations	48	41	64	53	44
6	Adjusting work goals when necessary	43	42	48	37	47
Counterproductive work behavior						
1	Displaying excessive negativity (e.g., complaining, making problems bigger than they are) *	62	57 [a]	55 [b]	61 [c]	86 [abc]
2	Doing things that harm your organization (e.g., not following rules, discussing confidential information) *	54	44 [ab]	73 [ac]	51 [c]	66 [b]
3	Doing things that harm your co-workers or supervisor (e.g., arguing, leaving work for others to finish)	52	58	43	47	54
4	Purposely making mistakes *	48	54 [a]	50	31 [ab]	54 [b]

Notes: * = significant difference between profession subgroups.
[abcd] = Denote which subgroups significantly differed from each other, for example, in item 3 of task performance the score of the researchers (29%) differed statistically significant from the score of the managers (64%) and HRM (66%).

Table 3. The relative weight (scale 0 – 100) of each dimension in IWP ratings, in total and per profession

		Profession			
Dimension	**Total** (n=253) Mean (SD)	Researchers (n=113) Mean (SD)	Managers (n=48) Mean (SD)	HRM (n=54) Mean (SD)	OHP (n=38) Mean (SD)
Task performance *	36 (13)	39 (15) [a]	33 (10) [a]	35 (13)	33 (10)
Contextual performance	22 (8)	22 (8)	23 (8)	23 (8)	23 (7)
Adaptive performance	20 (8)	19 (8)	20 (6)	22 (9)	22 (10)
Counterproductive work behavior	21 (13)	20 (12)	24 (11)	20 (15)	22 (13)
Total	100	100	100	100	100

Notes: * = significant difference between profession subgroup.
[a] Marginally significant difference (p = 0.058).

Discussion

The main goal of the current study was to gain consensus on how to measure IWP, which would enable the development of a standardized, generic, short instrument. Four broad, generic dimensions of IWP were used as a theoretical basis: task performance, contextual performance, adaptive performance, and counterproductive work behavior. Using a multi-disciplinary approach, possible employee behaviors or actions (indicators) were identified for each dimension, via a review of the literature, existing questionnaires, and data from interviews with experts. In total, 128 unique IWP indicators were identified, of which 23 were considered most relevant for measuring IWP, based on notable consensus among experts. On average, task performance received greatest weight when rating an employee's work performance. Contextual performance, adaptive performance, and counterproductive work behavior received almost equal weightings. There was agreement on 84% of the indicators between experts from different professional backgrounds. Furthermore, experts agreed on the relative weight of each IWP dimension in rating work performance. However, researchers weighed task performance slightly higher than managers. Almost half of the experts believed in the possibility of developing a completely generic questionnaire of IWP.

A multitude of measurement instruments aiming to measure IWP (or a similar construct such as presenteeism or productivity) were identified in a systematic search. Considering the large number of questionnaires (81), it is not

surprising that most IWP indicators were identified from questionnaires. Far more indicators were identified for contextual performance than for the other dimensions of IWP, although contextual performance was not rated higher than other dimensions. As task performance rated the highest; one could expect more indicators to be found for that dimension. This finding may indicate that task performance is a less complex and more uniform dimension to measure than contextual performance. However, it may also indicate that it is harder to think of generic behaviors for task performance than for contextual performance. This may mean that many task performance items are job-specific.

In the literature and questionnaire reviews, an indicator often (if not most often) used for assessing task performance was quantity of work [7]. Surprisingly, quantity of work was not selected as one of the most important indicators of task performance in the current study. In fact, it was selected by only 13% of the participants as an important indicator of IWP. This finding could be due to our sample containing relatively few participants from trade and industrial work. Alternatively, it could be due to the fact that quantity of work is captured in being result-oriented. While being result-oriented was not mentioned in the literature or questionnaires, it was selected as an important indicator for task performance in the current study, mainly by managers and human resources managers. These findings indicate that it may be more important to look at other indicators than work quantity to assess task performance, such as work quality or being result-oriented.

Based on the current findings, some of the most often used IWP scales do not incorporate all relevant indicators, or incorporate irrelevant indicators. Scales often used to assess contextual performance include for example Podsakoff and MacKenzie [13] or Van Scotter and Motowidlo [14]. The former focused on measuring altruism, conscientiousness, sportsmanship, courtesy, and civic virtue. The latter focused on measuring interpersonal facilitation and job dedication. The first dimension of both scales is mainly operationalized by helpful behaviors, such as helping others who have heavy work loads. In the current study, helping others was not one of the most relevant behaviors for contextual performance (selected by 16% of participants). Two of the top three contextual performance behaviors identified in the current study (showing responsibility and accepting and learning from feedback) are not directly included in either of these questionnaires.

Adaptive performance is a new and upcoming dimension in the work performance literature [7]. Except for the Job Adaptability Index developed by Pulakos, Arad, Donovan, and Plamondon [30], few researchers have identified

indicators or developed measures of adaptive performance. Adaptive performance behaviors, such as resiliency, coming up with creative solutions to difficult, novel problems, and keeping job skills and job knowledge up-to-date, were found to be very relevant for work performance in the current study. These types of behaviors can scarcely be found in existing literature or questionnaires. The aforementioned findings may represent new and contemporary developments in the concept of work performance or in today's job requirements.

Scales often used to assess counterproductive work behavior include for example Bennett and Robinson [15] or Spector et al. [31]. The former authors focus on measuring deviance directed at the organization (organizational deviance) and deviance directed at members of the organization (interpersonal deviance). The latter authors focus on measuring sabotage (e.g., damaging company equipment), withdrawal (e.g., taking longer breaks), production deviance (e.g., doing work incorrectly), theft (e.g., stealing company property), and abuse (e.g., making fun of someone at work). In the literature, these behaviors have also often been used to describe counterproductive work behavior [7]. This is largely in line with the findings of our current study, where displaying excessive negativity, doing things that harm your organization, doing things that harm your co-workers or supervisor, and purposely making mistakes, were found to be the most important indicators of counterproductive work behavior.

Strengths and limitations

The present study has several strengths. To our knowledge, this is the first study that has systematically identified the numerous indicators used for measuring IWP. Indicators of IWP were derived from reviews of the literature, questionnaires, and interviews with experts from multiple disciplines. In addition, during the selection of indicators, participants were given the opportunity to suggest additional items. This minimized the chance of missing indicators. Also, this is one of the first studies that aimed to gain consensus on which indicators were most important for measuring IWP. In selecting the most important indicators, not only researchers, but also stakeholders from practice were involved, which improved the practical representativeness and applicability of the findings.

The present study has some limitations. The first and second author categorized each indicator into one of the generic IWP dimensions from the list of 128 unique IWP indicators. This categorization may not be valid for every job, as the place of an indicator may differ depending on the context. For example, in some jobs,

communicating effectively may be an aspect of contextual performance, while in others jobs it may be an aspect of task performance. Future research will need to determine whether the indicators belong to the expected dimensions, and whether this categorization is similar over jobs. Far more indicators were identified for contextual performance than for the other dimensions. For practical reasons, it was decided to split the contextual performance indicators into two sub-dimensions (30 indicators at the interpersonal level and 30 indicators at the organizational level). From each sub-dimension, experts were asked to select the six most important indicators. This may have resulted in an overrepresentation of indicators from one of these sub-dimensions.

In addition, the extent to which participants were able to think on a generic level when selecting indicators is debatable. Participants were asked to select the most important indicators keeping in mind all types of professions, not just in their own profession. However, the extent to which participants were able to transcend their own company or field of work remains questionable. Despite the high similarities in ratings between professions, there were also some differences. For example, researchers found being customer-oriented significantly less important than other experts and managers found taking initiative significantly more important than other experts. This finding may indicate that, to a certain extent, the relevance of an indicator may depend on the job being evaluated or the evaluator. In general, however, the high agreement (84% or higher) between all subgroups indicates good generalizability of the findings over jobs. Nevertheless, future research should further establish whether the relevance of indicators is similar across jobs, as well as across companies and countries. Furthermore, we only looked at the number of times an indicator was selected as relevant and we did not ask the participants to rate each indicator on importance. The latter would have been too time-consuming for participants. Therefore, we cannot be sure that the most frequently selected indicator in our study was also the indicator judged to be most important by participants. However, it seemed reasonable to assume that an indicator selected by more than 40% of the participants is an important indicator for the measurement of IWP.

The decision to deem indicators relevant when selected by more than 40% of the participants may seem somewhat arbitrary. This decision was made for two reasons. First, this was based on a graph of the data, where 40% seemed to be a natural cut-off point. Second, ideally there should be a minimum of three items contributing to one dimension [32]. What would happen if the cut-off point was set

at 50% or 30% was also examined. For example, when using a 50% cut-off point, only two relevant indicators for task performance remained. When using a 30% cut-off point, 9 relevant indicators for task performance appeared. Therefore, to construct a short, but comprehensive questionnaire, a cut-off point of 40% was deemed to be sufficient.

Conclusion

Research on IWP would benefit greatly from a standardized, generic, short measurement instrument. In the current study, 128 unique IWP indicators were identified, of which 23 indicators were considered most relevant for measuring IWP, based on notable consensus among experts. This provides an important step towards the development of a standardized, generic, short instrument. Hopefully, results of the current study remove some of the uncertainty regarding the definition and measurement of IWP, and brings us one step closer to unraveling IWP and its causes and consequences.

References

1. Escorpizo R. Understanding work productivity and its application to work-related musculoskeletal disorders. International Journal of Industrial Ergonomics. 2008; 38: 291-297.
2. Schultz AB, Chen CY, Edington DW. The cost and impact of health conditions on presenteeism to employers. Pharmacoeconomics. 2009; 27: 365-378.
3. Cancelliere C, Cassidy JD, Ammendolia C, Cote P. Are workplace health promotion programs effective at improving presenteeism in workers? a systematic review and best evidence synthesis of the literature. BMC Public Health. 2011; 11: 395.
4. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol. 2005; 90: 1241-1255.
5. Campbell J.P. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, editors. Handbook of industrial and organizational psychology. Palo Alto, CA, US: Consulting Psychologists Press; 1990, Vol.1 (2nd ed.). p. 687-732.
6. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: a policy-capturing approach. J Appl Psychol. 2002; 87(1): 66-80.
7. Koopmans L, Bernaards CM, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual frameworks of individual work performance - A systematic review. Journal of Occupational and Environmental Medicine. 2011; 53(8): 856-866.
8. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, editors. Personnel Selection in Organizations San Francisco, CA: Jossey Bass; 1993. p. 71-98.
9. Griffin MA, Neal A, Parker SK. A new model of work role performance: positive behavior in uncertain and interdependent contexts. Academy of Management Journal. 2007; 50: 327-347.
10. Koopman C, Pelletier KR, Murray JF, et al. Stanford Presenteeism Scale: health status and employee productivity. Journal of Occupational and Environmental Medicine. 2002; 44: 14-20.

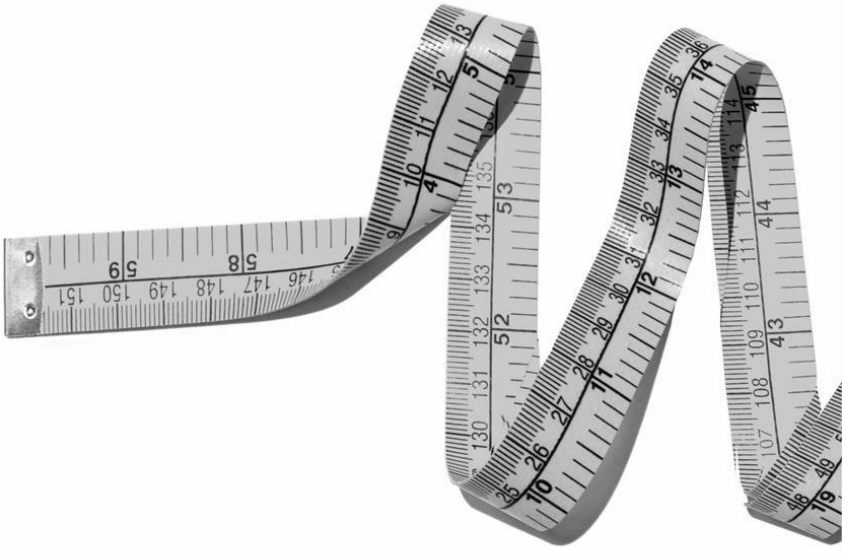
11. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproductibility of a work productivity and activity impairment instrument. Pharmacoeconomics. 1993; 4: 353-365.
12. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). Journal of Occupational and Environmental Medicine. 2003; 45: 156-174.
13. Podsakoff PM, MacKenzie SB. A second generation measure of organizational citizenship behavior. Indiana University, Bloomington; 1989.
14. Van Scotter JR, Motowidlo SJ. Interpersonal facilitation and job dedication as separate facets of contextual performance. Journal of Applied Psychology. 1996; 81(5): 525-531.
15. Bennett RJ, Robinson SL. Development of a Measure of Workplace Deviance. Journal of Applied Psychology. 2000; 85(3): 349-360.
16. LePine J.A., Erez A., Johnson D.E. The nature and dimensionality of organizational citizenship behavior: a critical review and meta-analysis. J Appl Psychol. 2002; 87(1): 52-65.
17. Beaton D., Bombardier C., Escorpizo R., Zhang W., Lacaille D., Boonen A., et al. Measuring worker productivity: frameworks and measures. J Rheumatol. 2009; 36(9): 2100-2109.
18. Evans C.J. Health and work productivity assessment: state of the art or state of flux? J Occup Environ Med. 2004; 46(6): S3-S11.
19. Hunt S.T. Generic work behavior: an investigation into the dimensions of entry-level, hourly job performance. Personnel Psychology. 1996; 49(1): 51-83.
20. Mattke S, Balakrishnan A, Bergamo G, Newberry SJ. A review of methods to measure health-related productivity loss. The American Journal of Managed Care. 2007; 13(4): 211-217.
21. Baarda, De Goede, Van Dijkum. Basisboek Statistiek met SPSS. Groningen/Houten, The Netherlands: Wolters-Noordhoff bv; 2003.
22. Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. Journal of Management. 1991; 17(3): 601-617.
23. Karidi MV, Papakonstantinou K, Stefanis N, Zografou M, Karamouzi G, Skaltsi P, et al. Occupational abilities and performance scale: Reliability-validity

assessment factor analysis. Social Psychiatry and Psychiatric Epidemiology. 2005; 40(5): 417-424.

24. Griffiths RD. A standardized assessment of the work behavior of psychiatric patients. British Journal of Psychiatry. 1973; 123: 403-408.
25. Lerner D, Amick BC, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Medical Care. 2001; 39(1): 72-85.
26. Endicott J, Nee J. Endicott Work Productivity Scale (EWPS): A new measure to assess treatment effects. Psychopharmacol Bull. 1997; 33(1): 13-16.
27. Stewart WF, Ricci JA, Leotta C, Chee E. Validation of the work and health interview. Pharmacoeconomics. 2004; 22: 1127-1140.
28. Wang G, Netemeyer RG. Salesperson creative performance: Conceptualization, measurement, and nomological validity. Journal of Business Research. 2004; 57(8): 805-812.
29. Greenslade J.H., Jimmieson N.L. Distinguishing between task and contextual performance for nurses: Development of a job performance scale. J Adv Nurs. 2007; 58(6): 602-611.
30. Pulakos ED, Arad S, Donovan MA, Plamondon KE. Adaptability in the Workplace: Development of a Taxonomy of Adaptive Performance. Journal of Applied Psychology. 2000; 85(4): 612-624.
31. Spector P.E., Fox S., Penney L.M., Bruursema K., Goh A., Kessler S. The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? J Vocat Behav. 2006; 68: 446-460.
32. De Vet H, Terwee C, Mokkink L, Knol D. Measurement in Medicine. Cambridge University Press; 2011.

Part II

Field-testing Phase



Chapter 4

Development of an Individual Work Performance Questionnaire

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt,
Stef van Buuren, Allard J. van der Beek, Henrica C.W. de Vet

International Journal of Productivity and Performance Management.
2013; 62(1):6-28

DOI 10.1108/17410401311285273

4

Abstract

Objective: The goal of the current study was to develop a generic and short questionnaire to measure work performance at the individual level – the Individual Work Performance Questionnaire (IWPQ). The IWPQ was based on a four-dimensional conceptual framework, in which individual work performance consisted of task performance, contextual performance, adaptive performance, and counterproductive work behavior.

Methods: After pilot-testing, the 47-item IWPQ was field-tested amongst a representative sample of 1,181 Dutch blue, pink, and white collar workers. Factor analysis was used to examine whether the four-dimensional conceptual framework could be confirmed. Rasch analysis was used to examine the functioning of the items in more detail. Finally, it was examined whether generic scales could be constructed.

Results: A generic, three-dimensional conceptual framework was identified, in which individual work performance consisted of task performance, contextual performance, and counterproductive work behavior. Generic, short scales could be constructed that fitted the Rasch model.

Conclusion: A generic, short questionnaire can be used to measure individual work performance across occupational sectors. In future versions of the IWPQ, more difficult items should be added to improve discriminative ability at the high ranges of the scale.

Introduction

Individual work performance (IWP) is a relevant and often used outcome measure of studies in the occupational setting. In the past decades, a great deal of research in fields such as management, occupational health, and industrial-organizational psychology has been devoted to discovering the predictors and effects of IWP. Only later did attention arise for defining the construct of IWP and understanding its underlying structure [e.g., 1, 2]. After all, a clear definition and theoretical framework of IWP is a prerequisite for valid measurement of the construct.

IWP was defined by Campbell [3] as "behaviors or actions that are relevant to the goals of the organization." Thus, IWP focuses on behaviors or actions of employees, rather than the results of these actions. In addition, behaviors should be under the control of the individual, thus excluding behaviors that are constrained by the environment [1]. In order to measure IWP, it is important to determine its underlying structure. Traditionally, the main focus of the IWP construct has been on *task performance*, which can be defined as the proficiency with which individuals perform the core substantive or technical tasks central to his or her job [3]. Behaviors used to describe task performance often include work quantity and quality, job skills, and job knowledge [1, 3].

Although it has long been recognized that IWP is a multidimensional construct [3, 4], only more recently has the role of employee behaviors beyond task performance received full attention [e.g., 1, 2, 5]. It is now generally agreed upon that, in addition to task performance, the IWP domain consists of *contextual performance* and *counterproductive work behavior* [e.g., 1, 6]. Contextual performance can be defined as behaviors that support the organizational, social and psychological environment in which the technical core must function [5]. Behaviors used to describe contextual performance include for example demonstrating effort, facilitating peer and team performance, cooperating, and communicating [1, 3]. Counterproductive work behavior can be defined as behavior that harms the well-being of the organization [1]. It includes behaviors such as absenteeism, off-task behavior, theft, and substance abuse [7].

A recent review by Koopmans et al. [7] has identified the new and upcoming dimension of *adaptive performance* in IWP frameworks [e.g., 8-10]. This dimension focuses on the growing interdependency and uncertainty of work systems and the corresponding change in the nature of IWP. Adaptive performance can be defined as

the extent to which an individual adapts to changes in the work role or environment [10].

Numerous scales have been developed to measure the dimensions of IWP. For example, Williams and Anderson [11] developed a short and generic task performance scale, which measured behaviors such as adequately completing assigned duties, fulfilling prescribed responsibilities, and performing tasks that are expected of the employee. Scales used to assess contextual performance are those developed by for example Podsakoff and MacKenzie [12] or Van Scotter and Motowidlo [13]. The former focuses on measuring altruism, conscientiousness, sportsmanship, courtesy, and civic virtue. The latter focuses on measuring interpersonal facilitation and job dedication. Scales used to assess counterproductive work behavior were developed by for example Bennett and Robinson [14] or Spector et al. [15]. The former authors focus on measuring organisational and interpersonal deviance The latter authors focus on measuring sabotage (e.g., damaging company equipment), withdrawal (e.g., taking longer breaks), production deviance (e.g., doing work incorrectly), theft (e.g., stealing company property), and abuse (e.g., making fun of someone at work). A scale developed to measure adaptive performance is the Job Adaptability Index (JAI) by Pulakos et al. [8]. It measures, for example, whether employees are able to solve problems creatively, to deal with uncertain or unpredictable work situations, and to learning new tasks, technologies and procedures.

Several limitations can be observed in the scales developed to measure dimensions of IWP. Most strikingly, none of these scales measure all dimensions of IWP together. As a result, they fail to incorporate the complete range of individual behaviors at work. This requires the researcher to search for, compare, and combine different scales to get a complete picture of IWP.

The task of deciding which scale(s) to use, is complicated by the fact that scales often operationalize the same dimension differently. This entrusts the researcher with the difficult task of deciding which operationalization is most appropriate and relevant for his or her study population. The different operationalizations are partly due to different conceptualizations of the dimensions, and partly due to specific populations being used to develop and refine the scales. For example, the task performance scale by Williams and Anderson [11] was based on a sample of employees with a technical/professional background, and the contextual performance scales by Podsakoff and MacKenzie [12] or Van Scotter and

Motowidlo [13] were based on a sample of petrochemical employees and U.S. Airforce mechanics, respectively.

The use of separate scales to measure the dimensions of IWP has given rise to another problem, namely that of *antithetical items* [2]. That is, items overlapping in content can be found in scales measuring different dimensions. This is especially the case for contextual performance and counterproductive work behavior scales. Many contextual performance scales include counterproductive behaviors (e.g., "Takes undeserved work breaks") that are reverse scored, and some counterproductive scales include functional behaviors (e.g., "Volunteers to finish a project for a coworker who is sick") that are reverse scored. However, contextual performance and counterproductive work behavior are not the opposite ends of one scale. The absence of counterproductive behaviors is not identical to good contextual performance, and likewise, the presence of functional behaviors is not identical to low counterproductivity. The inclusion of antithetical items is problematic because it magnifies the strength of the correlation between contextual and counterproductive scales, and perhaps more importantly, reduces the content validity of the scales.

The goal of the current study was to develop a generic and short questionnaire of IWP – the Individual Work Performance Questionnaire (IWPQ) – that overcomes the previously mentioned limitations. This questionnaire measures all IWP dimensions, has a standardized operationalization that is developed and refined based on a generic population, and includes no antithetical items. The methods section discusses the developmental process of the IWPQ. It describes the field-testing in a generic population and the analysis of the resultant data. The results section presents the results of the field-testing and the construction of the generic, short IWPQ. Subsequently, the most important findings are discussed, strengths and limitations of the research are addressed, and avenues for future research are proposed. Finally, the conclusions support the use of a generic, short questionnaire of IWP.

Methods

Individual Work Performance Questionnaire (IWPQ)

The IWPQ version 0.1 was based on a four-dimensional conceptual framework, in which IWP consists of four dimensions: task performance, contextual performance, adaptive performance, and CWB [7]. For each dimension, one scale was developed. The operationalization of the scales was based on a study by Koopmans et al. [16]. In this study, all possible indicators of the IWP dimensions were first identified from the literature, existing questionnaires, and expert interviews. Antithetical items were removed from the dimensions. This resulted in a list of 128 unique indicators of IWP. Subsequently, agreement amongst experts from different professional backgrounds and countries was reached on the most relevant, generic indicators per IWP dimension. The 23 relevant, generic indicators were included in the IWPQ scales. In addition, the task performance scale included work quantity as a relevant indicator. Although it was not selected as one of the most relevant indicators [16], for theoretical reasons we considered this an essential indicator of IWP. For each indicator, one to three questionnaire items were chosen, resulting in the 47-item IWPQ (Table 2). The task performance scale consisted of 13 questionnaire items (e.g.: *"How do you rate the quality of your own work?"*), contextual performance of 16 (e.g.: *"I came up with creative ideas at work"*), adaptive performance of 8 (e.g.: *"I have demonstrated flexibility"*), and CWB of 10 (e.g.: *"I complained about unimportant matters at work"*).

Pilot-testing

A pilot study amongst 54 researchers was conducted to optimize clarity, readability, and face validity of the IWPQ. The 54 researchers were employees of TNO (Netherlands Organization for Applied Scientific Research) and VU University medical center. In addition, think-aloud protocols were held with six persons (three researchers, one secretary, one nurse, and one manager). Based on the findings, clarity and readability of the items were improved. One main revision was reducing the answer categories from seven to five categories, as participants indicated that the differences between some answer categories were unclear. Another main revision was extending the recall period from 4 weeks to 3 months, to assure that most situations had likely taken place, and including a *"not applicable"* answer category for some questions, as many participants indicated that a situation may not have taken place in the past 4 weeks. To assess face validity, participants were asked

whether they thought the questionnaire actually measured IWP, whether any questions were redundant, and whether any important questions were missing. Most participants indicated that the face validity of the IWPQ was good. As a final check, the VU University Language Center screened the full questionnaire for readability and correct use of language.

Recall period and rating scales

All items had a recall period of 3 months and a 5-point rating scale. Rating scale labels were adapted to the specific item. Quality and quantity of work was rated from *"insufficient"* to *"very good"* (items 1 and 4), quality and quantity of work compared to last years was rated from *"much worse"* to *"much better"* (items 2 and 5), and decreased quality and quantity of work was rated from *"never"* to *"often"* (items 3 and 6). On the remaining items, participants rated the frequency of their behavior. Frequency ratings were preferred over agreement ratings, because agreement ratings generally require individuals to rate whether he or she is likely to engage in each behavior, and may assess attitude toward the behavior rather than actual behavior [2]. Frequency ratings require individuals to recall and mentally calculate how often one engaged in each behavior [17], and were therefore considered to be more valid. A problem with self-ratings of performance is that persons are inclined to judge their own performance favorably (the *leniency effect* [18]), and this produces ceiling effects in the scales. As a result, detecting improvement or distinguishing among high levels of performance is almost impossible. One method to counteract this effect is to shift the center of the scale, so that the average point is not in the middle but rather to the left of the scale [19]. For these reasons, the remaining task, contextual, and adaptive behaviors (items 7 to 38) were rated from "*seldom," "sometimes," "frequently," "often,"* to *"always."* As the counterproductive behaviors (items 39 to 49) were expected to produce floor rather than ceiling effects, the center of this scales was shifted to the right, ranging from *"never," "seldom," "sometimes," "frequently,"* and *"often."*

Field-testing

The IWPQ was tested in a study amongst a representative sample of 1,181 Dutch workers. An internet panel organization recruited the respondents. The internet panel consisted of Dutch adults who were willing to participate in research projects in exchange for a small financial reward. First, respondents filled out their gender, age, education, and type of occupation. Second, they completed the 47-item IWPQ.

Finally, respondents rated the understandability of the IWPQ and the applicability of the IWPQ to their occupation on a 5-point scale ranging from *"bad"* to *"very good."*

Data analysis of the field-test

Understandability and applicability

In order to determine whether participants found the IWPQ items understandable, and applicable to their occupation, the mean score and standard deviation on these questions were calculated. One-way analyses of variance were performed to examine whether there were differences between occupational sectors in understandability or applicability. Post-hoc tests with Bonferroni correction were performed to determine which occupational groups differed from each other.

Conceptual framework

In order to test whether the four-dimensional conceptual framework could be confirmed across occupational sectors, factor analysis (principal components) with varimax rotation was performed in SPSS 17. Beforehand, task performance items 3, 6, 10, and 13, and CWB items 1 to 10 were coded reversely (0 as 4, 1 as 3, 2 as 2, 3 as 1, 4 as 0) so that a low score meant low work performance and a high score meant high work performance. Fourteen IWPQ items had a *"not applicable"* category, which was entered as a missing value. During factor analysis, missing values were substituted by the mean value of an item, so that no individuals had to be deleted from the analysis. Score ranges of the items were examined for floor or ceiling effects (> 15% at the extreme values [20]). Also, inter-item correlations were examined. Items that correlate very low (<0.20) with all other items are problematic because they have no relationship to any other items, and should be deleted. Items that correlate very high (>0.90) with another item should also be considered carefully because they are almost identical to the other item, and one may be deleted.

The Kaiser-Meyer-Olkin's (KMO) Measure of Sampling Adequacy (should be >0.50) and Bartlett's Test of Sphericity (should be <0.05) were performed to test whether the variables in the dataset were sufficiently correlated to apply factor analysis. The results of the factor analysis were used to construct unidimensional scales. The factor loadings determined which items were retained in a scale. Items loading high on a factor (>0.40) for all occupational sectors, were retained. Prerequisite was that items loaded high on only one factor, as overlapping items hinder interpretation and scoring of factors.

Rasch analysis

To examine the functioning of the items in more detail, each scale was examined using Rasch analysis [21], a specific type of Item Response Theory (IRT). The analysis was performed separately for each scale, because Rasch analysis must be performed on a unidimensional scale. In comparison with Classical Test Theory (CTT), the Rasch model assesses a wider range of measurement properties, increasing the information available about a scale's performance [22, 23]. For example, Rasch analysis provides information on item difficulty (items are hierarchically ordered based on difficulty, expecting that if a person with a certain ability scores well on a difficult item, then that person scores well on easier items as well), response category ordering (does the category ordering of polytomous items work as expected), and differential item functioning (DIF; do subgroups in the sample respond differently to items). Analyses were conducted using RUMM2020 software [24].

- *Model fit*

  Data fit the Rasch model when observed responses are equivalent or do not greatly differ from responses expected by the Rasch model. The following fit statistics test model fit: 1) Chi-square fit, 2) item fit residuals, and 3) person fit residuals. The Chi-square fit statistic is an item-trait interaction score, reflecting the property of invariance across the trait. Generally, a nonsignificant Chi-square fit statistics indicates model fit. However, this statistic is highly sample size dependent, and in large samples it is almost certain to show significance because of the power of the test [25, 26]. RUMM2020 provides the option to reduce the sample by randomly selecting a specified number of persons from the existing sample. Therefore, model fit for the total sample was also tested by setting the sample size at 200 [27]. Item and person fit residuals represent the residuals between the observed and expected values for items and persons. Ideally, these should have a mean of approximately 0 and an SD of 1 [23].

- *Reliability*

  The person separation index (PSI) estimates the internal consistency of a scale. PSI is similar to Cronbach's alpha [28], only it uses the logit scale estimates as opposed to the raw scores. It is interpreted in a similar manner, that is, a minimum value of 0.70 is required for group use and 0.85 for individual use [23].

- *Improving fit*
  Multiple statistics determine which items should be removed to improve fit of a scale. Items with a high fit residual (> 2.5) are first candidates for deletion. Second, items with inadequate targeting are candidates for deletion. Third, items with a low slope are candidates for deletion, because they discriminate poorly between persons with low and high work performance. Furthermore, the content of the items is taken into account, making sure to retain items with important content. Item reduction is an iterative process, in which one item is removed at a time and fit re-estimated accordingly [20].

- *Category ordening*
  In addition to good model fit, the data has to satisfy several assumptions of the Rasch model. For one, Rasch analysis assumes that when using polytomous answer categories, a higher category reflects an increase in the underlying ability. If appropriate category ordening does not occur, the thresholds between adjacent answer categories are disordered [23].

- *Differential Item Functioning*
  Rasch analysis assumes that a scale functions consistently, irrespective of subgroups within the sample being assessed. Differential Item Functioning (DIF) affects model fit when different groups within the sample respond in a different manner to an item, despite equal levels of the underlying characteristic being measured [23].

- *Local independence*
  Rasch analysis assumes that the response to an item is independent of responses to other items, after controlling for the person's ability. When the answer to one item determines the answer to another item, there is a breach in local independence. Such breaches are identified through the residual correlation matrix, by looking for residual correlations ≥ 0.40. Local independence is often used to give an indication of unidimensionality of a scale [23].

- *Targeting of the scales*
  The person-item threshold map reveals the location of the persons and the items on a linear scale that runs from -5 to +5, with 0 being the average item

difficulty. This indicates how well targeted the items are for persons in the sample [23]. An equal distribution of items is desired if the instrument has to discriminate between persons at various ranges on the scale. Examination of the distribution of the items over the scale shows whether there is scarceness of items, i.e. gaps at certain locations on the scale.

Results

Participants

1,181 Dutch workers filled in the 47-item IWPQ in June 2011. Participants were all employed, and aged 18 to 65+ years. Almost half of the participants (49.5%) was female. The sample consisted of blue collar workers (manual workers, e.g.: carpenter, mechanic, truck driver), pink collar workers (service workers, e.g.: hairdresser, nurse, teacher), and white collar workers (office workers, e.g.: manager, architect, scientist). The specific jobs were classified into occupational sectors based on the Standard Jobs Classification of Statistics Netherlands (CBS). Table 1 presents further participant characteristics.

Table 1. Gender, age and education level of the 1,181 participants

		Occupational sector		
	Total sample	**Blue collar**	**Pink collar**	**White collar**
N	1,181 (100%)	368 (31%)	421 (36%)	392 (33%)
Gender *(% female)*	49.5%	16.3%	79.3%	48.7%
Age				
18-24 years	6%	5%	9%	2%
25-34 years	17%	13%	16%	23%
35-44 years	27%	28%	25%	29%
45-54 years	31%	31%	32%	30%
55-64 years	18%	22%	18%	16%
65+ years	1%	1%	0%	1%
Education level				
Primary education	1%	1%	1%	0%
Secondary education	30%	48%	34%	9%
Middle-level applied education	32%	39%	40%	17%
Higher professional education	37%	10%	25%	74%
Unknown	1%	2%	1%	0%

Understandability and applicability

Participants rated the understandability of the items as good to very good (M=3.2, SD=0.6 on a 0-4 scale). Blue collar workers (M=3.2, SD=0.7) found the items slightly less understandable than pink (M=3.3, SD=0.6) and white collar workers (M=3.3, SD=0.7), $F(2,1178) = 4.037$, $p < 0.05$. However, this difference is too small to be considered practically relevant. Participants rated the applicability of the items to their occupation as reasonable to good (M=2.6, SD=0.9 on a 0-4 scale). There were no differences between occupational sectors regarding the applicability of the items to their occupation, $F(2,1178) = 2.071$, $p > 0.05$.

Conceptual framework

Thirty-eight of the 47 items showed ceiling effects, i.e. more than 15% of the responses at the high end of the scale. Especially CWB items (recoded) showed ceiling effects, ranging up to 96.6% of the scores at the extreme value. None of the items showed very low (>0.20) or very high (>0.90) inter-item correlations. Fourteen items had a *"not applicable"* category, which was used by 14% of the respondents, on average.

For each occupational sector, the inter-item correlations were appropriate for factor analysis, with Kaiser-Meyer-Olkin's Measure of Sampling Adequacy being >0.90, and Bartlett's Test of Sphericity showing a *p-value* <0.001. The scree plots identified three factors for blue and white collar workers, and four factors for pink collar workers. For all occupational sectors, the *task performance* scale consisted of task performance items 3, 7 to 9, 11, 12, and contextual performance items 1, 2 and 5 (see Table 2). In addition, contextual performance items 4 and 6 were retained for blue collar workers. Task performance items 1, 2, 4 and 13 were retained for pink collar workers. Task performance items 1, 3, 6, 13, and contextual performance item 3, 4 and 6 were retained for white collar workers. For all occupational sectors, the *contextual performance* scale consisted of contextual performance items 7 to 10, 12 to 14, and adaptive performance items 1 to 8. In addition, contextual performance item 15 was retained for white collar workers. For blue and white collar workers, the *counterproductive* scale consisted of CWB items 1 to 10. For pink collar workers, this scale was split into two factors: a *minor CWB* factor (items 1 to 5), and a *serious CWB* factor (items 6 to 10).

Table 2. Raw mean scores (M) and standard deviations (SD) on the understandability and applicability items, and the Individual Work Performance Questionnaire (IWPQ) items (on a 0-4 scale)

Items		Rating scale (0 – 4)	Total sample M (SD)	Occupational sector		
				Blue Collar M (SD)	Pink Collar M (SD)	White collar M (SD)
Understandability and applicability						
1	How understandable were the questions?	*bad – very good*	3.2 (0.6)	3.2 (0.7)	3.3 (0.6)	3.3 (0.7)
2	How appropriate were the questions for your occupation?	*bad – very good*	2.6 (0.9)	2.5 (0.9)	2.6 (0.9)	2.6 (0.9)
Dimension: Task performance						
TP1	How do you rate the quality of your own work in the past 3 months?	*insufficient – very good*	3.0 (0.7)	3.1 (0.7) [1]	3.0 (0.7) [2]	3.0 (0.7) [2]

Table 2. Continued.

TP2	Compared to last year, I judge the quality of my work in the past 3 months to be…	*much worse – much better*	2.4 (0.6)	2.3 (0.5) [1]	2.4 (0.7) [2]	2.4 (0.7) [1]
TP3	How often was the quality of your work below what it should have been in the past 3 months?	*never – often*	1.0 (0.7)	0.9 (0.7) [2]	0.9 (0.7) [2]	1.1 (0.7) [2]
TP4	How do you rate the quantity of your own work in the past 3 months?	*insufficient – very good*	3.0 (0.9)	3.0 (0.8) [1]	3.0 (0.8) [2]	2.9 (0.9) [1]
TP5	Compared to last year, I judge the quantity of my work in the last 3 months to be…	*much worse – much better*	2.4 (0.8)	2.4 (0.7) [1]	2.4 (0.8) [1]	2.5 (0.8) [1]
TP6	How often was the quantity of your work less than it should have been in the past 3 months?	*never – often*	1.0 (0.9)	0.9 (0.9) [1]	1.0 (0.9) [1]	1.0 (0.9) [2]
TP7 *	I managed to plan my work so that it was done on time.	*seldom – always*	3.0 (1.0)	3.1 (0.9)	3.0 (1.0)	2.8 (1.0)

Table 2. Continued

TP8	I worked towards the end result of my work.	*seldom – always*	3.2 (0.8)	3.3 (0.8) [2]	3.3 (0.8) [2]	3.1 (0.9) [2]
TP9 *	I kept in mind the results that I had to achieve in my work.		3.3 (0.9)	3.2 (0.9)	3.3 (0.8)	3.3 (0.8)
TP10	I had trouble setting priorities in my work.		0.8 (1.0)	0.7 (1.0) [1]	0.8 (1.0) [1]	0.9 (0.9) [1]
TP11 *	I was able to separate main issues from side issues at work.		2.8 (1.0)	2.9 (1.1)	2.9 (1.0)	2.8 (1.0)
TP12 *	I was able to perform my work well with minimal time and effort.		2.4 (1.0)	2.7 (0.9)	2.5 (1.1)	2.2 (1.0)
TP13	It took me longer to complete my work tasks than intended.		1.1 (1.0)	0.8 (0.9) [1]	0.9 (1.0) [2]	1.4 (1.0) [2]

Table 2. Continued

Items		Rating scale (0 – 4)	Total sample M (SD)	Occupational sector		
				Blue Collar M (SD)	Pink Collar M (SD)	White collar M (SD)
Dimension: Contextual performance						
CP1	I was able to meet my appointments.	*seldom - always*	3.2 (0.7)	3.2 (0.7) [3]	3.3 (0.7) [2]	3.2 (0.7) [3]
CP2	I was able to fulfil my responsibilities.		3.4 (0.7)	3.4 (0.7) [2]	3.4 (0.7) [3]	3.3 (0.7) [2]
CP3	Collaboration with others went well.		3.2 (0.7)	3.2 (0.7) [1]	3.2 (0.8) [1]	3.1 (0.7) [3]
CP4	Others understood me well, when I told them something.		3.0 (0.7)	3.1 (0.7) [3]	3.1 (0.7) [1]	3.0 (0.7) [2]
CP5	I understood others well, when they told me something.		3.1 (0.7)	3.1 (0.7) [2]	3.1 (0.7) [2]	3.0 (0.6) [2]
CP6	Communication with others led to the desired result.		2.9 (0.8)	2.9 (0.9) [3]	3.0 (0.8) [1]	2.9 (0.8) [3]
CP7	I came up with creative ideas at work.		2.3 (1.0)	2.2 (1.1) [2]	2.2 (1.1) [2]	2.4 (0.9) [2]

CP8	I took the initiative when there was a problem to be solved.	*seldom - always*	2.5 (1.0)	2.5 (1.0) [2]	2.5 (1.0) [2]	2.7 (0.9) [2]
CP9	I took the initiative when something had to be organized.		2.1 (1.1)	2.0 (1.2) [2]	2.1 (1.1) [3]	2.2 (1.0) [3]
CP10 *	I started new tasks myself, when my old ones were finished.		2.8 (1.1)	2.7 (1.1)	2.8 (1.1)	2.9 (1.0)
CP11	I asked for help when needed.		1.7 (1.1)	1.5 (1.1) [1]	1.6 (1.1) [1]	1.9 (1.1) [1]
CP12	I was open to criticism of my work.		2.6 (1.0)	2.4 (1.2) [3]	2.6 (1.1) [2]	2.6 (1.0) [3]
CP13	I tried to learn from the feedback I got from others on my work.		2.4 (1.1)	2.4 (1.1) [3]	2.6 (1.0) [2]	2.7 (1.0) [3]
CP14 *	I took on challenging work tasks, when available.		2.4 (1.1)	2.4 (1.1)	2.4 (1.1)	2.5 (1.0)
CP15	I think customers/clients/patients were satisfied with my work.		3.1 (0.6)	3.2 (0.6) [1]	3.2 (0.6) [1]	3.0 (0.6) [2]
CP16	I took into account the wishes of the customer/client/patient in my work.		3.4 (0.7)	3.4 (0.7) [1]	3.5 (0.7) [1]	3.1 (0.8) [1]

Table 2. Continued

Items		Rating scale (0 – 4)	Total sample M (SD)	Occupational sector		
				Blue Collar M (SD)	Pink Collar M (SD)	White collar M (SD)
Dimension: Adaptive performance						
AP1 *	I worked at keeping my job knowledge up-to-date.	*seldom - always*	2.1 (1.2)	2.0 (1.3)	2.2 (1.2)	2.0 (1.1)
AP2 *	I worked at keeping my job skills up-to-date.		2.4 (1.1)	2.3 (1.2)	2.4 (1.1)	2.3 (1.0)
AP3	I have demonstrated flexibility.		3.1 (0.8)	2.7 (1.0) [2]	3.2 (0.8) [2]	3.0 (0.8) [2]
AP4	I was able to cope well with difficult situations and setbacks at work.		2.6 (1.0)	1.2 (1.0)	2.6 (1.0) [2]	2.5 (0.9) [3]
AP5	I recovered fast, after difficult situations or setbacks at work.		2.7 (0.9)	2.8 (1.0) [3]	2.7 (0.9) [2]	2.7 (0.9) [3]
AP6 *	I came up with creative solutions to new problems.		2.3 (1.0)	2.3 (1.1)	2.3 (1.0)	2.4 (0.9)

Table 2. Continued

AP7	I was able to cope well with uncertain and unpredictable situations at work.		2.6 (0.9)	2.6 (1.0) [2]	2.7 (0.9) [3]	2.6 (0.9) [2]
AP8	I easily adjusted to changes in my work.		2.8 (0.9)	2.9 (1.0) [2]	2.7 (0.9) [3]	2.8 (0.9) [2]
				Occupational sector		
Items		**Rating scale (0 – 4)**	**Total sample M (SD)**	**Blue Collar M (SD)**	**Pink Collar M (SD)**	**White collar M (SD)**
Dimension: Counterproductive work behavior						
CWB1 *	I complained about unimportant matters at work.	*never - often*	1.0 (0.9)	0.9 (0.9)	0.9 (0.9)	1.2 (0.9)
CWB2 *	I made problems greater than they were at work.		0.6 (0.7)	0.5 (0.7)	0.6 (0.7)	0.8 (0.8)
CWB3 *	I focused on the negative aspects of a work situation, instead of on the positive aspects.		0.9 (0.9)	0.8 (0.8)	0.9 (0.9)	1.1 (0.9)
CWB4 *	I spoke with colleagues about the negative aspects of my work.		1.4 (1.0)	1.2 (1.0)	1.3 (1.0)	1.5 (0.9)

Table 2. Continued

CWB5 *	I spoke with people from outside the organization about the negative aspects of my work.	*never - often*	1.0 (1.0)	0.8 (0.9)	1.0 (1.0)	1.1 (1.0)
CWB6	I purposely worked slowly.		0.2 (0.6)	0.3 (0.6) [2]	0.2 (0.6) [2]	0.2 (0.6) [2]
CWB7	I purposely left my work so that someone else had to finish it.		0.1 (0.5)	0.2 (0.5) [2]	0.1 (0.4) [2]	0.1 (0.5) [2]
CWB8	I behaved rudely towards someone at work.		0.3 (0.6)	0.3 (0.6) [2]	0.2 (0.5) [2]	0.3 (0.6) [2]
CWB9	I quarrelled with my colleagues, manager, or customers.		0.3 (0.6)	0.3 (0.6) [2]	0.3 (0.6) [2]	0.3 (0.6) [2]
CWB10	I purposely made mistakes.		0.1 (0.3)	0.1 (0.3) [2]	0.0 (0.3) [2]	0.1 (0.3) [2]

[1] Item removed from the scale based on factor analysis.

[2] Item removed from the scale to improve model fit.

[3] Item removed from the scale because it was job-specific.

* Items that were included in the generic scales.

Rasch analysis of the scales per occupational sector

To examine the functioning of the items in more detail, Rasch analysis was performed for each scale, per occupational sector. After deleting misfitting items (see Table 2), all the scales showed good model fit (Table 3, analyses 1-10). For all occupational sectors, the *task performance* scale included planning and organizing work (TP7), result-oriented working (TP9), prioritizing (TP11), and working efficiently (TP12). In addition, for blue collar workers, this scale included showing responsibility (CP1), and communicating effectively (CP4 and CP6). For pink collar workers, this scale also included showing responsibility (CP2). For white collar workers, this scale also included showing responsibility (CP1), cooperating with others (CP3), and communicating effectively (CP6).

For all occupational sectors, the *contextual performance* scale included taking initiative (CP10), taking on challenging work tasks (CP14), keeping job knowledge and skills up-to-date (AP1 and AP2), and coming up with creative solutions to novel, difficult problems (AP6). In addition, for blue collar workers, this scale included accepting and learning from feedback (CP12 and CP13) and showing resiliency (AP3 and AP5). For pink collar workers, this scale also included taking initiative (CP9). For white collar workers, this scale also included taking initiative (CP9), accepting and learning from feedback (CP12 and CP13), and showing resiliency (AP4 and AP5).

For all occupational sectors, the *counterproductive* scale included displaying excessive negativity (CWB1, 2, 3), and doing things that harm your organization (CWB4 and 5). There were no sector-specific items. For all occupational sectors, the CWB items 6-10 showed a low location and slope. The person-item map revealed that all these item thresholds were located lower than any of the persons in the sample. It was therefore decided to delete the CWB items 6-10.

Table 3. Summary of Rasch analyses for the occupational sectors and for the total sample, per IWPQ scale

Analysis no., description		Item fit residual, mean ±SD	Person fit residual, mean ±SD	Item-trait total Chi-square		PSI
				X^2 (df)	p	
Blue collar workers (n=368)						
Task performance						
1	TP7, 9, 11, 12, CP1, 4, 6	0.52 ± 1.51	-0.50 ± 1.36	72.77 (63)	0.19	0.82
Contextual performance						
2	CP10, 12-14, AP1-3, 5, 6	0.57 ± 0.96	-0.36 ± 1.35	90.92 (81)	0.21	0.85
CWB						
3	CWB1-5	-0.07 ± 1.00	-0.34 ± 0.97	42.76 (40)	0.35	0.84
Pink collar workers (n=421)						
Task performance						
4	TP7, 9, 11, 12, CP2	0.34 ± 1.49	-0.38 ± 0.98	49.05 (40)	0.15	0.82
Contextual performance						
5	CP9-10 CP14 AP1, 2, 4, 6-8	0.53 ± 0.91	-0.44 ± 1.44	65.34 (81)	0.90	0.88
Minor CWB						
6	CWB1-5	0.02 ± 1.13	-0.34 ± 1.00	58.65 (45)	0.08	0.85
Serious CWB						
7	CWB6-10	-0.48 ± 0.79	-0.22 ± 0.44	39.72 (20)	0.005	0.76
White collar workers (n=392)						
Task performance						
8	TP7, 9, 11, 12, CP1, 3, 6	-0.11 ± 0.72	-0.40 ± 1.11	69.21 (63)	0.28	0.80
Contextual performance						
9	CP9-10 CP12-14 AP1-2, 4-6	0.39 ± 1.32	-0.47 ± 1.60	104.35 (90)	0.14	0.81
CWB						
10	CWB1-5	0.21 ± 1.54	-0.32 ± 1.01	36.39 (40)	0.63	0.81

Table 3. Continued

Total sample (n=1,181)						
Task performance						
11	TP7, 9, 11, 12	0.20 ± 0.88	-0.44 ± 0.98	107.16 (32)	< 0.001	0.78
Contextual performance						
12	CP10, 14, AP1, 2, 6	0.40 ± 2.67	-0.54 ± 1.29	75.10 (45)	0.003	0.79
CWB						
13	CWB1-5	0.00 ± 1.90	-0.35 ± 1.01	76.28 (40)	< 0.001	0.84
Total sample (n=200)						
Task performance						
14	TP7, 9, 11, 12	0.20 ± 0.88	-0.44 ± 0.98	19.73 (32)	0.96	0.78
Contextual performance						
15	CP10, 14, AP1, 2, 6	0.40 ± 2.67	-0.54 ± 1.29	13.35 (45)	0.99	0.79
CWB						
16	CWB1-5	0.00 ± 1.90	-0.35 ± 1.01	14.87 (40)	0.99	0.84

Rasch analysis of the generic scales

Generic, short scales were constructed by including only those items that fitted the Rasch model for all occupational sectors (Table 2). These scales represent the IWPQ version 0.2. For the task performance scale, this included planning and organizing work (TP7), result-oriented working (TP9), prioritizing (TP11), and working efficiently (TP12). For the contextual performance scale, this included taking initiative (CP10), taking on challenging work tasks (CP14), keeping job knowledge and skills up-to-date (AP1 and AP2), and coming up with creative solutions to novel, difficult problems (AP6). For the counterproductive scale, this included displaying excessive negativity (CWB1, 2, 3), and doing things that harm your organization (CWB4 and 5).

Model fit

When testing the Rasch model for the total sample, the generic scales showed some misfit (analyses 11-13), as indicated by the significant Chi-square fit statistics. However, when setting the sample size at 200 [27], the Chi-square fit statistics became nonsignificant, indicating good model fit (analyses 14-16). Additionally, when testing the generic scales separately per occupational sector, the Chi-square fit statistics indicated good model fit (analyses not shown). This indicated that the

previously significant Chi-square fit statistic was caused by the power of the test, and that the data does in fact fit the Rasch model. The PSI ranged from 0.78 in the task performance scale to 0.84 in the CWB scale.

Category ordening

We examined whether items showed appropriate category ordening. Only the task performance item result-oriented working (TP9) demonstrated disordered thresholds. The answer categories 1 ("*sometimes*") and 2 ("*frequently*") were entirely overlapped by answer categories 0 ("*seldom*") and 3 ("*often*"), as shown in Figure 1. This indicated that there was no location on the scale (and therefore, no level of task performance) that "*sometimes*" or "*frequently*" were more likely to be selected than "*seldom*" or "*often*." Thus, for this item, a higher answer category did not necessarily reflect an increase in work performance. It was decided not to collapse any answer categories, because only one item showed disordered thresholds and the mean scores for categories showed the expected order [19, 23].

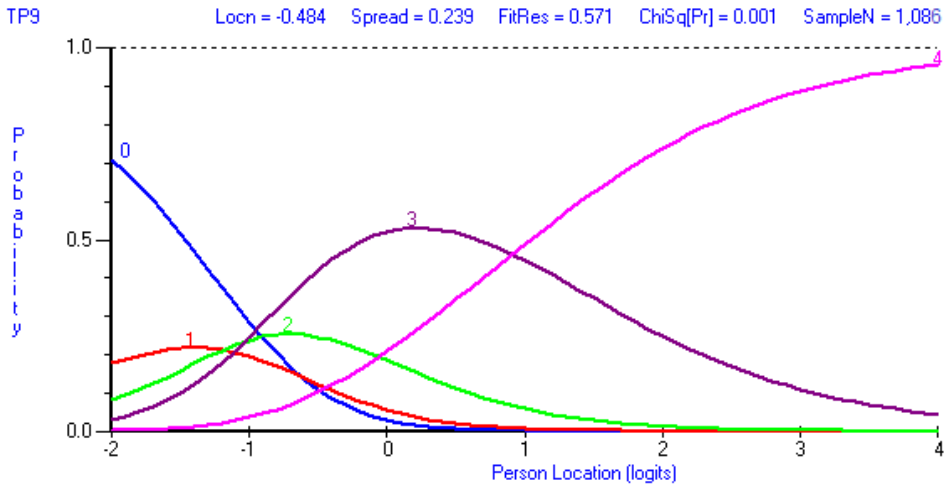


Figure 1. Category probability curve showing disordered thresholds for result-oriented working (TP9).

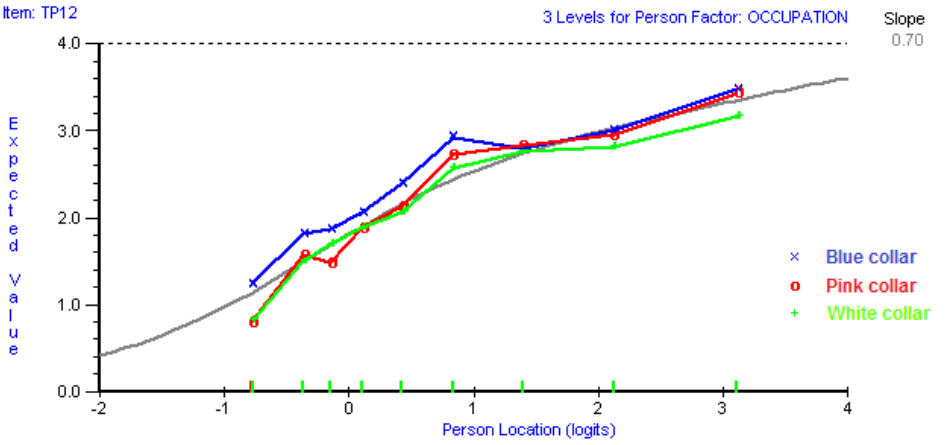


Figure 2. Item characteristic curves showing DIF between occupational sectors for working efficiently (TP12).

Differential Item Functioning (DIF)

We examined whether subgroups within the sample (occupational sector, gender, age) responded to items differently, despite equal levels of the underlying characteristic being measured. DIF was detected between occupational sectors for result-oriented working (TP9), and for working efficiently (TP12). Result-oriented working was harder for blue collar workers than for pink or white collar workers, whereas working efficiently was easier for blue collar workers than for pink and white collar workers. Also, DIF was detected between gender for working efficiently (TP12) and doing things that harm your organization (CWB5). Both were easier for males than for females.

A questionnaire consisting of many items with significant DIF may lead to biased scores for certain subgroups, and in future versions of the questionnaire, these items should be improved, or replaced by items free from DIF [29]. However, DIF tests are sensitive, and DIFs found in large samples may be statistically significant, but of little practical relevance [20]. DIF plots were used to examine whether the DIF effects were substantial. Figure 2 shows the item characteristic curves (ICCs) for item TP12, an example of the most serious DIF found in this study. For all identified DIF items, the ICCs were judged to be close together, and therefore, the DIF effects were considered to be of little practical relevance.

Local independence

We examined whether there were breaches in local independence of items, by looking for residual correlations ≥ 0.40. In the task performance scale, planning and organizing work (TP7) and prioritizing (TP11) showed negative response dependency (-0.42). Also, result-oriented working (TP9) and working efficiently (TP12) showed negative response dependency (-0.41). In the contextual performance scale, both keeping job knowledge up-to-date (AP1) and keeping job skills up-to-date (AP2) showed negative response dependency with taking initiative (CP10), taking on challenging work tasks (CP14), and coming up with creative solutions to novel, difficult problems (AP6) (ranging from -0.43 to -0.52). In the CWB scale, displaying excessive negativity (CWB1) and harming your organization (CWB5) showed negative response dependency (-0.41), as did displaying excessive negativity (CWB2) and harming your organization (CWB4; -0.43).

The findings of negative response dependency were likely a technical artifact of the Rasch model, caused by the low degrees of freedom in the generic scales. When the number of items in a scale is low, the Rasch model will generally

find negative response dependencies. This can be illustrated by the following example: in a two-item scale, a sumscore of 3 can come about in two different ways. Namely, a person scores 1 on the first item and 2 on the second item, or a person scores 2 on the first item and 1 on the second item. The difference between each item must be -1. Consequently, the residual correlations will always be negative [30]. In addition, the negative response dependency may partly be caused by the large sample size. If the number of persons is very large, all observed correlations will be statistically significantly different from 0, even when items fit the Rasch model perfectly [30]. These explanations were supported by the finding that the negative response dependencies disappeared in the job-specific scales, where the degrees of freedom were higher, and the sample size was smaller.

Person-item targeting

To get an indication of how well targeted the items were for the persons in the sample, the person-item threshold maps were examined. First, the person-item threshold maps showed that, especially for task performance and CWB, most persons were located at the higher end of the performance scale (see Figure 3). Second, the person-item maps showed that for all scales, the items were reasonably well distributed over the whole range of the scale. However, as most persons were located at the higher end of the performance scale, the discriminative ability of the IWPQ could be improved by including more items that measure work performance at the higher end of the performance scale.

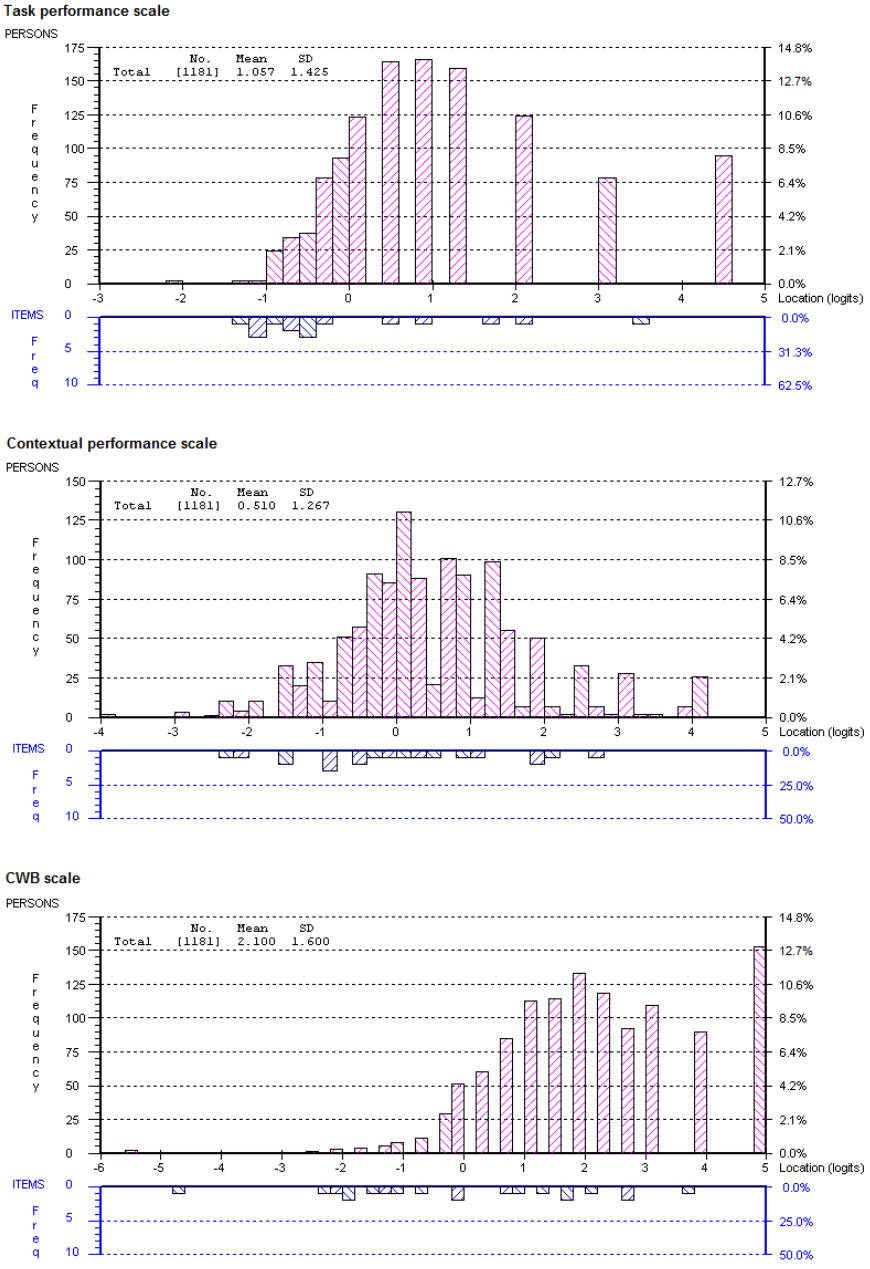


Figure 3. The person-item threshold maps showing the distribution of persons and items for the generic task performance, contextual performance, and CWB scales.

Discussion

Conceptual framework

The Individual Work Performance Questionnaire (IWPQ) 0.1 was based on a four-dimensional conceptual framework [7]. Instead, factor analyses showed that a three-dimensional IWP framework was generalizable across occupational sectors. In this framework, IWP consisted of the dimensions of task performance, contextual performance, and counterproductive work behavior (CWB). Although several studies have argued for adaptive performance as a separate dimension of IWP [e.g. 8, 9, 10], the current study did not support this proposition. Adaptive performance did not appear to be a separate dimension, but rather an aspect of contextual performance. Whereas contextual behaviors can be thought of as proactive, and adaptive behaviors as reactive [7], both can be considered supporting the organizational, social and psychological environment in which the technical core functions. They are both extra-role behaviors that do not directly contribute to the central job tasks, but do make it easier for employees to perform their central job tasks. In this view, it is not strange that the contextual and adaptive performance dimensions are one and the same. Although adaptive performance is relatively new to the field and it is too soon to draw firm conclusions, the findings of the current study indicate that adaptive performance is an aspect of contextual performance. The increasing attention for adaptive behaviors at work may reflect a shift in the content domain of contextual performance, to better suit the nature of today's work, which requires increasingly rapid adaptation to new situations and changing environments. In addition, six items hypothesized to belong to contextual performance (showing responsibility, communicating effectively and cooperating with others), appeared to belong to task performance. This finding likely also reflects the changing nature of today's work, in which the distinction between task and contextual performance behaviors becomes more blurred. Behaviors previously regarded as contextual behaviors, are now implicitly or explicitly seen as central to the job.

Generic and job-specific questionnaire items

The current research indicates that some items are relevant and generalizable across occupational sectors, whereas other items "work better" for specific occupational sectors. The IWPQ 0.2 could be constructed with generic scales that fitted the Rasch model well. The task performance scale included indicators measuring planning and organizing work, result-oriented working, prioritizing, and working efficiently. The

contextual performance scale included indicators measuring taking initiative, taking on challenging work tasks, keeping job knowledge and skills up-to-date, and coming up with creative solutions to novel, difficult problems. The CWB scale included indicators measuring displaying excessive negativity, and doing things that harm your organization.

The results of the current study indicated that the work quality and quantity indicators did not fit well with the other indicators of task performance. In literature and in existing questionnaires, however, these are traditionally two of the most often measured indicators of task performance [e.g., 1, 16]. Based on the conceptual definition of IWP (behaviors or actions that are relevant to the goals of the organization, and under control of the individual), the IWPQ focused on measuring indicators reflecting employee behaviors as opposed to the effectiveness of these behaviors. Work quality and quantity may actually reflect the effectiveness of employee behaviors rather than employee behaviors in themselves. Although the effectiveness of employee behaviors is certainly important from an organization's standpoint, strictly conceptual it should not be part of IWP questionnaires measuring employee behaviors. In addition, measures of effectiveness are likely to be more reflective of individual differences in abilities or skills (e.g., cognitive ability, social skill), and are frequently influenced by factors outside the control of the individual (e.g., technical problems, economic influences) [31].

Also, there was discrepancy between answers on *serious* CWB items (doing things that harm your co-workers or supervisor, and purposely making mistakes) and *minor* CWB items (displaying excessive negativity, doing things that harm the organization). This was most evident for pink collar workers, for whom the CWB dimension was split into two separate dimensions of *minor* and *serious* CWB. In all Rasch analyses, *serious* CWB items showed extreme ceiling effects, very low locations, and very low slopes. This could be due to the actual low occurrence of these behaviors, or due to worker's reluctance to honestly admit to serious CWB (social desirability). Thus, the current findings show that when aiming to assess IWP in a general working population, including *serious* CWB items may not be the best way to do this.

Generic scales

Generic scales could be constructed, supporting the use of an IWP questionnaire that can be utilized in all types of jobs. Generic scales pose considerable advantages in research, such as ease of administration and comparability between groups.

Although the generic scales showed good model fit, in some cases, job-specific scales may be preferred over generic scales. The job-specific scales showed a somewhat better fit, and a higher reliability, than the generic scales. Consequently, job-specific scales may be better able to spread out persons in the sample. Depending on their goal, researchers could choose to use a generic questionnaire (e.g., in nationwide surveys), or a job-specific questionnaire (e.g., in specific companies). Due to its generic nature, the IWPQ is not recommended for use in individual evaluations, assessments, and/or feedback.

Occupational sectors, and men and women, were found to respond differently to several items. A questionnaire consisting of many items with DIF may lead to biased scores for certain subgroups, because it is harder for them to achieve a good score on the questionnaire, despite equal levels of ability. Ideally, one should not compare the scores of subgroups when there are items with substantial DIF in the scale. However, DIF tests are sensitive [20], and the DIF effects identified in this study were considered to be of little practical relevance. Therefore, comparisons between occupational sectors, gender, and age groups on the IWPQ are justified.

Self-report questionnaire

The IWPQ was developed as a self-report questionnaire. Several downsides accompany self-reporting of performance, as opposed to objective measures or peer- or managerial ratings. First, self-ratings have a lower correlation with objective performance than managerial ratings. Jaramillo et al. [32] showed that managerial ratings correlated 0.44 with objective performance, whereas self-reports correlated 0.34 with objectives measures. Also, low correlations between self- and managerial ratings of performance are generally found, with meta-analyses reporting correlations between 0.19 [32] and 0.35 [33]. Second, self-ratings are known to show *leniency effects* [18]. That is, people are naturally motivated to present themselves in a favorable, socially desirable light. As a result, self-ratings of performance are generally one half to one standard deviation higher than ratings by peers or managers [18].

Nevertheless, self-report scales were chosen for several reasons. First, in many occupations, objective measures of performance are not easily obtainable [32]. Especially for knowledge work or high complexity jobs, direct measures of countable behaviors or outcomes such as production quantity or number of errors made, are almost impossible. Second, employees often have more *opportunity to observe* their own behaviors than peers or managers do [18]. This may be especially

true for counterproductive behaviors, because most of these behaviors are intended to be private and, hence, unobservable. It follows that peers or supervisors have little basis for judging many counterproductive behaviors [2]. A recent study by Berry [34] found that self-reports of counterproductive work behavior are actually more viable than other-ratings, with self-raters reporting engaging in more counterproductive behaviors than other-raters reported them engaging in. Third, peers or managers rate an employee's performance on basis of their general impression of the employee [2, 35]. This effect is named the *halo effect*. As a result, scores on the different dimensions of IWP are more similar and inter-correlations between the dimensions are overestimated. Finally, compared to objective measures or managerial ratings, self-reports have practical advantages such as ease of collection, issues of confidentiality, and less problems with missing data [36].

Strengths and limitations

The development of the IWPQ was based on thorough theoretical and practical examination. Care was taken to include generic indicators that covered the entire domain of IWP, that were equally relevant across occupational sectors, and that did not show overlapping content between dimensions. To guarantee this, thorough research about potential indicators was conducted before constructing the questionnaire [7, 16]. In addition, a reflective model was used to construct the questionnaire, in which the indicators were manifestations of the construct being measured. This implies that the indicators will correlate with each other, and also that they may replace each other, i.e. they are interchangeable. For that reason, it is not disastrous to miss some items that are also good indicators of the construct [20].

Another strength of the present study is that it is the first to develop or evaluate an IWP questionnaire using Rasch analysis. This offered unique insights into the IWPQ scale characteristics. Rasch analysis ensured that key measurement assumptions, such as appropriate category ordening, local independence, and differential item functioning, were tested. In addition, Rasch analysis has particular value in the development of new questionnaires, specifically in guiding item reduction [22]. It ensures that the items are well distributed over the whole range of the work performance scale. CTT techniques of item reduction rely on item-total correlations and/or indices of internal consistency, which can have unfortunate effects on the sensitivity of questionnaires and their ability to provide valid scores at the extremes of the measurement range. In CTT, items at the extremes of the measurement range are often discarded because too many or too few persons affirm

them. In reality, these “extreme” items may be the most important in a scale – extending the range of coverage of the construct [22].

The present study has some limitations as well. First, the IWPQ has not yet proven to be generalizable to managerial ratings. As mentioned before, low correlations between self- and managerial ratings of performance are generally found. Also, different factor structures have been found among self- and managerial ratings [37, 38]. Due to the *halo effect*, supervisors rate their employee’s performance in a more general way, leading to less discrimination between different dimensions of IWP for managerial ratings than for self-ratings. Thus, a simpler factor structure may be found for managerial ratings than for self-ratings. The convergence in scores between the different ratings sources, as well as generalizability of the factor structure of the IWPQ to managerial ratings, needs further examination.

Second, despite the shifted center of the rating scales, many persons scored high on the IWPQ items. This showed up in the item mean scores and in the Rasch analysis, where many persons had a high location on the person-item map. The high scores could be caused by the tendency of persons to evaluate themselves in a favorable light (*leniency effect*). Alternatively, the items may simply not be difficult enough for the persons in the sample. Especially for the task performance and CWB scale, there were too few items to measure the higher range of the scale. As a result, it is harder to discriminate among persons with high task performance and persons with low CWB, and to detect changes amongst these groups. In order to improve the discriminative ability of the IWPQ at the high ranges of the scale, adding extra answer categories is not an option. This will only test the response tendencies of the individual’s willingness to give extreme answers, and to what extent they can distinguish between the different answer categories. However, extra items could be formulated which cover the higher range of the ability scale [20]. This will show whether high scores were caused by the lack of difficult items, or whether a leniency effect is at play.

Conclusion

The aim of this study was to develop a generic and short questionnaire to measure work performance at the individual level. The Individual Work Performance Questionnaire (IWPQ) was developed, in which IWP consisted of the three dimensions of task performance, contextual performance, and CWB. The operationalization of the IWPQ scales was based on relevant and generic indicators, and the scales were refined based on a large, generic sample using the latest

statistical techniques. Short scales were constructed consisting of items that were relevant across all occupational sectors, supporting the use of a generic measure of IWP. Future research will need to focus on further developing and testing the reliability and validity of the IWPQ. The construct validity, sensitivity to change, and interpretability of the IWPQ need to be examined. One of the main adjustments to be made to the IWPQ is to formulate extra items, which cover the higher range of the ability scale. This will improve the questionnaire's discriminative ability, and sensitivity to change. Overall, the IWPQ facilitates researchers in measuring IWP more easily and comprehensively. In addition, unified measurement of IWP will increase comparability of studies. In the future, the IWPQ will hopefully contribute towards establishing the predictors and effects of IWP even more accurately and completely.

References

1. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. J Appl Psychol. 2002;87(1):66-80.
2. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol. 2005;90:1241-55.
3. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, eds. Handbook of industrial and organizational psychology. Palo Alto, CA, US: Consulting Psychologists Press. 1990:687-732.
4. Austin JT, Villanova P. The criterion problem: 1917-1992. Journal of Applied Psychology. 1992;77(6):836-74.
5. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, editors. Personnel Selection in Organizations. San Francisco, CA: Jossey Bass; 1993. p. 71-98.
6. Viswesvaran C, Ones DS. Perspectives on models of job performance. International Journal of Selection and Assessment. 2000;8(4):216-26.
7. Koopmans L, Bernaards CM, Hildebrandt VH, Schaufeli WB, De Vet, H,C,W, Van der Beek AJ. Conceptual frameworks of individual work performance - A systematic review. Journal of Occupational and Environmental Medicine. 2011 08;53(8):856-66.
8. Pulakos ED, Arad S, Donovan MA, Plamondon KE. Adaptability in the workplace: Development of a taxonomy of adaptive performance. Journal of Applied Psychology. 2000;85(4):612-24.
9. Sinclair RR, Tucker JS. Stress-CARE: An integrated model of individual differences in soldier performance under stress. In: Britt TW, Castro CA, Adler AB, editors. Military Life: The Psychology of Serving in Peace and Combat (Vol. 1): Military Performance. Westport, CT: Praeger Security International; 2006. p. 202-31.
10. Griffin MA, Neal A, Parker SK. A new model of work role performance: Positive behavior in uncertain and interdependent contexts. Academy of Management Journal. 2007;50:327-47.

11. Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. Journal of Management. 1991;17(3):601-17.
12. Podsakoff PM, MacKenzie SB. A second generation measure of organizational citizenship behavior. Indiana University, Bloomington.; 1989.
13. Van Scotter JR, Motowidlo SJ. Interpersonal facilitation and job dedication as separate facets of contextual performance. Journal of Applied Psychology. 1996;81(5):525-31.
14. Bennett RJ, Robinson SL. Development of a measure of workplace deviance. Journal of Applied Psychology. 2000;85(3):349-60.
15. Spector PE, Fox S, Penney LM, Bruursema K, Goh A, Kessler S. The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? Journal of vocational behavior. 2006;68:446-60.
16. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Measuring individual work performance - identifying and selecting indicators. Work: A Journal of Prevention, Assessment and Rehabilitation. 2013;45(3).
17. Schwarz N, Oyserman D. Asking questions about behavior: Cognition, communication, and questionnaire construction. American Journal of Evaluation. 2001;22:127-60.
18. Van der Heijden BIJM, Nijhof AHJ. The value of subjectivity: Problems and prospects for 36-degree appraisal systems. The International Journal of Human Resource Management. 2004;15(3):493-511.
19. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development, 4th ed. Oxford University Press; 2008.
20. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge University Press; 2011.
21. Rasch G. Probabalistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
22. Tennant A, McKenna SP, Hagell P. Application of rasch analysis in the development and application of quality of life instruments. Value in Health. 2004;7:S22-6.
23. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? when should it be applied, and what should one look for in a rasch paper? Arthritis & Rheumatism (Arthritis Care & Research). 2007 12/15;57(8):1358-62.

24. Andrich D, Lyne A, Sheridan B, Luo G. RUMM 2020. Pert, WA, Australia: RUMM Laboratory; 2003.
25. Traub RE. A priori considerations in choosing an item response model. In: Hambleton RK, editor. Applications of Item Response Theory. Educational Research Institute of British Columbia; 1983. p. 57-70.
26. Lundgren Nilsson A, Tennant A. Past and present issues in rasch analysis: The functional independence measure (FIM™) revisited. Journal of Rehabilitation Medicine. 2011;43(10):884-91.
27. Andrich D, Styles IM. Distractors with information in multiple choice items: A rationale based on the rasch model. In: Smith E, Stone G, editors. Criterion referenced testing: Using Rasch measurement Models. Maple Grove, Minnesota: JAM Press; 2009. p. 24-70.
28. Cronbach IJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-333.
29. Westers P, Kelderman H. Examining differential item functioning due to item difficulty and alternate attractiveness. Psychometrika. 1991;57(1):107-18.
30. RUMM Laboratory. "Factor Analysis and Negative PCA Values." Available from: http://www.rummlab.com.au/faq12.html (accessed 15 March 2012).
31. Penney LM, David E, Witt LA. A review of personality and performance: Identifying boundaries, contingencies, and future research directions. Human Resource Management Review. 2011;21:297-310.
32. Jaramillo F, Carrillat FA, Locander WB. A meta-analytic comparison of managerial ratings and self-evaluations. Journal of Personal Selling & Sales Management. 2005;XXV(4):315-28.
33. Harris MM, Schaubroeck J. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology. 1988;41:43-62.
34. Berry CM, Carpenter NC, Barratt CL. Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. Journal of Applied Psychology. 2012;97(3):613-36.
35. Viswesvaran C, Schmidt FL, Ones DS. Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. Journal of Applied Psychology. 2005;90(1):108-31.
36. Schoorman DF, Mayer RC. The value of common perspectives in self-reported appraisals: You get what you ask for. Organizational Research Methods. 2008;11(1):148-59.

37. Thornton GC. Psychometric properties of self-appraisals of job performance. Personnel Psychology. 1980;33:263-71.
38. Spector PE, Bauer JA, Fox S. Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? Journal of Applied Psychology. 2010;97(4):781-90.

Chapter 5

Improving the Individual Work Performance Questionnaire using Rasch Analysis

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt, Stef van Buuren, Allard J. van der Beek, Henrica C.W. de Vet

Journal of Applied Measurement. 2014; 15(2)

5

Abstract

Recently, the Individual Work Performance Questionnaire (IWPQ) version 0.2 was developed using Rasch analysis. The goal of the current study was to improve targeting of the IWPQ scales by including additional items. The IWPQ 0.2 (original) and 0.3 (including additional items) were examined using Rasch analysis. Additional items that showed misfit or did not improve targeting were removed from the IWPQ 0.3, resulting in a final IWPQ 1.0. Subsequently, the scales showed good model fit and reliability, and were examined for key measurement requirements (e.g., category ordening, unidimensionality, and differential item functioning). Finally, calculation and interpretability of scores were addressed. Compared to its previous version, the final IWPQ 1.0 showed improved targeting for two out of three scales. As a result, it can more reliably measure workers at all levels of ability, discriminate between workers at a wider range on each scale, and detect changes in individual work performance.

Introduction

Individual work performance (IWP) is a relevant and often used outcome measure of studies in the occupational setting. In the past decades, a great deal of research in fields such as management, occupational health, and industrial-organizational psychology has been devoted to discovering predictors and effects of IWP. However, only later attention has arisen for better conceptualizing and measuring IWP itself [e.g., 1, 2].

IWP can be defined as "*behaviors or actions that are relevant to the goals of the organization*" [3]. Thus, IWP focuses on behaviors or actions of employees, rather than the results of these actions. In addition, behaviors should be under the control of the individual, thus excluding behaviors that are constrained by the environment [2]. Since long, IWP is considered to be a multidimensional construct [3, 4]. Based on several reviews of the literature [2, 5, 6], it can be concluded that IWP consists of three broad dimensions. The first dimension, *task performance*, traditionally has received most attention, and can be defined as "the proficiency with which individuals perform the core substantive or technical tasks central to his or her job" [3]. The second dimension of IWP is *contextual performance*, defined as "behaviors that support the organizational, social and psychological environment in which the technical core must function" [7]. The third dimension of IWP is *counterproductive work behavior*, defined as "behavior that harms the well-being of the organization" [2].

Numerous scales have been developed to measure IWP. However, several limitations can be observed in these scales. First, and most strikingly, none of them measure all dimensions of IWP. As a result, there is no questionnaire available that incorporates the complete range of individual behaviors at work. Second, scales often use different operationalizations of the same dimensions, either due to different conceptualizations or different developmental or target populations. This makes it difficult to select the most appropriate and relevant scale. Third, scales measuring different dimensions often show items overlapping in content – called antithetical items [1].

To overcome the afore mentioned limitations, the Individual Work Performance Questionnaire (IWPQ) 0.2 was recently developed [8]. The IWPQ incorporates all three dimensions of IWP, whose operationalization was developed and refined based on a generic population (workers in all types of occupations), and includes no antithetical items. The IWPQ is a generic instrument, thus, it is suitable

for workers in all types of occupations (i.e. blue, pink, and white collar workers). Short scales for each dimension were constructed using Rasch analysis [9]. Rasch analysis offers detailed insight into scale characteristics, and therefore, has particular value in the development of new questionnaires [10]. The IWPQ scales showed good fit to the Rasch model, and satisfied key measurement requirements of the Rasch model, such as local independence, and unidimensionality.

One of the main purposes of the IWPQ is to detect changes in work performance, for example in interventions. In order to reliably measure change, the IWPQ should be able to measure persons at all levels of ability (from low to high IWP). Rasch analysis provides information on whether a questionnaire can measure persons at all levels of ability, in the form of person-item distribution maps. However, these showed that the targeting of the items to the persons was suboptimal [8]. An equal distribution of the items over the scales is desired for reliably measuring persons at all levels of ability, and for discriminating between persons at various ranges on the scale [11]. For the task and contextual performance scales, there were insufficient items located at the higher range of the scale (i.e. difficult items), while for the counterproductive work behavior scale, there were insufficient items sensitive to the lower range of the scale (i.e. easy items). As a consequence, the IWPQ is less able to discriminate workers with high task and contextual performance, and less able to discriminate workers low counterproductive performance.

The goal of the current study was to improve the targeting of the IWPQ. It was hypothesized that improved targeting could be achieved by formulating additional items that cover the locations of the scales where there was a scarceness of items.

Methods

Individual Work Performance Questionnaire (IWPQ)

Compared to the 14-item IWPQ version 0.2 [8], the IWPQ 0.3 was adjusted by adding items that should be located at the higher range of the task and contextual performance scales (i.e. difficult items), and items that should be located at the lower range of the counterproductive work behavior scale (i.e. easy items). Three items were formulated by the authors for task performance, seven for contextual performance, and three for counterproductive work behavior. This resulted in the 27-item IWPQ version 0.3 (see Table 2). The task performance (TP) scale consisted of 7 items (e.g.: “I managed to plan my work so that it was done on time”), contextual

performance (CP) of 12 items (e.g.: "I started new tasks myself, when my old ones were finished"), and counterproductive work behavior (CWB) of 8 items (e.g.: "I complained about unimportant matters at work"). Within each scale, items were presented to participants in randomized order, to avoid order effects. The TP and CP scales had a 5-point rating scale ranging from *seldom, sometimes, frequently, often,* to *always.* The CWB rating scale ranged from *never, seldom, sometimes, frequently,* to *often.* All items had a recall period of 3 months.

Participants

The IWPQ 0.3 was tested amongst a representative sample of Dutch workers, who were selected via a large internet panel organization. The internet panel consisted of Dutch adults who were willing to participate in research projects in exchange for a small reward. Workers from three occupational sectors were selected: blue collar (manual workers, e.g.: carpenter, mechanic, truck driver), pink collar (service workers, e.g.: hairdresser, nurse, teacher), and white collar workers (office workers, e.g.: manager, architect, scientist). Participants' gender, age, completed education level, and type of occupation were provided by the internet panel organization.

Data analysis

First, score ranges of the IWPQ items were checked for floor or ceiling effects (> 15% at the extreme values [11]). Inter-item correlations, Kaiser-Meyer-Olkin's (KMO) Measure of Sampling Adequacy (should be > 0.50), and Barlett's Test of Sphericity (should be < 0.05) were examined to test whether the items were sufficiently correlated to apply factor analysis. Principal components analysis with varimax rotation was performed in SPSS 20, to determine whether the three-dimensional conceptual framework of the IWPQ could be confirmed.

To examine the functioning of the items in further detail, each scale was examined using Rasch analysis [9]. The Rasch model assumes that the probability of a given respondent affirming an item is a logistic function of the difference between the person's ability and the item difficulty. In the Rasch model, items are hierarchically ordered based on difficulty, expecting that if a person with a certain ability scores well on a difficult item, then that person scores well on easier items as well. The polytomous Andrich rating scale model [12] was used, and analyses were conducted in RUMM2030 [13].

Model fit

If observed responses are equivalent or do not greatly differ from the expected responses from the model, then data are said to fit the Rasch model. The following fit statistics were used to test model fit: 1) Chi-square fit, 2) item fit residuals, and 3) person fit residuals. The Chi-square fit statistic is an item-trait interaction score, and reflects the property of invariance across the trait. Generally, Chi-square fit statistics should be nonsignificant, indicating model fit. However, this statistic is highly dependent on sample size, and in large samples it is almost certain to show statistical significance because of the high power of the test [14]. Therefore, model fit for the total sample was tested by randomly setting the sample size at 500 [15]. Item and person fit residuals represent the residuals between the observed and expected values for items and persons. Ideally, these should have a mean of approximately 0 and a standard deviation (*SD*) of 1.

Reliability

Furthermore, the person separation index (PSI) was examined. The PSI is an estimate of the internal consistency of a scale, and is similar to Cronbach's alpha [16], only it uses the logit scale estimates as opposed to the raw scores. It is interpreted in a similar manner, that is, a minimum value of 0.70 is required for group use and 0.85 for individual use [10]. PSI also indicates how well the items separate, or spread out, the persons in the sample [17].

Targeting of the scales

The person-item threshold map reveals the location of the persons and the items on a linear scale that runs from –5 to +5, with 0 being the average item difficulty. This gives an indication of how well targeted the items are for persons in the sample [10]. An equal distribution of items is desired if the instrument has to discriminate between persons at various ranges on the scale. Ideally, the mean location of the persons is 0 and the *SD* is 1, indicating perfect targeting of the items to the persons.

Improving fit

Multiple statistics were examined to determine which items should be removed to improve fit of a scale. First, it was examined which items showed fit residuals outside the accepted values of < –2.5 or > 2.5. Second, as the goal of the current study was to improve targeting of the IWPQ, it was examined whether the additional items contributed to improved targeting of the scales. This was done by examining the item

locations. For the task and contextual performance scales, items with a high difficulty parameter (as indicated by a location > 0) improved targeting, whereas for the CWB scale, items with a low difficulty parameter (as indicated by a location < 0) improved targeting. Both item fit residuals and targeting were taken into account in deciding which items to remove from the scale. Item removal was an iterative process, with one item removed at a time and fit re-estimated accordingly.

Category ordening

In addition to good model fit, data has to satisfy several requirements of the Rasch model. For one, Rasch analysis assumes that when using polytomous answer categories, a higher category should reflect an increase in underlying ability. If appropriate category ordening does not occur, thresholds between adjacent answer categories are disordered [10].

Local independence

Also, Rasch analysis assumes local independence, i.e. that the response to an item is independent of responses to other items, after controlling for the person's ability. There can be two types of breaches in local independence: response dependency and multidimensionality. In response dependency, the response to one item depends on the responds to a previous item. Response dependency can be identified through the residual correlation matrix, by looking for residual correlations ≥ 0.30. Multidimensionality can be identified through a principal components analysis of the residuals. Besides the main Rasch factor, there should be no further associations between the items other than random associations [10].

Differential Item Functioning

Finally, Rasch analysis assumes that a scale functions consistently, irrespective of subgroups within the sample being assessed. Differential item functioning (DIF) can affect model fit when different groups within the sample respond in a different manner to an item, despite equal levels of the underlying characteristic being measured [10]. In the current study, DIF for gender, age, and occupational sector was examined.

Results

Participants

In January 2012, 1,424 Dutch workers filled in the 27-item IWPQ. Participants were all employed, and aged 17 to 69 years. Less than half of the participants (42.4%) was female. The sample consisted of 442 blue collar, 540 pink collar, and 442 white collar workers. Table 1 presents further sample characteristics.

Table 1. Sample characteristics

	Total sample (*N* = 1,424)	**Blue collar (*n* = 442)**	**Pink collar (*n* = 540)**	**White collar (*n* = 442)**
Gender (% female)	42.4	14.0	65.6	42.5
Age (%)				
17-34 years	22.2	19.9	23.0	23.5
35-44 years	26.2	22.6	30.7	24.2
45-54 years	29.6	29.4	28.9	30.5
55-69 years	22.0	28.1	17.4	21.8
Education level (%)				
Primary	3.1	5.4	3.3	0.5
Secondary	38.1	54.5	40.7	18.6
Middle-level applied	29.7	34.4	34.4	19.2
Higher professional	28.5	4.8	21.3	61.1
Unknown	0.6	0.9	0.2	0.7

IWPQ

Conceptual framework

Table 2 shows the means (and *SD*s) of the IWPQ items. The score distributions of the IWPQ items were examined for floor or ceiling effects (> 15% of responses at the extreme categories). Four task performance items and two contextual performance items showed ceiling effects. All CWB items showed floor effects (Table 2). The inter-item correlations were appropriate for factor analysis, with the Kaiser-Meyer-Olkin's measure of sampling adequacy being > 0.90, and Bartlett's test of sphericity showing a *p*-value < 0.001. Based on the scree plot, the three-dimensional conceptual framework of the IWPQ was confirmed. All items loaded on the expected factors.

Table 2. Items of the Individual Work Performance Questionnaire (IWPQ)

Items		Mean	*SD*	% floor	% ceiling
Task performance scale					
In the past 3 months…					
TP1	I managed to plan my work so that it was done on time.	2.80	0.95	2.1	23.2
TP2 *	My planning was optimal.	2.47	0.98	3.4	13.2
TP3	I kept in mind the results that I had to achieve in my work.	3.11	0.81	0.8	34.3
TP4	I was able to separate main issues from side issues at work.	2.83	0.82	0.7	19.3
TP5 **	I knew how to set the right priorities.	2.87	0.77	0.6	19.0
TP6	I was able to perform my work well with minimal time and effort.	2.32	1.00	4.8	9.5
TP7 *	Collaboration with others was very productive.	2.48	0.89	2.6	9.2
Contextual performance scale					
In the past 3 months…					
CP1 *	I took on extra responsibilities.	2.24	1.09	6.0	11.5
CP2	I started new tasks myself, when my old ones were finished.	2.57	1.13	5.6	23.1
CP3	I took on challenging work tasks, when available.	2.32	1.08	6.4	12.6
CP4	I worked at keeping my job knowledge up-to-date.	2.28	1.15	7.9	14.6
CP5	I worked at keeping my job skills up-to-date.	2.42	1.02	4.6	13.0
CP6	I came up with creative solutions to new problems.	2.31	0.98	3.4	9.6
CP7 *	I kept looking for new challenges in my job.	2.12	1.10	7.6	10.8
CP8 **	I did more than was expected of me.	2.51	0.99	2.8	15.7
CP9 *	I actively participated in work meetings.	2.25	1.20	10.9	14.5
CP10 **	I actively looked for ways to improve my performance at work.	2.30	1.00	3.9	10.5

Table 2. Continued

CP11 **	I grasped opportunities when they presented themselves.	2.40	1.03	3.7	13.6
CP12 **	I knew how to solve difficult situations and setbacks quickly.	2.43	0.91	2.2	9.6
Counterproductive work behavior scale					
In the past 3 months…					
CWB1	I complained about unimportant matters at work.	0.97	0.85	33.0	0.4
CWB2	I made problems greater than they were at work.	0.71	0.76	44.9	0.3
CWB3	I focused on the negative aspects of a work situation, instead of on the positive aspects.	1.10	0.86	26.1	0.6
CWB4	I spoke with colleagues about the negative aspects of my work.	1.56	1.02	17.2	2.9
CWB5	I spoke with people from outside the organization about the negative aspects of my work.	1.21	1.05	31.5	2.2
CWB6 **	I did less than was expected of me.	0.71	0.73	42.1	0.4
CWB7 **	I managed to get off from a work task easily.	0.98	0.78	29.3	0.4
CWB8 **	I sometimes did nothing, while I should have been working.	0.80	0.82	42.1	0.5

Note. * additional items that were retained, ** additional items that were not retained.

Rasch analysis

To examine the functioning of the items in further detail, each scale was examined using Rasch analysis. In Table 3, the summary fit statistics for the IWPQ 0.2 (original items), 0.3 (including additional items), and 1.0 (final version) are presented per scale.

Model fit, reliability, targeting, and improving fit

Task performance

Model fit was tested with a sample size of 500, to avoid significance due to a large sample size [19]. The scale showed good model fit for both the IWPQ 0.2 (p = 0.65)

and IWPQ 0.3 (p = 0.38), see Table 3. Ideally, the person and item fit residual mean and *SD* are close to 0 and 1, indicating perfect fit of the data to the Rasch model. When comparing the IPWQ 0.2 and 0.3, the mean location of the persons decreased from 1.24 to 1.13, indicating slightly better targeting of the IWPQ 0.3. The item fit residual *SD* increased from 1.97 to 3.18, indicating greater misfit amongst the items in version 0.3. The PSI increased from 0.71 to 0.82, indicating higher reliability for the IWPQ 0.3.

First, it was examined which items showed fit residuals outside the accepted values of < –2.5 or > 2.5. Item 5 ("I knew how to set the right priorities") had a slightly large negative fit residual (–2.87), whereas item 7 ("Collaboration with others was very productive") had a large positive fit residual (6.17). Second, the location of the additional items was examined. Item 2 ("My planning was optimal") had a location of 0.48, and, thus, improved targeting of the scale. Items 5 and 7 had locations of –0.63 and 0.57, respectively. Based on these findings, item 5 was first removed from the scale, because it did not improve targeting. After this, item 7 was also removed from the scale, because it still showed a large positive fit residual (4.86), and it deteriorated model fit. Subsequently, the final 5-item task performance scale was established, showing good model fit (p = 0.92) and a PSI of 0.81.

Contextual performance

The scale showed good model fit for the IWPQ 0.2 (p = 0.96) and 0.3 (p = 0.43), see Table 3. When comparing the IWPQ 0.2 and 0.3, the mean location of the persons indicated equal targeting. The item fit residuals increased from 2.02 to 3.88, indicating greater misfit amongst the items in version 0.3. The person fit residuals increased from 1.68 to 2.09, indicating greater misfit amongst the persons in version 0.3. The PSI value increased from 0.77 to 0.90, indicating higher reliability for version 0.3.

Four items (1, 2, 4, and 9) showed large positive fit residuals, and three items (3, 7, and 11) showed large negative fit residuals. Second, the additional items 8 ("I did more than was expected of me"), 11 ("I grasped opportunities when they presented themselves") and 12 ("I knew how to solve difficult situations and setbacks quickly") did not improve targeting, as evidenced by their negative locations (–0.32, –0.16 and –0.17, respectively), and were therefore removed from the scale. After their deletion, additional item 10 ("I actively looked for ways to improve my performance at work") also showed a low location (–0.06), and was removed from the scale. After this, the item fit residuals still indicated some misfit

among the items. The items 1 and 9 still showed large positive fit residuals (2.53 and 4.77), and items 3 and 7 still showed large negative fit residuals (–5.28 and –3.31). However, because all four items had a positive location (0.07, 0.17, 0.02, and 0.21, respectively), and contributed to model fit, they were retained in the scale. This resulted in the final 8-item contextual performance scale, showing good model fit (p = 0.37) and a PSI of 0.85.

Counterproductive work behavior (CWB)

The scale showed good model fit for the IWPQ 0.2 (p = 0.92) and 0.3 (p = 0.89), see Table 3. When comparing the IWPQ versions 0.2 and 0.3, the mean location of the persons decreased from –1.69 to –1.80, indicating slightly worse targeting for version 0.3. The item fit residuals increased from 1.10 to 1.87, indicating greater misfit amongst the items in version 0.3. The PSI value increased from 0.74 to 0.79, indicating higher reliability for version 0.3.

CWB item 2 (“I made problems greater than they were at work”) showed a large negative fit residual (–2.92). Second, it was examined whether the three additionally formulated items had negative item locations, i.e. improved targeting. However, none of the additional items did (locations of 0.45, 0.29, and 0.27, respectively), and they were removed from the scale. The item and person fit residuals indicated no further misfit, and the previously misfitting item now also showed an acceptable fit residual. The original 5-item CWB scale remained, showing good model fit (p = 0.92) and a PSI of 0.74.

Table 3. Summary test of fit statistics presented per scale (n=500), for the IWPQ versions 0.2 (original items), 0.3 (including additional items), and 1.0 (final version)

	Items location, mean ± SD	Item fit residual, mean ± SD	Persons location, mean ± SD	Person fit residual, mean ± SD	Item-trait total Chi-square		PSI
					Chi-square (df)	*p*	
Task performance scale							
IWPQ 0.2	0.00 ± 0.67	–0.22 ± 1.97	1.24 ± 1.49	–0.45 ± 1.00	24.54 (28)	0.65	0.71
IWPQ 0.3	0.00 ± 0.62	0.03 ± 3.18	1.13 ± 1.41	–0.49 ± 1.29	65.72 (63)	0.38	0.82
IWPQ 1.0	0.00 ± 0.63	0.02 ± 1.86	1.12 ± 1.46	–0.46 ± 1.10	32.44 (45)	0.92	0.81
Contextual performance scale							
IWPQ 0.2	0.00 ± 0.15	0.39 ± 2.02	0.47 ± 1.28	–0.73 ± 1.68	29.90 (45)	0.96	0.77
IWPQ 0.3	0.00 ± 0.19	0.40 ± 3.88	0.48 ± 1.28	–0.68 ± 2.09	75.21 (72)	0.37	0.9
IWPQ 1.0	0.00 ± 0.17	0.39 ± 3.24	0.38 ± 1.21	–0.63 ± 1.77	110.05 (108)	0.43	0.85

Table 3. Continued

	Items location, mean ± SD	Item fit residual, mean ± SD	Persons location, mean ± SD	Person fit residual, mean ± SD	Item-trait total Chi-square		PSI
					Chi-square (df)	*p*	
Counterproductive work behavior scale							
IWPQ 0.2	0.00 ± 0.69	0.30 ± 1.10	–1.69 ± 1.44	–0.39 ± 1.09	32.55 (45)	0.92	0.74
IWPQ 0.3	0.00 ± 0.59	0.12 ± 1.87	–1.80 ± 1.29	–0.38 ± 1.27	58.32 (72)	0.89	0.79
IWPQ 1.0*	0.00 ± 0.69	0.30 ± 1.10	–1.69 ± 1.44	–0.39 ± 1.09	32.55 (45)	0.92	0.74

Note. * The counterproductive work behavior scale version 0.2 and 1.0 contained the same items.

Category ordening

After reaching the final IWPQ 1.0, key measurement requirements of the Rasch model were tested. First, appropriate category ordening was examined. Out of all 18 items, only task performance item 3 ("I kept in mind the results that I had to achieve in my work") demonstrated disordered thresholds, for pink collar workers. Answer category 1 (*sometimes*) was entirely overlapped by the other answer categories, as shown in Figure 1. This indicated that for this item, there was no location on the scale (and therefore, no level of task performance) that pink collar workers were more likely to select *sometimes* than the other answer categories. It was decided not to collapse any answer categories, because only one item showed disordered thresholds, this occurred for only one occupational sector, and the mean scores for categories did show the expected order [10, 18].

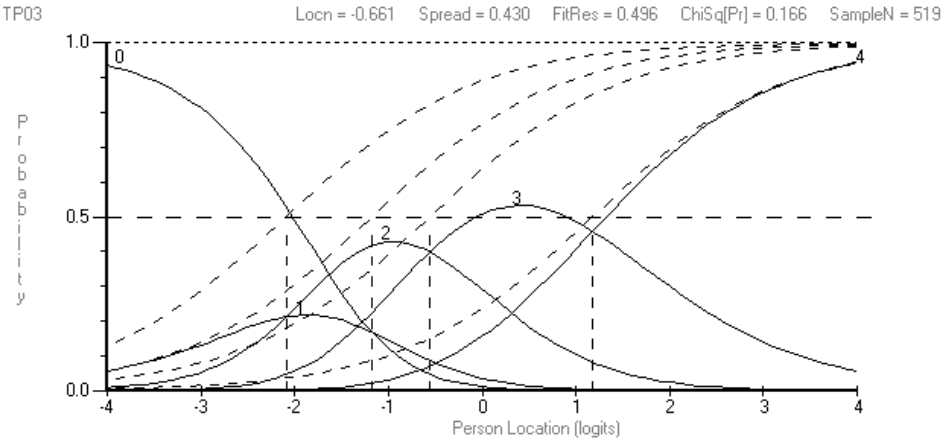


Figure 1. The category probability curves showing disordered thresholds for task performance item 3, for pink collar workers. The latent dichotomous responses (dotted lines) represent the observed responses for each answer category. The category characteristic curves (solid lines) represent the probability that the answer category will be selected, depending on the person location. The dotted, vertical lines indicate the thresholds between two answer categories.

Local independence

There was a slight negative response dependency between task performance items 3 and 6 (r = -0.33), and 1 and 7 (r = -0.32). Also, negative response dependency was identified for CWB item 1 with 4 and 5 (r = -0.34 and -0.37), and for CWB item 2 with

4 and 5 (r = -0.37 and -0.33). The identified negative residual correlations were not worrisome, and were not considered to violate the assumption of local independence.

To estimate the degree of multidimensionality, for each scale, two subsets of items (positively and negatively loaded items on PC1) were created. These two sets of items were used to make separate person estimates, and independent t-tests were performed to determine whether these two subsets of items lead to significantly different person estimates (95% CI). The two subsets of items did not produce significantly different person estimates for any of the scales (< 5%), indicating unidimensionality.

Differential Item Functioning

Finally, we examined whether subgroups (gender, age, and occupational sectors) within the sample responded to items differently, despite equal levels of ability. In the task performance scale, uniform DIF was detected between age groups for item 6 ("I was able to perform my work well with minimal time and effort"). Workers aged 17 to 35 years found this item easier than older workers. Thus, with equal levels of task performance, younger workers scored higher on this item, than older workers. Uniform DIF was detected between occupational sectors for task performance items 3 ("I kept in mind the results that I had to achieve in my work") and 6 ("I was able to perform my work well with minimal time and effort"). The first item was easier for white collar workers than for blue and pink collar workers, whereas the second item was easier for blue collar workers than for pink and white collar workers. The DIF for the occupational sectors cancelled each other out slightly, but overall, favored white collar workers. This meant that white collar workers scored higher on the scale than blue or pink collar workers, with equal levels of task performance.

In the contextual performance scale, uniform DIF was detected between occupational sectors for the items 1 ("I took on extra responsibilities") and 9 ("I actively participated in work meetings"). The first item was easier for blue collar workers than for pink and white collar workers, whereas the second item was easier for white collar workers than for blue and pink collar workers. However, these effects may cancel each other out, and when comparing the person location means per occupational sector, the difference was not significant (p = 0.70).

In the CWB scale, non-uniform DIF for gender was detected for item 2 ("I made problems greater than they were at work"). At the same level of CWB, females scored higher on this item than males. Uniform DIF for age was detected for item 4

("I spoke with colleagues about the negative aspects of my work"). At the same level of CWB, older workers scored higher on this item than younger workers.

Targeting

For the IWPQ 0.2 task and contextual performance scales, it was observed that most persons were located at the higher range of the ability scale, and there were insufficient items located at this range of the scale. For the CWB scale, most persons were located at the lower range of the ability scale, and there were insufficient items located at this range of the scale (Figure 2).

For the IWPQ 1.0 task and contextual performance scales, it was observed that the persons were located more towards the center of the ability scale (reflected in a lower mean person score, see Table 3), and the item thresholds were distributed more evenly across the scale (reflected in more thresholds at the higher range of the scales; Figure 3). The information curve also covers more of the person distribution. This indicated improved person-item targeting. However, for task performance, there was still some scarceness of the items at the highest end of the scales, indicating that it is hard to distinguish amongst top task performers. For the CWB scale, targeting remained the same. Although the item thresholds were distributed quite evenly across the scale, most persons were located at the lower range of the ability scale. Compared to the person locations, there were insufficient items at the lowest end of the scale, indicating that it is hard to distinguish amongst the lowest counterproductive performers.

Calculating scores

For the subscales, a mean score can be calculated by adding the item scores, and dividing their sum by the number of items in the subscale. Mean subscale scores were chosen because they are easier to understand as their values are in the same range (0-4) as the item scores. One overall IWPQ score cannot be calculated, as the valid calculation of a sumscore requires unidimensionality [19]. Furthermore, summing results in a loss of information about the underlying separate dimensions.

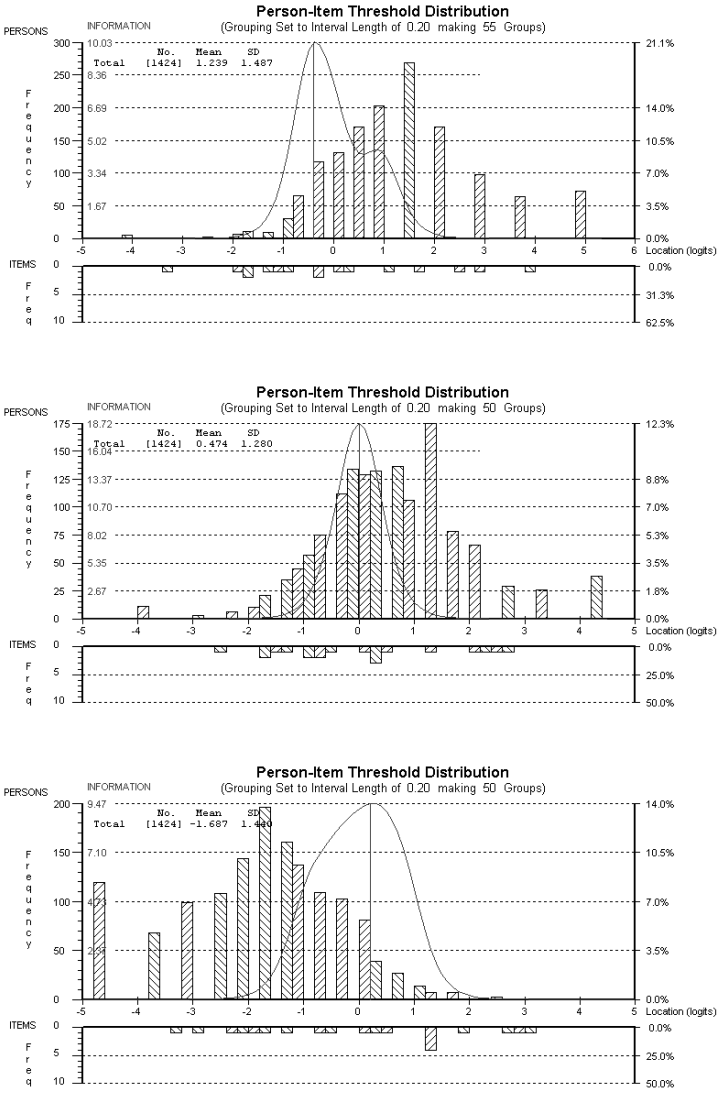


Figure 2. From top to bottom: person-item threshold maps representing the targeting of the IWPQ 0.2 task performance, contextual performance, and counterproductive work behavior scale, respectively. The top distribution in each map shows the persons, and the bottom distribution shows the item thresholds. The curve in the person distribution represents the information function.

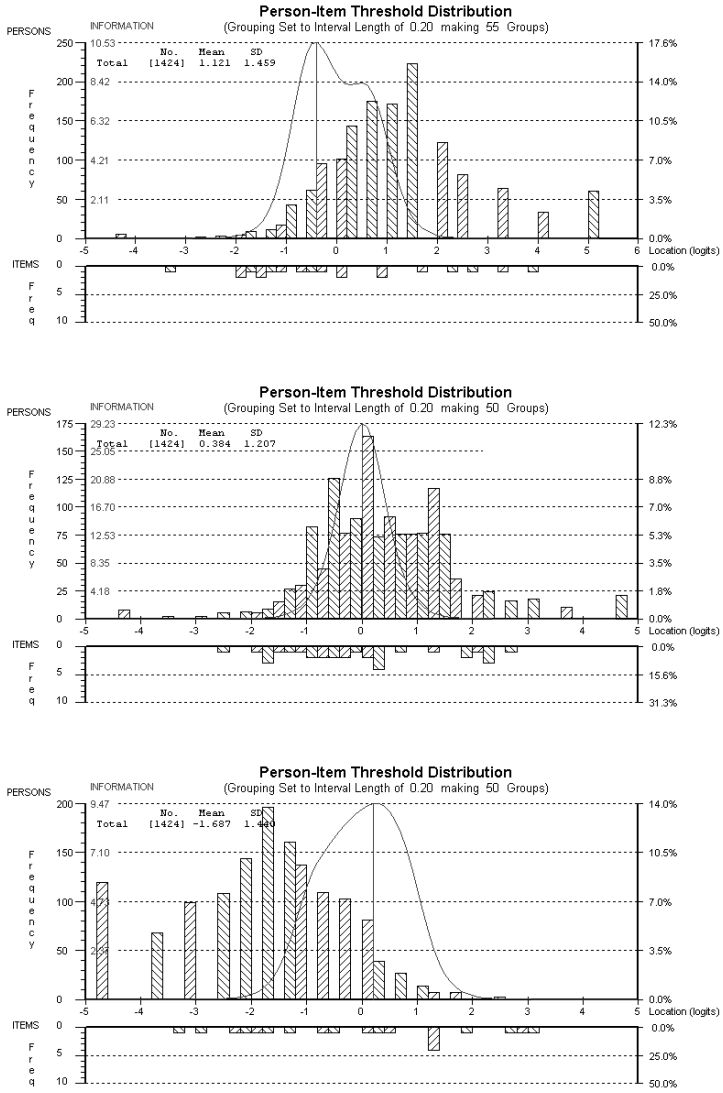


Figure 3. From top to bottom: person-item threshold maps representing the targeting of the IWPQ 1.0 task performance, contextual performance, and counterproductive work behavior scale, respectively. The top distribution in each map shows the persons, and the bottom distribution shows the item thresholds. The curve in the person distribution represents the information function. Please note that the counterproductive work behavior scale contains the same items in versions 0.2 and 1.0, and thus, targeting is the same.

Interpretation

Finally, we consider the interpretability of the IWPQ, defined as "the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations - to an instrument's quantitative scores or change in scores" [20]. As the current study used a large, representative sample of workers, the scores obtained in the present study are considered to be generalizable, and are thus considered norm scores. However, because DIF was identified for occupational sectors, norm scores are presented separately for each occupational sector. The distribution of scores presented in Table 4 can serve as a guide for interpretability. An interpretation of the scores, based on percentiles, is given from "very high" to "very low" performance. The interpretability of change scores remains a question for future research.

Table 4. Distributional properties and interpretation of the IWPQ scale scores (ordinal), per occupational sector

	Blue collar			**Pink collar**			**White collar**		
	TP	**CP**	**CWB**	**TP**	**CP**	**CWB**	**TP**	**CP**	**CWB**
Mean	2.77	2.30	1.03	2.68	2.31	1.09	2.55	2.34	1.21
SD	0.62	0.82	0.63	0.63	0.76	0.71	0.63	0.72	0.66
% 0 score	0.2	0.5	8.8	0.4	0.7	10.2	0.5	0.5	5.7
% 100 score	4.8	1.6	0.2	3.5	1.9	0.2	1.4	0.9	0.2
Interpretation									
"Very low"									
(≤ 10th	≤	≤	≤	≤	≤	≤	≤	≤	≤
percentile)	2.00	1.25	0.20	1.83	1.25	0.00	1.83	1.37	0.40
"Low"	2.01	1.26	0.21	1.84	1.26	0.01	1.84	1.38	0.41
(10th - 25th	-	-	-	-	-	-	-	-	-
percentile)	2.49	1.74	0.59	2.32	1.74	0.59	2.16	1.87	0.79
"Average"	2.50	1.75	0.60	2.33	1.75	0.60	2.17	1.88	0.80
(25th - 75th	-	-	-	-	-	-	-	-	-
percentile)	3.16	2.99	1.39	2.99	2.87	1.59	2.99	2.87	1.59
"High"	3.17	3.00	1.40	3.00	2.88	1.60	3.00	2.88	1.60
(75th - 90th	-	-	-	-	-	-	-	-	-
percentile)	3.49	3.24	1.79	3.49	3.12	1.99	3.32	3.24	1.99
"Very high"									
(≥ 90th	≥	≥	≥	≥	≥	≥	≥	≥	≥
percentile)	3.50	3.25	1.80	3.50	3.13	2.00	3.33	3.25	2.00

Discussion

Developing a measurement instrument is an iterative process, and there should be enough time for proper field-testing, further adaptation and re-evaluation before the final instrument is arrived at [11]. Often, however, there is insufficient time and funds to do this, and the instrument is used in research or practice straight away, making the threshold for adaptations, understandably, high. Strength of the IWPQ is that time was taken to improve the quality and functioning of the IWPQ, before it being applied in research or practice. In previous research, suboptimal targeting of the IWPQ version 0.2 was identified [8]. Therefore, the goal of the current study was to improve the targeting of the IWPQ, in order to more reliably measure persons at all levels of ability, enabling the instrument to more reliably detect changes in their IWP over time. The current study presents the IWPQ version 1.0, with generic, short scales that showed good fit to the Rasch model. Improved targeting of the task and contextual performance scales was achieved, by adding new items to the scales.

To our knowledge, the current study is one of the first studies attempting to improve the targeting of a measurement instrument. In the fields of social science and health science, attention for Rasch analysis has picked up in recent years. Various questionnaires, which were originally developed using classical test theory, have been re-evaluated with Rasch analysis [e.g., 17, 19, 21]. The main goals of these studies were to examine whether the questionnaires met key measurement requirements of the Rasch model, and whether they could be shortened by removing misfitting items. Often, these questionnaires do not meet key measurement requirements of the Rasch model, such as appropriate category ordening, unidimensionality, and differential item functioning. Several studies found that the questionnaire under examination showed suboptimal targeting, with most questionnaires exhibiting considerable ceiling effects [e.g., 21-23]. While some authors suggest that this suboptimal targeting could be improved by adding new items, to our knowledge, so far, none have actually attempted this.

Floor effects

In the current study, improved targeting of the CWB scale was not achieved, and floor effects remained for this scale. However, we cannot be sure whether this floor effect is a true characteristic of the population (an actual low occurrence of these behaviors in the workplace), or whether this is a shortcoming of the measurement instrument (unable to pick up low CWB). Furthermore, there are obvious problems

with social desirability: workers might be reluctant to admit that they engage in CWBs. Especially in longitudinal studies, floor effects could be problematic, because workers who score low on CWB at baseline, cannot show any further improvement (thus, even less CWB). However, it is important to consider whether we actually want to discriminate low counterproductive workers any further. After all, the main goal of the scale may be to discriminate workers that show moderate or high CWB, and to detect their improvements (decreases in CWB).

Misfitting items

Despite good model fit, not all items showed fit residuals within the acceptable limits. In the contextual performance scale, two items ("I took on extra responsibilities" and "I actively participated in work meetings"*)* had large positive fit residuals, indicating low levels of discrimination. Differential item functioning (DIF) between occupational sectors was identified for these items, which may have caused their large fit residuals. Two other items ("I kept looking for new challenges in my job" and "I took on challenging work tasks, when available") showed large negative fit residuals, indicating high levels of discrimination. The reason for their misfit is unclear. It is possible, however, that the large negative fit residuals are an artifact of the Rasch model, as a compensation for the two large positive fit residuals. Despite the large fit residuals of the items, they contributed to model fit and targeting of the scales, and were therefore retained.

Differential Item Functioning

Furthermore, differential item functioning (DIF) was identified for several items. A questionnaire consisting of many items with DIF may lead to biased scores for certain subgroups, because it is easier for them to achieve a good score on the questionnaire, despite equal levels of ability. For example, it is slightly easier for white collar workers to obtain a good score on the task performance scale, despite the fact that their level of task performance may be equally high as blue and pink collar workers. Ideally, one should not compare the scores of subgroups when there are items with substantial DIF in the scale. However, we must keep in mind that DIF analyses are very sensitive to sample size, and that even small amounts of DIF may be found to be statistically significant in large samples [11]. The maximum amount DIF identified in the IWPQ was 0.55 on the –5 to +5 Rasch ability scale, and it can be questioned whether this difference is practically relevant.

If we want a generic questionnaire that is comparable across genders, age groups, and occupational sectors, the items displaying DIF should be removed from the IWPQ. However, as one of the main purposes of the IWPQ is to detect changes over time, we chose to retain the items with DIF in order to obtain optimal targeting. Whether good targeting or comparability across subgroups is more important, depends on the purpose of the measurement instrument. If the goal of a measurement instrument is to detect changes over time, adequate targeting is most important. If the goal is to compare subgroups within a sample, items free from DIF are most important. In its current form, the IWPQ is suitable for all occupational sectors, is able to reliably measure persons at all levels of ability and to detect changes within persons or groups over time (e.g., in workplace intervention studies). However, because of differential item functioning, the IWPQ might be less apt for making comparisons between different groups (e.g., comparing carpenters and dentists on IWP). Thus, the IWPQ is generic in the sense that the same questionnaire can be distributed to workers from all occupational sectors. However, different cut-off points should be used when interpreting scores for workers from different occupational sectors. In addition, when using Rasch analysis, scores for the different occupational sectors are calculated differently. Thus, workers from different occupational sectors can have the exact same answers on the items in a scale, but still obtain different scale scores due to DIF.

Group versus individual use

The reliability of the IWPQ scales varied from 0.74 for the CWB scale to 0.85 for the contextual performance scale. As a minimum value of 0.70 is required for group use and 0.85 for individual use [10], all scales are appropriate for group comparisons. Our sample consisted of a large, representative population of workers from diverse occupational sectors in The Netherlands. This makes it likely that our findings are generalizable to a larger working population, and allows the scores obtained in the current study to be used as norm scores for the occupational sectors. The IWPQ is not recommended for use in individual evaluations, assessments, and/or feedback.

Future research

Future research will need to focus on further testing the reliability and validity of the IWPQ. Specifically, the construct validity of the IWPQ needs to be examined, as well as its sensitivity to change as a result of interventions. Also, the interpretability of change scores warrants attention. What is the smallest change the IWPQ can detect

(beyond measurement error), and when is a change practically relevant? So far, the IWPQ has only been tested in the Dutch language and population. To support widespread use of the IWPQ, a main concern should be to validate the IWPQ in other languages (especially in English), as well as in other countries and cultures.

Conclusion

The current study presents the IWPQ version 1.0, with generic, short scales that showed good fit to the Rasch model and satisfied key measurement requirements. Compared to its previous version, the IWPQ 1.0 showed improved targeting for two out of three scales. As a result, it can more reliably measure workers at all levels of ability, discriminate between workers at a much wider range on each scale, and detect changes in IWP.

References

1. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol. 2005;90:1241-55.
2. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. J Appl Psychol. 2002;87(1):66-80.
3. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: M.D. Dunnette, and L.M. Hough (Eds), Handbook of industrial and organizational psychology, Vol.1 (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.; 1990, pp. 687-732.
4. Austin JT, Villanova P. The criterion problem: 1917-1992. Journal of Applied Psychology. 1992;77(6):836-74.
5. Viswesvaran C, Ones DS. Perspectives on models of job performance. International Journal of Selection and Assessment. 2000;8(4):216-26.
6. Koopmans L, Bernaards CB, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual frameworks of individual work performance: A systematic review. Journal of Occupational and Environmental Medicine. 2011;53(8):856-66.
7. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, editors. Personnel Selection in Organizations. San Francisco, CA: Jossey Bass; 1993. p. 71-98.
8. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Development of an individual work performance questionnaire. International Journal of Productivity and Performance Management. 2013;62(1):6-28.
9. Rasch G. Probabalistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
10. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? when should it be applied, and what should one look for in a rasch paper? Arthritis & Rheumatism (Arthritis Care & Research). 2007 12/15;57(8):1358-62.
11. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge University Press; 2011.

12. Andrich D. Rating formulation for ordered response categories. Psychometrika. 1978;43:561-73.
13. Andrich D, Sheridan B, Luo G. RUMM 2030: Rasch unidimensional models for measurement. Perth, Western Australia: RUMM Laboratory. 2009.
14. Lundgren Nilsson A, Tennant A. Past and present issues in rasch analysis: The functional independence measure (FIM™) revisited. Journal of Rehabilitation Medicine. 2011;43(10):884-91.
15. Andrich D, Styles IM. Distractors with information in multiple choice items: A rationale based on the rasch model. In: Smith E, Stone G, editors. Criterion referenced testing: Using Rasch measurement Models. Maple Grove, Minnesota: JAM Press; 2009. p. 24-70.
16. Cronbach IJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-333.
17. Lamoureux EL, Pallant JF, Pesudovs K, Hassell JB, Keeffe JE. The impact of vision impairment questionnaire: An evaluation of its measurement properties using rasch analysis. Invest Ophthalmol Vis Sci. 2006 11;47(11):4732-41.
18. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development, 4th ed. Oxford University Press; 2008.
19. Van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insights into the measurement properties of the neck disability index. Arthritis & Rheumatism (Arthritis Care & Research). 2009 04/15;61(4):544-51.
20. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63:737-45.
21. Garamendi E, Pesuvods K, Stevens MJ, Elliott DB. The refractive status and vision profile: Evaluation of psychometric properties and comparison of rasch and summated likert-scaling. Vision Research. 2006;46:1375-83.
22. Gothwal VK, Wright TA, Lamoureux EL, Pesuvods K. Rasch analysis of the quality of life and vision function questionnaire. Optometry and Vision Science. 2009;86(7).
23. Pesuvods K, Garamendi E, Keeves JP, Elliott DB. The activities of daily vision scale for cataract surgery outcomes: Re-evaluating validity with rasch analysis. Investigative ophthalmology & visual science. 2003;44(7):2892-9.

Part III

Validation Phase

Chapter 6

Construct Validity of the Individual Work Performance Questionnaire

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt,
Henrica C.W. de Vet, Allard J. van der Beek

Journal of Occupational and Environmental Medicine. 2014; 56(3)

DOI 10.1097/JOM.0000000000000113

6

Abstract

Objective: It is difficult to measure individual work performance comprehensively and generically. This study examines the construct validity of the recently developed Individual Work Performance Questionnaire (IWPQ).

Methods: 1,424 Dutch workers from three occupational sectors (blue, pink, and white collar) participated in the study. First, IWPQ scores were correlated with related constructs (convergent validity). Second, differences between known groups were tested (discriminative validity).

Results: First, IWPQ scores correlated weakly to moderately with absolute and relative presenteeism, and work engagement. Second, significant differences in IWPQ scores were observed for workers differing in job satisfaction, and workers differing in health.

Conclusion: Overall, the results indicate acceptable construct validity of the IWPQ. Researchers are provided with a reliable and valid instrument to measure IWP comprehensively and generically, amongst workers from different occupational sectors, with and without health problems.

Introduction

Individual work performance (IWP), defined as "*behaviors or actions that are relevant to the goals of the organization*" [1], is an important outcome in multiple research fields, as well as in practice. The conceptualization of IWP has a long history, and many frameworks have been proposed to describe the construct domain of IWP [e.g., 1-3]. The measurement of IWP has proven to be even more challenging, with numerous and diverse behaviors, actions, or results being applied as indicators of IWP [4]. Thus, despite the importance of IWP in research and practice, there is little consensus on how to conceptualize and measure IWP. This lack of consensus is undesirable, because valid measurement is a prerequisite for accurately establishing, for example, predictors of IWP, or effectiveness of interventions to improve IWP.

In the field of occupational health, there has been little attention for conceptualizing the IWP construct. The main focus was on sickness absenteeism or presenteeism, i.e., work absence or losses in IWP due to health impairments. In accordance, numerous instruments have been developed to measure sickness absenteeism or presenteeism, such as the Work Productivity And Impairment Questionnaire [5], Work Limitations Questionnaire [6], and the WHO Health and Performance Questionnaire [7].

In the field of work and organizational psychology, traditionally, the main focus of the IWP construct was on task performance, which can be defined as the proficiency with which individuals perform the core substantive or technical tasks central to his or her job [1]. It is now generally agreed upon that, in addition to task performance, the IWP domain consists of contextual performance and counterproductive work behavior [2,3,8]. Contextual performance can be defined as behaviors that support the organizational, social and psychological environment in which the technical core must function [9]. Counterproductive work behavior can be defined as behavior that harms the well-being of the organization [3]. In accordance, work and organizational psychologists have developed numerous scales to measure task performance [e.g., 10], contextual performance [e.g., 11], or counterproductive work behavior [e.g., 12].

It is evident that a multitude of instruments exists to measure IWP, or related constructs such as absenteeism or presenteeism. However, these existing scales show several limitations. Most strikingly, none of them measure all of the relevant dimensions of IWP together. Thus, they do not measure the full range of IWP. Also, scales measuring different dimensions can include antithetical items,

creating unjust overlap between these scales [13]. As a result, the content validity of these scales can be questioned. Furthermore, none of the scales appear suitable for generic use. The scales were developed for specific populations, such as employees with health problems [e.g., 5-7,13], or they were developed and refined based on employees with a specific occupation [e.g., 10,11].

Recently, the Individual Work Performance Questionnaire 1.0 [IWPQ; 14,15] was developed, in order to overcome limitations of existing questionnaires. A conceptual framework for the IWPQ was established based on a systematic review of the occupational health, work and organizational psychology, and management and economics literature [8]. The conceptual framework consists of three dimensions (task performance, contextual performance, and counterproductive work behavior), that represent the full range of IWP. No antithetical items were included in the questionnaire. Furthermore, the IWPQ was developed and refined based on a generic working population (i.e., blue, pink, and white collar workers), and is therefore suitable across occupational sectors, as well as for workers with and without health problems. The development, as well as the face, and structural validity of the IWPQ have been established in previous studies [4,14,15].

The current study expands research on the IWPQ by examining its construct validity. Construct validity refers to whether the instrument provides the expected scores, based on existing knowledge about the construct [16]. In the current study, construct validity was assessed by testing expectations about the relationship of the IWPQ scales with related constructs (convergent validity), and about differences between groups (discriminative validity).

Convergent validity

First, the IWPQ was correlated with the World Health Organization's Health and Performance Questionnaire [HPQ; 7], a validated questionnaire that intends to measure a similar construct. Only the HPQ presenteeism questions were administered. If scores on these questionnaires correlate strongly, then this indicates convergent validity [16].

Hypothesis 1a: The HPQ absolute presenteeism score, representing one's overall performance, was expected to show a strong positive correlation with the IWPQ task and contextual scales, and a strong negative correlation with the IWPQ counterproductive scale ($r > 0.50$ and $r < -0.50$, respectively [17]).

Hypothesis 1b: The HPQ relative presenteeism score, representing one's overall performance compared to the performance of most workers at the same job, was expected to show a weak positive correlation with the IWPQ task and contextual scales ($r < 0.30$), and a weak negative correlation with the IWPQ counterproductive scale ($r > -0.30$).

Second, the IWPQ was correlated with the Utrecht Work Engagement Scale [UWES; 18]. Work engagement is defined as "*a positive, fulfilling, work-related state of mind that is characterized by vigor, dedication, and absorption*" [19]. We consider work engagement and individual work performance to be related but not similar constructs. Several studies have shown that work engagement is moderately positive related to IWP [e.g., 20,21]. If scores on these questionnaires correlate moderately, then this indicates convergent validity.

Hypothesis 1c: The overall work engagement score, and its subscale scores, were expected to show a moderate positive correlation with the IWPQ task and contextual scales, and moderate negative correlation with the IWPQ counterproductive scale ($r = 0.30$-0.50 and $r = -0.30$- -0.50, respectively).

Discriminative validity

Another way to test the validity of the IWPQ is to examine whether it can differentiate known groups. Two known predictors of IWP are job satisfaction [e.g., 22] and health [e.g., 23]. Job satisfaction can be defined as "*an emotional state resulting from the evaluation or appraisal of one's job experiences*" [24]. A recent review [25] and meta-analysis [22] have shown that higher job satisfaction predicts higher IWP. Health can be defined as "*a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity*" [26]. Both mental health [e.g., 27] and physical health [e.g., 23] have shown to be predictors of IWP.

Hypothesis 2a: Persons high in job satisfaction were expected to show significantly higher IWPQ task and contextual scores, and lower IWPQ counterproductive scores, than persons low in job satisfaction.

Hypothesis 2b: Persons in good health were expected to show significantly higher IWPQ task and contextual scores, and lower IWPQ counterproductive scores, than persons in poor health.

Methods

Participants

Participants were selected from a representative sample of Dutch workers from three occupational sectors: blue collar (manual workers, e.g.,: carpenter, mechanic, truck driver), pink collar (service workers, e.g.,: hairdresser, nurse, teacher), and white collar workers (office workers, e.g.,: manager, architect, scientist). Participants were recruited from a large internet panel consisting of Dutch adults willing to participate in research projects in exchange for a small reward.

Measures

Individual work performance was measured using the Individual Work Performance Questionnaire 1.0 [15]. The IWPQ 1.0 consisted of 3 scales (task performance, contextual performance, and counterproductive work behavior) with a total of 18 items (see Table 1). Within each scale, items were presented to participants in randomized order, to avoid order effects. All items had a recall period of 3 months and a 5-point rating scale ("*seldom"* to *"always"* for task and contextual performance, *"never"* to *"often"* for counterproductive work behavior). For the IWPQ subscales, a mean score was calculated by adding the item scores, and dividing their sum by the number of items in the subscale. Hence, the IWPQ yielded three subscale scores that ranged between 0 and 4, with higher scores reflecting higher task and contextual performance, and higher counterproductive work behavior.

To examine convergent validity, two presenteeism questions of the HPQ [7] were used: "*On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours*" (item 1) and "*Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked"* (item 2). Both had a recall period of 3 months and an 11-point rating scale ("*worst performance"* to *"top performance").* The HPQ absolute presenteeism score was calculated by multiplying item 2 (*rating of overall job performance)* by 10. It has a lower bound of 0 (total lack of performance during time on the job) and an upper bound of 100 (no lack of performance during time on the job). The HPQ relative presenteeism score was calculated by dividing item 2 by item 1. It is restricted to the range of 0.25 to 2.00, where 0.25 is the worst relative performance (25% or less of other workers' performance) and 2.00 is the best performance (200% or more of other workers' performance [28]).

Work engagement was measured using the UWES-9 [18]. The UWES includes three scales (vigor, dedication, and absorption) and has been extensively validated [e.g., 19,20]. No recall period was specified and all items had a 7-point rating scale ("*almost never"* to *"always").* The UWES-9 yielded one total score and three subscale scores ranging between 0 and 6.

Job satisfaction and overall health were measured using one question each *("How do you rate your overall job satisfaction?"* and *"How do you rate your overall health?"*). Both had a recall period of 3 months and an 11-point rating scale (*"very low"* to *"very high").* Research has shown that a single-item measure of job satisfaction and health-related quality of life correlate highly with multi-item measures of job satisfaction and health-related quality of life, and can therefore be considered valid [e.g., 29-31].

Table 1. Items of the Individual Work Performance Questionnaire (IWPQ), with means and standard deviations (SD) on a 0-4 range

Items		**Mean**	**SD**
Task performance (TP) scale			
In the past 3 months…			
TP1	I managed to plan my work so that it was done on time.	2.80	0.95
TP2	My planning was optimal.	2.47	0.98
TP3	I kept in mind the results that I had to achieve in my work.	3.11	0.81
TP4	I was able to separate main issues from side issues at work.	2.83	0.82
TP5	I was able to perform my work well with minimal time and effort.	2.32	1.00
Contextual performance (CP) scale			
In the past 3 months…			
CP1	I took on extra responsibilities.	2.24	1.09
CP2	I started new tasks myself, when my old ones were finished.	2.57	1.13
CP3	I took on challenging work tasks, when available.	2.32	1.08
CP4	I worked at keeping my job knowledge up-to-date.	2.28	1.15
CP5	I worked at keeping my job skills up-to-date.	2.42	1.02
CP6	I came up with creative solutions to new problems.	2.31	0.98
CP7	I kept looking for new challenges in my job.	2.12	1.10
CP8	I actively participated in work meetings.	2.25	1.20

Table 1. Continued

Counterproductive work behavior (CWB) scale			
In the past 3 months…			
CWB1	I complained about unimportant matters at work.	0.97	0.85
CWB2	I made problems greater than they were at work.	0.71	0.76
CWB3	I focused on the negative aspects of a work situation, instead of on the positive aspects.	1.10	0.86
CWB4	I spoke with colleagues about the negative aspects of my work.	1.56	1.02
CWB5	I spoke with people from outside the organization about the negative aspects of my work.	1.21	1.05

Data analysis

Pearson's correlations of the IWPQ subscale scores with the HPQ absolute and relative presenteeism scores, and the UWES-9 scores, were calculated. In addition, the correlations were examined separately for each occupational sector (blue, pink, and white collar workers), to determine whether the correlation strengths differed across occupational sectors.

For discriminative validity, groups low and high in job satisfaction, and low and high in overall health, were created using quartiles. Analyses of variance (ANOVAs) were performed to examine whether the job satisfaction quartiles, and the overall health quartiles, significantly differed on the IWPQ scores. Post-hoc tests with Bonferroni correction were performed to determine which quartiles significantly differed from each other. Finally, the ANOVAs with Bonferroni correction were performed separately for each occupational sector (blue, pink, and white collar workers), to determine whether the findings were generalizable across occupational sectors. All analyses were performed in SPSS 20 [32].

Results

Participants

In January 2012, 1,424 Dutch workers filled in the IWPQ. Participants were all employed, and aged 17 to 69 years. Table 2 presents sample characteristics and mean (and SD) scores on the IWPQ scales.

Table 2. Sample characteristics, and mean and standard deviation (SD) scores on the IWPQ task performance, contextual performance, and counterproductive work behavior scales (range 0-4)

	Total sample (N = 1,424)	**Blue collar (n = 442)**	**Pink collar (n = 540)**	**White collar (n = 442)**
Gender (% female)	42.4	14.0	65.6	42.5
Age (%)				
17-34 years	22.2	19.9	23.0	23.5
35-44 years	26.2	22.6	30.7	24.2
45-54 years	29.6	29.4	28.9	30.5
55-69 years	22.0	28.1	17.4	21.8
Education level (%)				
Primary	3.1	5.4	3.3	0.5
Secondary	38.1	54.5	40.7	18.6
Middle-level applied	29.7	34.4	34.4	19.2
Higher professional	28.5	4.8	21.3	61.1
Unknown	0.6	0.9	0.2	0.7
	Total sample Mean (SD)	**Blue collar Mean (SD)**	**Pink collar Mean (SD)**	**White collar Mean (SD)**
Task performance	2.71 (0.66)	2.82 (0.65)	2.71 (0.66)	2.59 (0.63)
Contextual performance	2.31 (0.77)	2.30 (0.82)	2.31 (0.76)	2.34 (0.72)
Counterproductive work behavior	1.11 (0.67)	1.03 (0.63)	1.09 (0.71)	1.21 (0.66)

Convergent validity

The absolute presenteeism score showed a moderate positive correlation with the IWPQ task and contextual scales, and a weak negative correlation with the counterproductive scale (see Table 3). The relative presenteeism score showed a weak positive correlation with the IWPQ subscales. These patterns of correlations did not differ across occupational sectors.

The overall work engagement score showed a moderate positive correlation with the IWPQ task and contextual scales, and an almost moderate negative correlation with the counterproductive scale. The work engagement subscales scores showed a weak to moderate positive correlation with the IWPQ task and contextual scales, and a weak to moderate negative correlation with the counterproductive scale. Again, this pattern did not differ across occupational sectors.

Table 3. Convergent validity: Expected and observed correlations of the IWPQ task performance (TP), contextual performance (CP), and counterproductive work behavior (CWB) scales with the HPQ absolute and relative presenteeism scores, and work engagement scores

HPQ absolute presenteeism score	**Expected correlation**	**Observed correlation**
TP scale	Strong (> 0.50)	Moderate (0.39) ^
CP scale	Strong (> 0.50)	Moderate (0.33) ^
CWB scale	Strong (< -0.50)	Weak (-0.16) ^
HPQ relative presenteeism score		
TP scale	Weak (< 0.30)	Weak (0.09) *
CP scale	Weak (< 0.30)	Weak (0.11) *
CWB scale	Weak (> -0.30)	Weak (0.07) *
UWES scores		
TP scale	Moderate (0.30-0.50)	Weak to moderate (0.26-0.35) ^
CP scale	Moderate (0.30-0.50)	Moderate (0.40-0.43) *
CWB scale	Moderate (-0.30- -0.50)	Weak to moderate (-0.23- -0.31) ^

Note. * Expectation confirmed.
^ Expectation party confirmed.

Discriminative validity

Job satisfaction quartiles differed significantly on task performance, contextual performance, and counterproductive work behavior (see Figure 1). There was a clear dose-response relationship between job satisfaction and individual work performance. Persons high in job satisfaction showing higher task and contextual performance, and lower counterproductive work behavior, than persons low in job satisfaction. All quartiles significantly differed from each other.

Overall health quartiles differed significantly on task performance, contextual performance, and counterproductive work behavior (see Figure 1). Again, there was a clear dose-response relationship between overall health and individual work performance. Persons in good overall health showing higher task and contextual performance, and lower counterproductive work behavior, than persons in poor overall health. Almost all quartiles significantly differed from each other. Again, these patterns did not differ across occupational sectors.

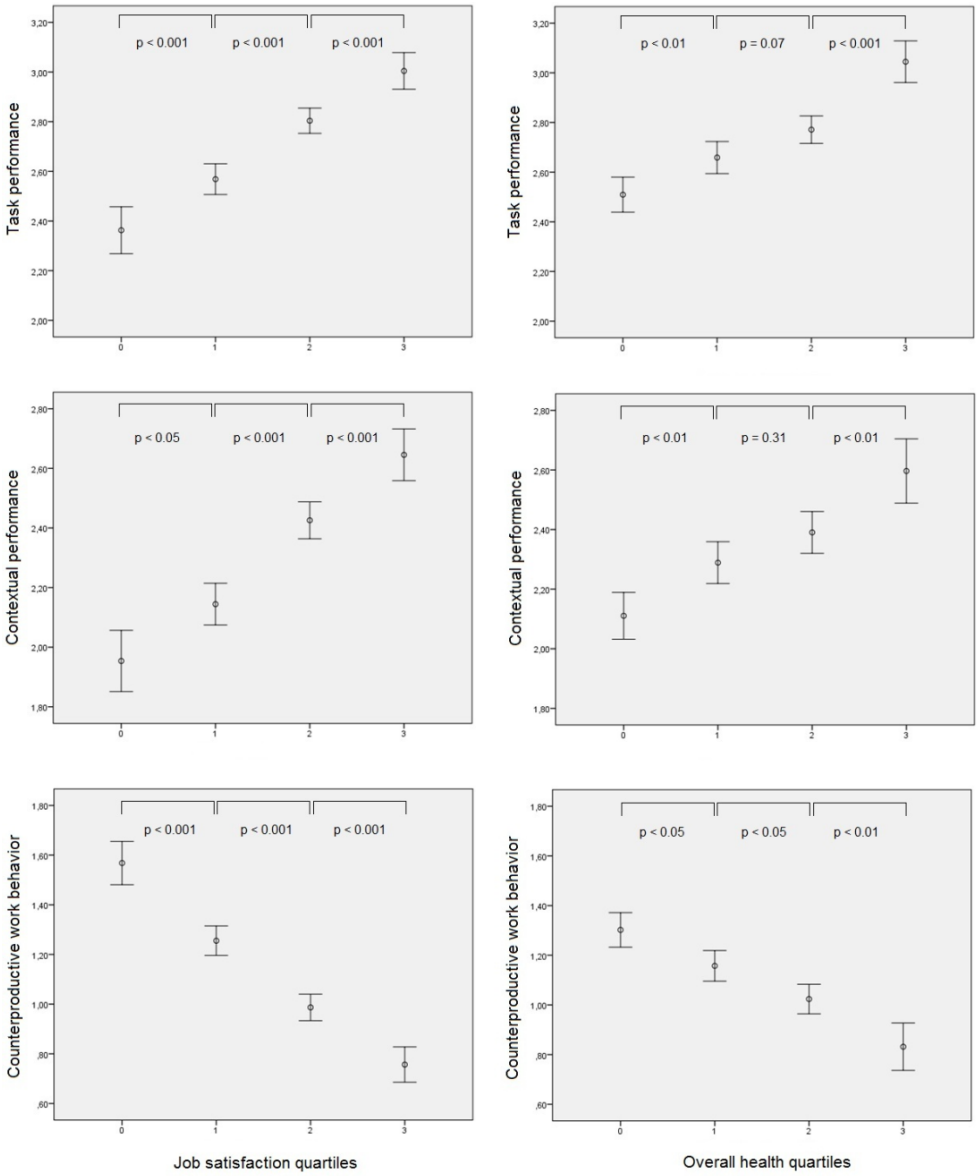


Figure 1. Discriminative validity of the IWPQ subscale scores (range 0-4) for job satisfaction quartiles and overall health quartiles (in the plots, the dot represents the mean score and the lines the 95% confidence interval).

Discussion

The current study expands research on the recently developed Individual Work Performance Questionnaire (IWPQ) [15] by examining its construct validity. Based on existing knowledge about the construct, the IWPQ largely provided the expected scores with regard to relationships to scores of other instruments, and with regard to differences between relevant groups. Terwee et al. [33] propose a standard of good construct validity when 75% of the hypotheses are confirmed. In the current study, 10 out of 15 (66.6%) hypotheses (5 constructs x 3 IWPQ scales) were confirmed. Overall, these findings indicate acceptable construct validity.

The correlations between the IWPQ subscale scores and the HPQ absolute presenteeism score were less strong than expected. Possibly, these weaker correlations were caused by limitations accompanying the administration of the HPQ in the current study. First, only the two presenteeism questions were administered in the current study. Although Kessler and colleagues [28] say that these questions can be abstracted from the complete HPQ, to our knowledge, the validity of these questions alone is unknown. The HPQ memory priming questions, intended to improve the accuracy of report, were omitted. Instead, the IWPQ questions were believed to be sufficient priming questions for participants to be able to give an overall rating of their work performance. Second, a longer recall period (3 months) was employed in the current study, instead of the original recall period (4 weeks). This was done to increase comparability between the questionnaires, and to avoid participants from needing to change their recall period while completing the questionnaires.

Another possible explanation for the weaker correlations is that the IWPQ concentrates on actual performance (what one typically does), whereas the HPQ absolute presenteeism question assesses actual performance in relation to possible performance (what one can maximally do). This subtle difference may make the two constructs less comparable than a priori expected, warranting moderate correlations. Furthermore, multi-item measures in which items measure the same construct are usually more reliable than single item measures [e.g., 31]. As IWP is a multi-dimensional construct, one overall question may not adequately reflect the full range of individual behaviors at work. The content that people have (or don't have) in mind when answering this question, may be substantially different for different persons, or from what researchers consider aspects of IWP. A finding to support this, is that the counterproductive work behavior scale correlated weakly with the HPQ

absolute presenteeism score. This suggests that people hardly take into account their counterproductive work behaviors, when rating their overall work performance. In the IWPQ, a more complete and generic picture of IWP is given, providing information about the separate components of IWP (a 'profile').

As expected, weak correlations of the IWPQ scores with the HPQ relative presenteeism score were found. This makes sense as the IWPQ does not ask workers to compare themselves with other workers, whereas this is the focus of the HPQ relative presenteeism score. Furthermore, participants may be uncomfortable rating the performance of their colleagues, or they may not have colleagues with similar jobs [34].

As expected, the IWPQ task and contextual performance scores showed moderate positive correlations with work engagement. The counterproductive work behavior score showed a weak to moderate negative correlation with work engagement. Although this last correlation was slightly lower than expected, the range of correlations was so close to expectations that this was not considered worrisome. The finding that engaged workers display more task and contextual performance behaviors, and less counterproductive work behaviors, is in line with previous studies which have shown that work engagement is positively related to IWP [e.g., 20,21]. Interestingly, the current study showed that work engagement was more strongly correlated with contextual behaviors, than with task or counterproductive behaviors. Thus, engaged workers mainly benefit their team or organization by engaging in behaviors that are not directly part of their central job tasks, but that do support the organizational, social, and psychological environment in which the central job tasks are performed.

Finally, the IWPQ was able to discriminate between relevant groups. Consistent with expectations, persons high in job satisfaction showed higher task and contextual scores, and lower counterproductive work behavior scores, than persons low in job satisfaction. Also, persons high in overall health showed higher IWPQ task and contextual scores, and lower IWPQ counterproductive work behavior scores, than persons low in overall health. The findings that satisfied workers, and healthy workers, perform better, is consistent with previous research [e.g., 22,23]. These findings also support the longstanding notion that happy, healthy workers are productive workers [e.g., 35].

Construct validity of related questionnaires

The construct validation process that we employed in the current study, appears to be similar to, or better than, the construct validation process of related questionnaires. Within the area of occupational health, several self-report questionnaires have been developed to measure losses in work performance (presenteeism or absenteeism). Contrary to the IWPQ, these questionnaires are mainly intended for persons with health complaints. The most used questionnaires in this area include the HPQ [7], Work Limitations Questionnaire [WLQ; 6] and Work Productivity and Activity Impairment questionnaire [WPAI; 5]. Although the construct validity of these questionnaires has been reasonably to well established, a limitation of their construct validation process was that no a priori expectations were specified on the strengths of the correlations, and often, only the significance of a correlation was presented, and not the strength of a correlation.

Within work and organizational psychology, numerous self-report scales have been developed to measure task performance [e.g., 10], contextual performance [e.g., 11], or counterproductive work behavior [e.g., 12]. These scales are mainly used for establishing the determinants of IWP. Concerning their validity, the main focus has been on examining the relationships between the different scales [e.g., 36]. Relationships to scores of other instruments, or differences between relevant groups, hardly appear to have been examined. Thus, there is little information of the construct validity of these questionnaires.

Lack of a golden standard

A type of validity that could not be examined in the present study, is criterion validity, defined as “the degree to which scores of an instrument are an adequate reflection of a gold standard” [37]. This is because a perfect gold standard seldom exists in practice [16], as was the case for the construct under study here. A perfect gold standard may be an objective measure of individual work performance, however, these are very hard to obtain in practice [38]. Especially for knowledge work or high complexity jobs, direct measures of countable behaviors or outcomes such as production quantity or number of errors made, are almost impossible.

Although construct validity is often considered to be less powerful than criterion validation, with strong theories and specific and challenging expectations, it is possible to acquire substantial evidence that the measurement instrument is measuring what it purports to measure [16]. One of the strengths of the current study, is that a clear conceptual framework of IWP was present, and specific and

challenging expectations based on theory or literature findings were formulated. This makes it possible to draw firm conclusions about the construct validity of the IWPQ scores.

Future research

Construct validation is an ongoing process [16], and therefore, more research should be conducted to create a strong web of evidence to support the validity of the IWPQ. Strictly speaking, one can never state that a measurement instrument is valid, only that it provides valid scores in the specific situation in which it has been tested [16]. Future research will also need to examine additional properties of the IWPQ, such as its sensitivity to change, and interpretability of change scores. Also, it can be hypothesized that determinants of IWP affect the separate dimensions differently (as was shown in the current study for work engagement). Therefore, relationships of determinants to the separate dimensions of IWP can be further examined, as well as outcomes related to the separate dimensions of IWP. Other interesting avenues include relating IWPQ scores to objective outcomes (e.g., absenteeism data, monetary outcomes), or by comparing self-, other-, and supervisor-ratings (360 degrees performance rating).

Conclusion

Based on the results of the current study, the IWPQ 1.0 showed acceptable overall construct validity. Its convergent validity proved to be sufficient, and its discriminative validity very good. The IWPQ provides researchers with a reliable and valid instrument to measure IWP generically, amongst workers from different occupational sectors, and workers with and without health problems.

References

1. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In Dunnette MD, Hough LM, eds. Handbook of industrial and organizational psychology, Vol.1 (2nd ed.). Palo Alto, CA, US: Consulting Psychologists Press; 1990:687-732.
2. Viswesvaran C, Ones DS. Perspectives on Models of Job Performance. International Journal of Selection and Assessment. 2000;8(4):216-226.
3. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: a policy-capturing approach. J Appl Psychol. 2002;87(1):66-80.
4. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, van der Beek AJ. Measuring Individual Work Performance: Identifying and Selecting Indicators. Work: Journal of Prevention, Assessment & Rehabilitation. 2013;45(3).
5. Reilly MC, Zbrozek AS, Dukes EM. The Validity and Reproducibility of a Work Productivity and Activity Impairment Instrument. Pharmacoeconomics. 1993;4(5).
6. Lerner D, Amick BC, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Medical Care. 2001;39(1):72-85.
7. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). Journal of Occupational and Environmental Medicine. 2003;45:156-174.
8. Koopmans L, Bernaards CB, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual frameworks of individual work performance: a systematic review. Journal of Occupational and Environmental Medicine. 2011;53(8):856-866.
9. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, eds. Personnel Selection in Organizations. San Francisco, CA: Jossey Bass; 1993. p. 71-98.
10. Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. Journal of Management. 1991;17(3):601-617.
11. Podsakoff PM, MacKenzie SB. A second generation measure of organizational citizenship behavior. Indiana University, Bloomington; 1989.

12. Bennett RJ, Robinson SL. Development of a Measure of Workplace Deviance. J Appl Psychol. 2000;85(3):349-360.
13. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol. 2005;90:1241-1255.
14. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Development of an Individual Work Performance Questionnaire. International Journal of Productivity and Performance Management. 2013;62(1):6-28.
15. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Improving the Individual Work Performance Questionnaire using Rasch Analysis. Journal of Applied Measurement. 2014;15(2).
16. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. Cambridge University Press; 2011.
17. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New Jersey: Lawrence Erlbaum Associates; 1988.
18. Schaufeli WB, Bakker AB. Utrecht Work Engagement Scale: Preliminary Manual. 2003;Version 1.
19. Schaufeli WB, Bakker AB, Salanova M. The Measurement of Work Engagement With a Short Questionnaire: A Cross-National Study. Educational and Psychological Measurement. 2006;66:701-716.
20. Bakker AB, Schaufeli WB, Leiter MP, Taris TW. Work Engagement: An Emerging Concept in Occupational Health Psychology. Work & Stress. 2008;22(3):187-200.
21. Demerouti E, Cropanzano R. From Thought to Action: Employee Work Engagement and Job Performance. In: Bakker AB, Leiter MP, eds. Work Engagement: A Handbook of Essential Theory and Research. New York: Psychology Press; 2010.
22. Harrison DA, Newman DA, Roth PL. How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. Academy of Management Journal. 2008;49(2):305-325.
23. Schultz AB, Edington DW. Employee Health and Presenteeism: A Systematic Review. Journal of Occupational Rehabilitation. 2007;17:547-579.
24. Locke EA. The nature and causes of job satisfaction. In: Dunnette MD, editor. Handbook of industrial and organizational psychology. Palo Alto, CA: Consulting Psychologists Press; 1976:1297-1349.

25. Judge TA, Bono JE, Thoresen CJ, Patton GK. The Job Satisfaction-Job Performance Relationship: A Qualitative and Quantitative Review. Psychological Bulletin. 2001;127(3):376-407.
26. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.
27. Lerner D, Mosher Henke R. What Does Research Tell Us About Depression, Job Performance, and Work Productivity? Journal of Occupational and Environmental Medicine. 2008;50(4):401-410.
28. Kessler RC, Petukhova M, McInnes K, Ustun TB. Content and Scoring Rules for the WHO HPQ absenteeism and presenteeism questions. 2007.
29. Wanous JP, Reichers AE, Hudy MJ. Overall job satisfaction: how good are single-item measures? J Appl Psychol. 1997;82(2):247-252.
30. Nagy MS. Using a single-item approach to measure facet job satisfaction. Journal of Occupational and Organizational Psychology. 2002;75:77-86.
31. Cunny KA, Perri M. Single-item vs multiple-item measures of health-related quality of life. Psychological Reports. 1991;69:127-130.
32. IBM Corp. IBM SPSS Statistics for Windows, Version 20.0. 2011.
33. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60:34-42.
34. Bernaards CM, Proper KI, Hildebrandt VH. Physical Activity, Cardiorespiratory Fitness, and Body Mass Index in Relationship to Work Productivity and Sickness Absence in Computer Workers With Preexisting Neck and Upper Limb Symptoms. Journal of Occupational and Environmental Medicine. 2007;49(6):633-640.
35. Fischer CD. Why do lay people believe that satisfaction and performance are correlated? Possible sources of a commonsense theory. Journal of Organizational Behavior. 2003;24:753-777.
36. Spector PE, Bauer JA, Fox S. Measurement Artifacts in the Assessment of Counterproductive Work Behavior and Organizational Citizenship Behavior: Do We Know What We Think We Know? J Appl Psychol. 2010;97(4):781-790.
37. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology,

and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63:737-745.

38. Jaramillo F, Carrillat FA, Locander WB. A meta-analytic comparison of managerial ratings and self-evaluations. Journal of Personal Selling & Sales Management. 2005;XXV(4):315-328.

Chapter 7

Responsiveness of the Individual Work Performance Questionnaire

Linda Koopmans, Jennifer K. Coffeng, Claire M. Bernaards, Cécile R.L. Boot, Vincent H. Hildebrandt, Henrica C.W. de Vet, Allard J. van der Beek

Article submitted for publication

7

Abstract

Background: Individual work performance is an important outcome measure in studies in the workplace. Nevertheless, its conceptualization and measurement has proven challenging. To overcome limitations of existing scales, the Individual Work Performance Questionnaire (IWPQ) was recently developed. The aim of the current study was to gain insight into the responsiveness of the IWPQ.

Methods: Data were used from the Be Active & Relax randomized controlled trial. The aim of the trial was to investigate the effectiveness of an intervention to stimulate physical activity and relaxation of office workers, on need for recovery. Individual work performance was a secondary outcome measure of the trial. In total, 39 hypotheses were formulated concerning correlations between changes on the IWPQ scales and changes on similar constructs (e.g., presenteeism) and distinct constructs (e.g., need for recovery) used in the trial.

Results: 260 Participants completed the IWPQ at both baseline and 12 months of follow-up. For the IWPQ scales, 23%, 15%, and 38%, respectively, of the hypotheses could be confirmed. In general, the correlations between change scores were weaker than expected. Nevertheless, at least 85% of the correlations were in the expected direction.

Conclusions: Based on results of the current study, no firm conclusions can be drawn about the responsiveness of the IWPQ. Several reasons may account for the weaker than expected correlations. Future research on the IWPQ's responsiveness should be conducted, preferably in other populations and intervention studies, where greater changes over time can be expected.

Introduction

Individual work performance, defined as "*employee behaviors or actions that are relevant to the goals of the organization*" [1], is an important outcome measure in studies in the workplace. Nevertheless, its conceptualization and measurement has proven challenging. First, consensus on a clear definition and conceptual framework of individual work performance (IWP) was lacking. Considering the diversity in conceptual frameworks of IWP, it is not surprising that numerous instruments have been developed to measure (aspects of) IWP. Due to the diversity in conceptual frameworks, the content validity of these existing instruments is questionable (e.g., do they measure the full range of individual work performance?). Also, generic applicability of these instruments is limited, because they are often developed for specific populations (e.g., for a specific occupation, or for workers with health complaints).

To overcome the aforementioned limitations, the Individual Work Performance Questionnaire (IWPQ) was recently developed [2, 3]. The IWPQ is based on a three-dimensional conceptual framework of IWP, which was developed after a systematic review of the literature [4]. This framework includes the dimensions of task performance ("*the proficiency with which individuals perform the core substantive or technical tasks central to his or her job*" [1]), contextual performance ("*behaviors that support the organizational, social and psychological environment in which the technical core must function*" [5]), and counterproductive work behavior ("*behavior that harms the well-being of the organization*" [6]). The IWPQ is a generic instrument, thus, it is suitable for workers in all types of occupations (i.e., blue, pink, and white collar workers) and workers with and without health complaints.

An important purpose of the IWPQ is to assess changes in IWP. For example, we may want to examine fluctuations in IWP over time (e.g., due to age), follow the effects of negative factors on IWP over time (e.g., health problems), or identify successful methods to improve IWP (e.g., intervention studies). In order to do this, the IWPQ must be responsive to changes over time. Responsiveness can be defined as "the ability of an instrument to detect change over time in the construct to be measured" [7]. There is a lot of confusion about the concept over responsiveness, and many different definitions and measures have been proposed over the past decades [8]. In addition, or perhaps, as a result, responsiveness is a seldom examined issue. When assessing responsiveness, we focus on the validity of a *change score*,

which is estimated on the basis of two or more measurement points [8]. The aim of the current study was to gain insight into the responsiveness of the IWPQ.

Methods

Participants

Data were used from the Be Active & Relax "Vitality in Practice" (VIP) randomized controlled trial [9]. The aim of the Be Active & Relax trial was to investigate the effectiveness of an intervention to stimulate physical activity and relaxation of office workers, on need for recovery. By means of stimulating physical activity and relaxation, work-related outcomes (e.g., sickness absenteeism, work engagement and individual work performance) were also expected to improve. The trial included a 2x2 factorial design with four research arms. The four arms consisted of a combined social and physical environmental intervention, a social environmental intervention only, a physical environmental intervention only and a control group. For the purpose of the current study, data of all four groups were taken together. This study was approved by the Medical Ethics Committee of the VU University Medical Center, Amsterdam, The Netherlands. Full details of the design of the Be Active & Relax trial have been reported elsewhere [9].

Measures

Measurements took place at baseline (T0), and at 6 months (T1) and 12 months (T2) follow-up. Only the measurements at baseline and at 12 months (T2) were used to assess responsiveness of the IWPQ.

Individual work performance was measured using the Individual Work Performance Questionnaire (IWPQ) [3, 10]. The IWPQ consists of 18 questions in three scales: task performance, contextual performance, and counterproductive work behavior. The IWPQ had a recall period of 3 months and a 5-point rating scale ("*seldom"* to *"always"* for task and contextual performance, *"never"* to *"often"* for counterproductive work behavior). The psychometric properties of the IWPQ have been tested and results indicated good to excellent reliability for task performance ($\alpha = 0.78$), contextual performance ($\alpha = 0.85$) and counterproductive work behavior ($\alpha = 0.79$). The IWPQ has shown good face and structural validity [2, 3, 10], as well as sufficient convergent validity and good discriminative validity [11].

Presenteeism was assessed through self-report with the World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) [12].

Presenteeism was assessed by asking participants to rate their actual performance in relation to possible performance. The score represents percentage of performance, and has a lower bound of 0 (*total lack of performance*) and an upper bound of 100 (*top performance*). The reliability and validity of the HPQ was examined for several occupations, and showed good convergent validity. However, poor validity was found for white collar workers [12, 13].

Job satisfaction was assessed using one overall question on a 5-point rating scale from "*highly dissatisfied*" to "*very satisfied*." A single-item measure of job satisfaction has been found to correlate highly with job satisfaction scales, and was therefore considered valid [14, 15].

Work engagement was measured using the Utrecht Work Engagement Scale (UWES) [16]. The UWES consists of three scales (vigour, dedication, and absorption), and a total of 17 items assessed on a 7-polint scale ranging from "*never*" to "*always*." The total score was calculated by adding the means of each scale, and dividing the sum by three. The psychometric properties of this questionnaire have been tested and results indicated an acceptable reliability of vigour (α = 0.68-0.80), dedication (α = 0.91) absorption (α = 0.73-0.75), and the total score (α = 0.93), as well as acceptable convergent validity [16].

Work ability was assessed using one question ("*How do you rate your current work ability compared to lifetime best?*") from the Work Ability Index (WAI) [17], on an 11-point rating scale from 0 "*completely unable to work*" to 10 "*at its best*." The single-item question is very strongly associated with the total WAI, and has shown good predictive validity [18].

Performance rating by the manager was assessed by asking one question ("*How would your manager rate your overall job performance, compared to colleagues in a similar job?")* on a 5-point rating scale from "*much worse*" to "*much better*." This question was adapted from the WHO-HPQ [12] presenteeism question, and previously used in The Netherlands Working Conditions Survey [19]. The reliability and validity of this question is unknown.

Self-rated work quality and quantity were assessed using one question each ("*How do you rate the quality of your own work?" and "How do you rate the quantity of your own work?")* on a 5-point rating scale from "*insufficient*" to "*excellent*." The reliability and validity of these questions is unknown.

Need for recovery (NFR) was assessed using the Need for Recovery after Work scale [20]. This Dutch version of the Questionnaire on the Experience and Evaluation of Work (Dutch abbreviation: VBBA) consists of eleven dichotomous

items (yes/no), representing short-term effects of a day at work. The NFR score is a percentage score (0 to 100) of positive answers of those providing data for at least 8 of the 11 items. The Need for Recovery after Work scale has shown good reliability (α = 0.86-0.88), construct validity, and sensitivity to change in The Netherlands [20-22].

Physical activity was assessed using the Short Questionnaire to Assess Health Enhancing Physical Activity (SQUASH) [23]. Duration and intensity of active commuting, leisure time activities, sport activities, household activities, and physical activities at work (standing and walking), were assessed. For each domain, employees were asked to report the frequency (i.e., times per week), duration of activities (i.e., in minutes), and self-reported intensity (i.e., light, moderate or vigorous). Total scores for minutes per week spent on light, moderate, and vigorous physical activities were calculated. The SQUASH scores have shown reasonable reproducibility (r = 0.57-0.58) and validity against accelerometry (r = 0.45-0.67), which is comparable to other physical activity questionnaires [23, 24].

General health and vitality were measured using the Dutch version of the Rand-36 [25]. General health was measured by asking workers to indicate how they perceived their general health, on a 5-point scale from "*poor*" or "*excellent*." Vitality was measured with a scale of 5 items, asking workers to indicate how often they felt full of life, worn out, tired and full of energy, on a 6-point scale from "*never*" to "*always*." This scale was transformed to a 0-100 score, with higher scores indicating higher vitality. The Dutch version of the Rand-36 has shown good reliability for the vitality scale (α = 0.82) and had reasonable construct validity [25].

Exhaustion was measured using the OLdenburg BurnOut Inventory (OLBI) [26]. The OLBI consists of eight items on a 4-point scale ranging from "*totally disagree*" to "*totally agree*." A mean score was calculated. The OLBI has shown good reliability (α = 0.80-0.85) and reasonable convergent and discriminant validity in different occupational groups [26, 27].

Sickness absenteeism data were retrieved from company records, for the year prior to the intervention (i.e. baseline), and for the year of the intervention (i.e., 12 month follow-up). The score represents the number of workdays absent per year.

Hypotheses

A construct approach of responsiveness testing [8] was applied in the current study, which means that hypotheses were formulated concerning relationships between changes on the IWPQ and changes on other instruments used in the Be Active &

Relax trial. These were divided into hypotheses with similar constructs (e.g., presenteeism) and distinct constructs (e.g., need for recovery). Stronger correlations of the IWPQ scales were hypothesized with constructs similar to IWP than constructs distinct from IWP. Expectations were formulated per IWPQ scale, resulting in a total of 39 hypotheses (3 IWPQ scales x 13 constructs). If positive correlations were expected for task and contextual performance, negative correlations were expected for counterproductive work behavior, and vice versa (also see Table 2).

With similar constructs

The change in each IWPQ scale was expected to correlate moderately (0.30-0.50 or -0.50 - -0.30) with the change in presenteeism [11], job satisfaction [e.g., 28], work engagement [e.g., 29], work ability [e.g., 30], performance rating by the manager [31], work quality, and work quantity. Based on literature, the change in counterproductive work behavior was expected to correlate weakly or not at all (-0.20 – 0.20) with the change in the last three constructs [32].

With distinct constructs

The change in each IWPQ scale was expected to correlate weakly (0.20 – 0.30 or -0.30 –-0.20) with the change in need for recovery [e.g., 27, 33], physical activity [e.g., 34], general health [e.g., 35, 36], vitality [e.g., 37], and exhaustion [e.g., 38]. Finally, the change in each IWPQ scale was expected to correlate weakly or not at all (-0.20 – 0.20) with the change in sickness absenteeism [39, 40].

Data analysis

Pearson correlations between the change scores of each IWPQ scale and the change scores on the other constructs were calculated for the change scores from baseline (T0) to 12 months (T2). Only participants who completed the IWPQ at both T0 and T2 were included in the data analysis. Analyses were conducted in SPSS 20.0 [41].

Results

Descriptive statistics of the participants

Of the 412 participants in the Be Active & Relax trial, 260 participants (63%) completed the IWPQ at both baseline and 12 months. At baseline (n=260), participants had a mean age of 43.2 years (SD = 9.9), and 37% was female.

Descriptive statistics of the IWPQ scales and the other constructs

Table 1 presents the mean scores and standard deviations (SD) on the IWPQ scales and the other constructs at baseline (T0) and 12 months (T2). It also reports the mean and standard deviation (SD_{change}) of the change scores on the IWPQ scales and the other constructs from T0 to T2.

Table 1. Mean scores (and SD) and mean change scores (and SD_{change}) on the IWPQ scales and the similar/distinct constructs at baseline (T0) and 12 months (T2)

	T0 (baseline) Mean (SD)	**T2 (12 months)** Mean (SD)	**Change score T2-T0** Mean (SD_{change})
IWPQ (1-5)			
Task performance	3.46 (0.68)	3.63 (0.66)	0.17 (0.70)
Contextual performance	3.34 (0.71)	3.39 (0.79)	0.04 (0.69)
Counterproductive work behavior	2.23 (0.65)	2.16 (0.66)	-0.07 (0.64)
Similar constructs			
Presenteeism (0-100)	76.58 (8.76)	75.87 (10.62)	-0.79 (11.51)
Job satisfaction (1-5)	3.96 (0.73)	3.85 (0.75)	-0.11 (0.80)
Work engagement (1-7)	4.91 (0.85)	4.84 (0.93)	-0.07 (0.71)
Work ability (1-10)	7.79 (1.42)	7.70 (1.57)	-0.08 (1.56)
Performance rating by the manager (1-5)	3.41 (0.81)	3.46 (0.81)	0.06 (0.81)
Self-rated work quality (1-5)	3.83 (0.79)	3.63 (0.87)	-0.20 (0.95)
Self-rated work quantity (1-5)	3.87 (0.83)	3.74 (0.92)	-0.12 (0.95)
Distinct constructs			
Need for recovery (0-100)	32.20 (29.26)	27.78 (28.71)	-2.40 (23.70)
Physical activity (min/week)			
Light	1810.10 (1363.68)	1603.23 (1618.94)	-199.40 (1785.64)
Moderate	281.81 (254.19)	350.94 (633.98)	72.66 (629.00)
Vigorous	83.53 (160.15)	99.79 (272.90)	9.40 (266.15)
General health (1-5)	3.35 (0.85)	3.37 (0.84)	0.79 (1.53)
Vitality (0-100)	64.08 (18.84)	65.72 (17.97)	1.87 (15.17)
Exhaustion (1-4)	2.15 (0.48)	2.15 (0.46)	0.04 (0.40)
Sickness absenteeism (workdays absent per year)	7.55 (21.81)	7.37 (20.91)	0.55 (25.03)

Correlations between change scores

Table 2 presents the expected and observed correlations between the change scores of the IWPQ scales and the change scores of the other constructs. For task performance, 85% of the correlations were in the expected direction, and for contextual performance and counterproductive work behavior, 92% of the correlations were in the expected direction. However, in many cases, the correlations were weaker than expected.

For the task performance scale, 3 out of 13 (23%) hypotheses were fully confirmed. As expected, the change in task performance correlated moderately positive with the changes in vitality (r = 0.23), moderately negatively with the change in exhaustion (r = -0.23), and weakly negative with the change in absenteeism (r = -0.14).

For the contextual performance scale, 2 out of 13 (15%) hypotheses were fully confirmed. As expected, the change in contextual performance correlated moderately positive with the change in vitality (r = 0.29), and weakly negative with the change in absenteeism (r = -0.08). Furthermore, the correlation between the change in contextual performance and the changes in most of the similar constructs (e.g., presenteeism, work engagement, work ability) approached the 0.30 correlation strength.

For the counterproductive work behavior scale, 5 out of 13 (38%) hypotheses were fully confirmed. As expected, the change in counterproductive work behavior correlated weakly with the changes in rating by the manager (r = -0.02), work quality (r = -0.06), work quantity (r = 0.02), and absenteeism (r = -0.09), and moderately positive with the change in exhaustion (r = 0.23).

Table 2. Pearson correlations (E = expected, O = observed) between **change scores** of the IWPQ scales and similar/distinct constructs

	IWPQ scale		
	Task performance	**Contextual performance**	**Counterproductive work behavior**
Similar constructs			
Presenteeism	E: 0.30 – 0.50 O: 0.18	E: 0.30 – 0.50 O: 0.22	E: -0.50 – -0.30 O: -0.11
Job satisfaction	E: 0.30 – 0.50 O: 0.12	E: 0.30 – 0.50 O: 0.17	E: -0.50 – -0.30 O: -0.24
Work engagement	E: 0.30 – 0.50 O: 0.19	E: 0.30 – 0.50 O: 0.29	E: -0.50 – -0.30 O: -0.23
Work ability	E: 0.30 – 0.50 O: 0.16	E: 0.30 – 0.50 O: 0.26	E: -0.50 – -0.30 O: -0.23
Performance rating by the manager	E: 0.30 - 0.50 O: 0.16	E: 0.30 - 0.50 O: 0.22	E: -0.20 – 0.20 O: -0.02 *
Work quality	E: 0.30 – 0.50 O: 0.20	E: 0.30 – 0.50 O: 0.18	E: -0.20 – 0.20 O: -0.06 *
Work quantity	E: 0.30 – 0.50 O: 0.11	E: 0.30 – 0.50 O: 0.19	E: -0.20 – 0.20 O: 0.02 *
Distinct constructs			
Need for recovery	E: -0.30 – -0.20 O: -0.15	E: -0.30 – -0.20 O: -0.11	E: 0.20 – 0.30 O: 0.16
Physical activity	E: 0.20 – 0.30	E: 0.20 – 0.30	E: -0.30 – -0.20
Light	O: -0.09	O: -0.04	O: -0.07
Moderate	O: 0.03	O: 0.03	O: -0.07
Vigorous	O: -0.05	O: 0.00	O: -0.04
General health	E: 0.20 – 0.30 O: -0.07	E: 0.20 – 0.30 O: 0.08	E: -0.30 – -0.20 O: 0.02
Vitality	E: 0.20 – 0.30 O: 0.23 *	E: 0.20 – 0.30 O: 0.29 *	E: -0.30 – -0.20 O: -0.03
Exhaustion	E: -0.30 – -0.20 O: -0.23 *	E: -0.30 – -0.20 O: -0.13	E: 0.20 – 0.30 O: 0.23 *
Sickness absenteeism	E: -0.20 - 0.20 O: -0.14 *	E: -0.20 - 0.20 O: -0.08 *	E: -0.20 - 0.20 O: -0.09 *
Hypotheses:			
Confirmed	23%	15%	38%
In the right direction	85%	92%	92%

Note: E = expected correlation, O = observed correlation. * = Confirmed hypothesis.

Discussion

The aim of the current study was to examine the responsiveness of the IWPQ, i.e., the ability of the IWPQ to detect change over time. A total of 39 hypotheses were formulated concerning the relationships between changes on the IWPQ and changes on similar constructs (e.g., presenteeism) and distinct constructs (e.g., need for recovery) used in the Be Active & Relax trial. For the IWPQ task performance, contextual performance, and counterproductive work behavior scales, 23%, 15%, and 38%, respectively, of the hypotheses could be confirmed. As hypothesized, the correlations of the IWPQ scales were slightly stronger with similar constructs than with distinct constructs, on average. However, in general, the correlations between change scores were weaker than expected. Nevertheless, most of the correlations (at least 85%) were in the expected direction. Exceptions were the correlations between the change scores of task performance and light and intense physical activity (r = -0.09 and -0.05, respectively), task performance and general health (r = -0.07), contextual performance and light physical activity (r = -0.04), and counterproductive work behavior and general health (r = 0.02).

Several reasons may account for the weaker than expected correlations. First, the IWPQ questions may not be sensitive enough to pick up changes in IWP over time. Also, it is hard to say how a change from answer categories “regularly” to “often” can be achieved. What needs to be done to accomplish a change from “regularly” to “often,” e.g., in keeping your work results in mind? And what does this change mean? In sum, the questions of the IWPQ scales may lack discriminative ability. However, in the developmental phase of the IWPQ scales, Rasch analysis [42] was performed to make sure that those items with a high discrimination parameter (i.e., high slope) were retained in the IWPQ 1.0 [2, 3]. Also, in the construct validation phase of the IWPQ scales, the IWPQ 1.0 was able to differentiate between known groups [11]. This suggests that the items in the IWPQ scales should have enough discriminative ability to detect changes in IWP over time.

Possibly, low responsiveness of the IWPQ could be caused by ceiling and floor effects in the scales. Although previous examination of the IWPQ using Rasch analysis has shown that the items of the IWPQ are relatively well-distributed over the scales, persons continue to score relatively high on task performance (ceiling effect), and low on CWB (floor effect) [3]. This could be caused by the tendency of persons to evaluate and present themselves in a socially desirable, favorable way [43, 44]. As a consequence of the ceiling and floor effects, it becomes hard to detect

further improvements in task performance, and further decreases in CWB. Thus, the ability to detect changes at the high part of the task performance scale, and low part of the CWB scale, may be diminished.

Another possible reason for the lower than expected correlations may lie in the study population. As said before, the population in the current study consisted of relatively healthy, well-functioning office workers who, in general, scored high on constructs such as general health, presenteeism, and job satisfaction, and low on constructs such as need for recovery, exhaustion, and sickness absenteeism. This makes it hard to obtain or detect any further improvements in this population. Despite the use of an intervention, small changes on the constructs over the 12-month intervention period were obtained. When examining the scatterplots of the change scores, low spread on many constructs can be observed (i.e., dots clustered in the middle), and this can cause deflated correlations [8].

Finally, a reason for the lower than expected correlations may be that the intervention was not effective enough to obtain changes in IWP. The primary aim of the Be Active & Relax study was to investigate the effectiveness of an intervention to stimulate physical activity and relaxation of office workers, on need for recovery [9]. Indirectly, physical activity and relaxation were expected to improving IWP. However, and it may be that the intervention was not specific or intense enough to obtain improvements in IWP. Despite the fact that the intervention was not directly targeted at IWP, and despite high baseline levels on the constructs, a statistically significant increase in tasks performance (B = 0.2, 95% CI 0.0; 0.4), and a statistically significant decrease in contextual performance (B = -0.3, 95% CI -0.4; 0.1), were detected in the Be Active & Relax study [45]. The decrease in contextual performance could be explained by the fact that participants in the intervention groups were stimulated to engage in physical activity and relaxation during the workday, and this possibly could have reduced taking on extra work tasks, for example. Thus, this study showed that the IWPQ is able to detect statistically significant changes in individual work performance over time.

Assessment of responsiveness

As stated in the Introduction, there is a lot of confusion about the concept over responsiveness, and many different definitions and measures have been proposed over the past decades [8]. In addition, or perhaps, as a result, responsiveness is a seldom examined issue. For example, Abma et al. [46] reviewed the measurement properties of five self-report (health-related) work functioning instruments; the

EWPS, WLQ, SPS, WPS, and LEAPS. For all five instruments, the methodological quality of responsiveness testing was poor, or not studied. Of the instruments used in the current study, only the responsiveness of the Need for Recovery Scale was examined. Based on effect sizes, the responsiveness of this scale appeared to be good [21]. However, the responsiveness of the other questionnaires used in the current study remains unknown. This is a limitation of the responsiveness testing process, because responsiveness of a new questionnaire is tested against change scores of existing questionnaires, whose responsiveness is also unknown, and may be poor.

No golden standard or clear guidelines seem to exist for the assessment of responsiveness and the interpretation of results. De Vet and colleagues [8] stated that responsiveness is often examined based on inappropriate outcome measures, such as effect sizes or standardized response mean. They advise that responsiveness should be seen as a form of longitudinal validity, using either a criterion approach (if a gold standard is available) or a construct approach (testing hypotheses of change scores).

In addition to the lack of clarity on how responsiveness should be tested, there are no clear guidelines as to what the strength of correlations between change scores should be. A final reason for the large percentage of unconfirmed hypotheses in the current study, may be that the hypothesized correlations (r = 0.30-0.50) were too high to begin with. In line with Cohen [47], we interpreted a correlation coefficient over 0.50 as strong, 0.30 to 0.50 as moderate, 0.10 to 0.30 as weak, and below 0.10 as no relation between constructs at all. Often, Cohen's guidelines are used for cross-sectional correlations, i.e., when a correlation between two different measurement scores obtained at the same point in time is examined (thus, there is only one measurement). When it comes to correlations between change scores (multiple measurements), it is based on two measurements, and a double measurement error is involved. Due to this double measurement error, it seems reasonable that lower correlations may be expected. This issue has been addressed by other researchers. For example, Abma et al. [48] examined the responsiveness of the Work Role Functioning questionnaire, and they hypothesized correlation sizes around 0.20 to 0.30 with other constructs, because it was expected that many participants would show no changes, and based on results in earlier studies with similar questionnaires. For the constructs used in the current study, previous research has shown that, for example, the cross-sectional correlation between IWP and work engagement ranges between r = 0.30-0.50 [e.g., 49]. It is therefore

questionable whether correlations of r = 0.30-0.50 between their change scores can reasonably be expected. Such high correlations between change scores would likely be obtained for identical constructs, rather than similar (but not identical) constructs.

Recommendations for future research

The responsiveness of the IWPQ should be further examined in future research, to determine whether its responsiveness is truly low, or whether the low responsiveness found in the current study was caused by limitations of the current study. We therefore recommend examining the responsiveness of the IWPQ in different populations, preferably in populations with low(er) baseline levels on the constructs, where large(r) changes on the constructs over time can be expected. Suggestions for such populations could be a sample of workers with work-related musculoskeletal health problems, mental health problems, and/or low job satisfaction. An intervention study, which is directly aimed at improving IWP, could obtain greater changes in these populations, making it easier to detect changes in IWP and related constructs. Suggestions for such a study could be an intervention focusing on managerial style, technological improvements at work, and/or job skills training. Also, the responsiveness of the IWPQ should preferably be examined using other measurement instruments of which the responsiveness is known. Finally, the responsiveness of questionnaires deserves greater attention, and clear guidelines for assessing and interpreting responsiveness should be adopted. The guidelines proposed by Terwee et al. [50], Mokkink et al. [51], and De Vet et al. [8] could provide a good starting point for this.

Conclusion

Based on results of the current study, no firm conclusions can be drawn about the responsiveness of the IWPQ. Overall, most of the correlations between changes on the IWPQ scales and changes on other constructs were in the expected direction, although not as high as expected. This might indicate low responsiveness of the IWPQ. However, the weaker than expected correlations may also be accounted for by characteristics of the intervention study, such as the relatively healthy, well-functioning study population, and an intervention study that was not primarily aimed at IWP. Nevertheless, the IWPQ was able to show statistically significant changes in IWP during baseline and 12 months follow-up. Future research should provide more information about the responsiveness of the IWPQ, preferably in other populations and intervention studies.

References

1. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In Handbook of industrial and organizational psychology. Volume 1. 2nd edition. Edited by Dunnette MD & Hough, LM. (1990). Palo Alto, CA, US: Consulting Psychologists Press; 1990:687-755.
2. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Development of an individual work performance questionnaire. International Journal of Productivity and Performance Management 2013;62(1):6-28.
3. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Improving the individual work performance questionnaire using rasch analysis. Journal of Applied Measurement 2014;15(2).
4. Koopmans L, Bernaards CM, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual frameworks of individual work performance: A systematic review. Journal of Occupational and Environmental Medicine 2011;53(8):856-66.
5. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Personnel Selection in Organizations. Edited by Schmitt N, Borman WC. San Francisco, CA: Jossey Bass; 1993. p. 71-98.
6. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. J Appl Psychol 2002;87(1):66-80.
7. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 2010;63:737-45.
8. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge University Press; 2011.
9. Coffeng JK, Hendriksen IJM, Duijts SF, Proper KI, Van Mechelen W, Boot CRL. The development of the be active & relax vitality in practice (VIP) project and design of an RCT to reduce the need for recovery in office employees. BMC Public Health 2012;12:592.

10. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Measuring individual work performance: Identifying and selecting indicators. Work: Journal of Prevention, Assessment & Rehabilitation 2013; 45(3).
11. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Construct validity of the individual work performance questionnaire. Journal of Occupational and Environmental Medicine, 2014;56(3).
12. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, Pronk N, Simon G, Ustun TB, Wang P. The world health organization health and work performance questionnaire (HPQ). Journal of Occupational and Environmental Medicine 2003;45:156-74.
13. Kessler RC, Ames M, Hymel PA, Loeppke R, McKenas DK, Richling DE, Stang PE, Ustun TB. Using the world health organization health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. Journal of Occupational and Environmental Medicine 2004;46(6):S23-37.
14. Wanous J.P., Reichers AE, Hudy MJ. Overall job satisfaction: How good are single-item measures? Journal of Applied Psychology 1997;82(2):247-52.
15. Nagy MS. Using a single-item approach to measure facet job satisfaction. Journal of Occupational and Organizational Psychology 2002;75:77-86.
16. Schaufeli WB, Bakker AB. Utrecht Work Engagement Scale: Preliminary Manual. Occupational Health Psychology Unit, Utrecht University; 2003. Version 1.
17. Ilmarinen J. The work ability index (WAI). Occupational medicine (Oxford, England). 2007;57:160.
18. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. The work ability index and single-item question: Associations with sick leave, symptoms, and health - a prospective study of women on long-term sick leave. Scandinavian Journal of Work, Environment and Health 2010;36(5):404-12.
19. Koppes LLJ, De Vroome EMM, Mol MEM, Janssen BJM, Van den Bossche SNJ. The Netherlands working conditions survey. PlantijnCasparie, Almere: The Netherlands Organisation for Applied Scientific Research; 2008.
20. Van Veldhoven M, Broersen S. Measurement quality and validity of the "need for recovery scale". Occupational and environmental medicine 2003;60(Suppl 1):i3-9.

21. De Croon EM, Sluiter JK, Frings-Dresen MH. Psychometric properties of the need for recovery after work scale: Test-retest reliability and sensitivity to detect change. Occupational and environmental medicine 2006;63(3):202-6.
22. Van Veldhoven MJ, Sluiter JK. Work-related recovery opportunities: Testing scale properties and validity in relation to health. International Archive of Occupational and Environmental Health 2009;82(9):1065-75.
23. Wendel-Vos GC, Schuit AJ, Saris WH, Kromhout D. Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. Journal of clinical epidemiology 2003;56(12):1163-9.
24. Wagenmakers R, van den Akker-Scheek I, Groothoff JW, Zijlstra W, Bulstra SK, Kootstra JW, Wendel-Vos GC, Van Raaij JJ, Stevens M. Reliability and validity of the short questionnaire to assess health-enhancing physical activity (SQUASH) in patients after total hip arthroplasty. BMC Musculoskelet.Disord. 2008;17(9):141.
25. Van der Zee KI, Sanderman R. Het meten van de algemene gezondheidstoestand met de rand-36: Een handleiding (2nd ed.). Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken, NCG; 2012.
26. Demerouti E, Bakker AB, Vardakou I, Kantas A. The convergent validity of two burnout instruments: A multitrait-multimethod analysis. European Journal of Psychological Assessment 2003;19(1):12-23.
27. Demerouti E, Bakker AB. The oldenburg burnout inventory: A good alternative to measure burnout and engagement. In: Halbesleben JRB, editor. Handbook of Stress and Burnout in Health Care. Happauge, NY: Nova Science; 2008.
28. Judge TA, Bono JE, Thoresen CJ, Patton GK. The job satisfaction-job performance relationship: A qualitative and quantitative review. Psychological Bulletin 2001;127(3):376-407.
29. Bakker AB, Demerouti E. Towards a model of work engagement. Career Development International 2008;13(3):209-23.
30. Van den Berg TIJ. The role of work ability and health on sustaining employability [dissertation]. Erasmus University Rotterdam; 2010.
31. Harris MM, Schaubroeck J. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology 1988;41:43-62.
32. Dalal RS. A meta-analytic comparison of managerial ratings and self-evaluations. Journal of Personal Selling & Sales Management 2005;XXV(4):315-28.

33. Fritz C, Sonnentag S. Recovery, health, and job performance: Effects of weekend experiences. Journal of occupational health psychology 2005;10(3):187-99.
34. Pronk NP, Martinson B, Kessler RC, Beck AL, Simon GE, Wang P. The association between work performance and physical activity, cardiorespiratory fitness, and obesity. Journal of Occupational and Environmental Medicine 2004;46(1):19-25.
35. Schultz AB, Edington DW. Employee health and presenteeism: A systematic review. Journal of Occupational Rehabilitation 2007;17:547-79.
36. Boles M, Pelletier B, Lynch W. The relationship between health risks and work productivity. Journal of Occupational and Environmental Medicine 2004;46(7):737-45.
37. Wright TA, Bonett DG, Sweeney DA. Mental health and work performance: Results of a longitudinal field study. Journal of Occupational and Organizational Psychology 1993;66:277-84.
38. Wright TA, Cropanzano R. Emotional exhaustion as a predictor of job performance and voluntary turnover. Journal of Applied Psychology 1998;83(3):486-93.
39. Bycio P. Job performance and absenteeism: A review and meta-analysis. Human Relations 1992;45(2):193-220.
40. Morrow PC, McElroy JC, Laczniak KS, Fenton JB. Using absenteeism and performance to predict employee turnover: Early detection through company records. Journal of vocational behavior 1999;55:358-74.
41. IBM Corp. IBM SPSS statistics for windows, version 20.0. 2011.
42. Rasch G. Probabalistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
43. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology 2003;88(5):879-903.
44. Van der Heijden BIJM, Nijhof AHJ. The value of subjectivity: Problems and prospects for 36-degree appraisal systems. The International Journal of Human Resource Management 2004;15(3):493-511.
45. Coffeng JK, Hendriksen IJM, Duijts SF, Proper KI, Van Mechelen W, Boot CRL. Effectiveness of a combined social and physical environmental intervention on work-related outcomes in office employees, submitted for publication.

46. Abma FI, Van der Klink JJL, Terwee CB, Amick III BC, Bultmann U. Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. Scandinavian Journal of Work, Environment and Health 2012;38(1):5-18.
47. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New Jersey: Lawrence Erlbaum Associates; 1988.
48. Abma FI, van der Klink JJ, Bültmann U. The work role functioning questionnaire 2.0 (Dutch version): Examination of its reliability, validity and responsiveness in the general working population. Journal of Occupational Rehabilitation 2013;23(1):135-47.
49. Demerouti E, Bakker AB. Employee well-being and job performance: Where we stand and where we should go. In: Houdmont J, McIntyre S, editors. Occupational Health Psychology: European Perspectives on Research, Education and Practice. 1st ed. Maia: ISMAI Publications; 2006.
50. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Quality of Life Research 2003;12(4):349-62.
51. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, Knol DL, Bouter LM, De Vet HCW. Inter-rater agreement and reliability of the COSMIN (COnsensus-based standards for the selection of health status measurement instruments) checklist. BMC Medical Research Methodology 2010;10(82).

Chapter 8

Cross-Cultural Adaptation of the Individual Work Performance Questionnaire

Linda Koopmans, Claire M. Bernaards, Vincent H. Hildebrandt,
Debra Lerner, Henrica C.W. de Vet, Allard J. van der Beek

Article submitted for publication

8

Abstract

Objectives: The study objectives were to perform a cross-cultural adaptation of the Individual Work Performance Questionnaire (IWPQ) from the Dutch to the American-English language, and to assess the questionnaire's internal consistency and content validity in the American-English context.

Methods: The Dutch IWPQ was translated and adapted to the American-English language using guidelines of Beaton et al. (2000). The process consisted of five steps: a forward translation by two independent translators, synthesis, back-translation by two other independent translators, an expert committee review, and pilot-testing. During the pilot-testing, cognitive interviews with 40 American workers were performed, to examine the comprehensibility, applicability, and completeness of the American-English IWPQ.

Results: The questionnaire translation was conducted without major difficulties. The questionnaire instructions were slightly modified to aid interpretation in the American-English language. Inconsistencies with verb tense were identified, and it was decided to consistently use the simple past for the American-English version. In general, participants were positive on the comprehensibility, applicability and completeness of the questionnaire during the pilot-testing phase. Items TP3 and TP4 might warrant further attention. Furthermore, the study showed good results concerning the internal consistency of the American-English IWPQ (Cronbach's alphas for the scales between 0.79 and 0.89) and good content validity.

Conclusion: The results indicate that the cross-cultural adaptation of the American-English IWPQ was successful and that the measurement properties of the translated version are good.

Introduction

In today's world, it is increasingly important to maintain, improve, and optimize individual work performance (IWP) of employees. In Europe and the United States of America, for example, the 'grey wave' (i.e., accelerated growth of the older working population and a decline in numbers of the younger working population) and the economic recession force companies and employees to perform more or better work with less people. Also, due to the grey wave, the retirement age of older workers has been prolonged [e.g., 1]. Thus, their IWP has to be maintained until a later age. In order to accurately establish the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP, valid measurement of IWP is a prerequisite.

IWP, defined as "*behaviors or actions that are relevant to the goals of the organization*", is since long considered to be a multidimensional construct [2,3]. Based on several reviews of the literature [4-6], it can be concluded that IWP consists of three broad dimensions: task performance, contextual performance, and counterproductive work behavior. The first dimension, *task performance*, traditionally has received most attention, and can be defined as "the proficiency with which individuals perform the core substantive or technical tasks central to their job" [2]. The second dimension of IWP is *contextual performance*, defined as "behaviors that support the organizational, social and psychological environment in which the technical core must function" [7]. The third dimension of IWP is *counterproductive work behavior*, defined as "behavior that harms the well-being of the organization" [5].

Recently, the Individual Work Performance Questionnaire (IWPQ) [8,9] was developed in The Netherlands. The IWPQ is the first questionnaire to incorporate all relevant dimensions of IWP into one questionnaire. An advantage of this is that the content of each scale is fitted to the content of the other scales. As a result, the scales do not include antithetical items, that is, items overlapping in content [10]. Another advantage of the IWPQ is that it is generically applicable. Previous questionnaires to measure IWP were often developed for, or refined based on, specific populations, such as workers in specific jobs [e.g., 11,12] or workers with health problems [e.g., 13,14]. The IWPQ can be used for workers in blue, pink, and white collar jobs, and workers with and without health problems [e.g., 8,9].

Considering the advantages of the IWPQ as a comprehensive and generic measurement instrument of IWP, it seems especially suitable for examining the

effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP. In order for the IWPQ to be used outside of The Netherlands, it has to be cross-culturally adapted and validated. Because of possible cultural differences between countries, instruments need to be systematically translated, adapted, and validated before they can be used in other cultural contexts. Beaton et al. [15] have proposed a guideline for cross-cultural translation and adaptation, that consists of five steps: a forward translation by two independent translators, synthesis, back-translation by two other independent translators, an expert committee review, and pilot-testing. In the pilot-testing phase, cognitive interviews are held with people from the target population, in order to get an understanding of the comprehensibility, applicability, and completeness of the translated questionnaire. The objectives of the current study were to perform a cross-cultural adaptation of the Individual Work Performance Questionnaire (IWPQ) from the Dutch to the American-English language, and to assess the questionnaire's internal consistency and content validity in the American-English context.

Methods

Individual Work Performance Questionnaire

The Individual Work Performance Questionnaire (IWPQ) [8,9] measures "*employee behaviors or actions that are relevant to the goals of the organization*" [2]. The IWPQ consists of 18 items, divided into three scales: task performance, contextual performance, and counterproductive work behavior (see Table 1). All items have a recall period of 3 months and a 5-point rating scale ("*seldom"* to *"always"* for task and contextual performance, *"never"* to *"often"* for counterproductive work behavior). A mean score for each IWPQ scale can be calculated by adding the item scores, and dividing their sum by the number of items in the scale. Hence, the IWPQ yields three scale scores that range between 0 and 4, with higher scores reflecting higher task and contextual performance, and higher counterproductive work behavior.

Cross-cultural adaptation

The IWPQ's cross-cultural adaptation process followed the guidelines of Beaton et al. [15], pictured in Figure 1.

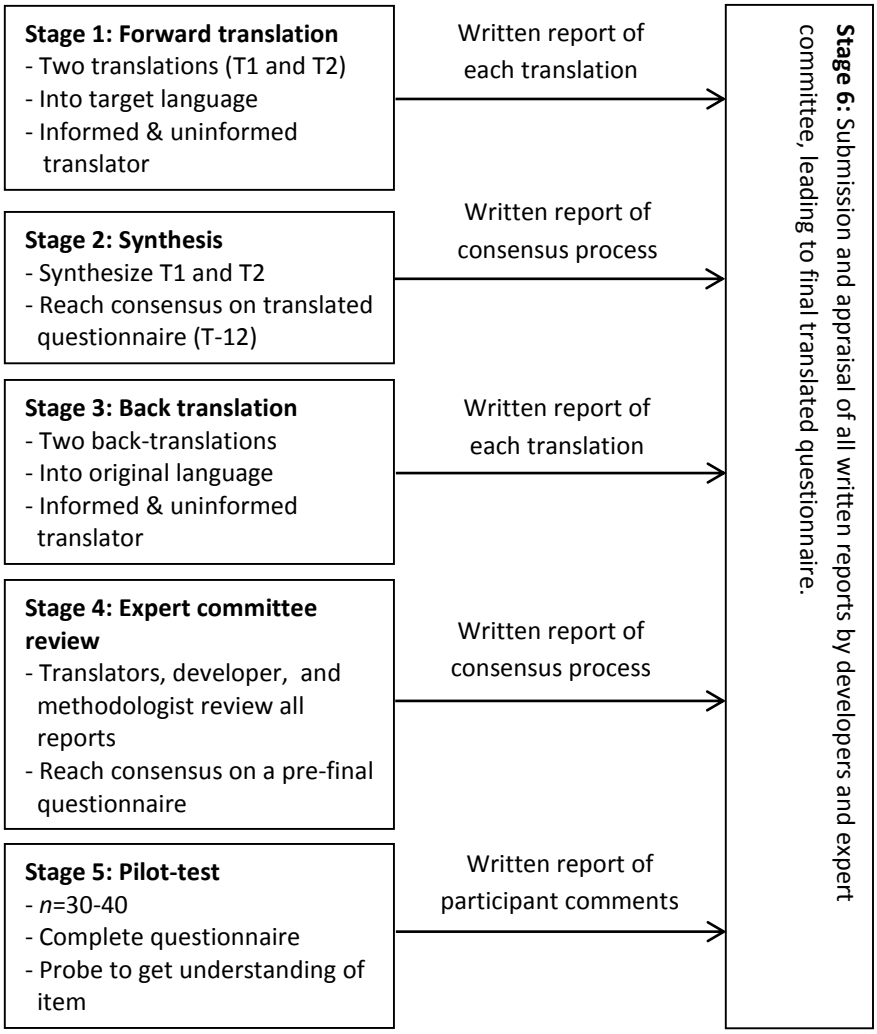


Figure 1. Recommended stages of the cross-cultural adaptation process, based on Beaton et al. [15].

Forward translation

The forward translation of the IWPQ's instruction, items, and answer categories, was performed by two independent translators. Both translators were bilingual, with American-English as their mother tongue. One translator had expertise on individual work performance, and the other translator was naive about the topic. Both translators wrote a report of the translation, containing challenging phrases and uncertainties, and considerations for their decisions.

Synthesis

The results of both translations (T1 and T2) were compared by the two translators and one researcher (LK). A written report documented the consensus process, the discrepancies, and how the discrepancies were resolved. The translators and the researcher reached consensus on one common American-English questionnaire (T-12).

Back translation

The common American-English questionnaire was back-translated into Dutch by two other independent translators. Both translators were bilingual, with Dutch as their mother tongue. One translator was naive about the topic, whereas the other translator had expertise on the topic. Both translators wrote a report of the translation, containing challenging phrases and uncertainties, and considerations for their decisions.

Expert committee review

All the translated versions were combined into one pre-final questionnaire by an expert committee. The expert committee consisted of the four translators, one researcher (LK), and one methodologist (HdV). Discrepancies between the original and translated versions were identified and discussed. Also, semantic, idiomatic, experiential and conceptual equivalences were evaluated. Again, a written report documented the consensus process, the discrepancies, and how the discrepancies were resolved. The expert committee reached consensus on a pre-final American-English version of the IWPQ.

Pilot-test

To examine the comprehensibility, applicability, and completeness of the translated questionnaire, a pilot-test was performed. A total of 40 participants were included

in the pilot-test. Inclusion criteria were: currently working (8 hours a week or more), aged 18-65 years, and able to read and understand the American-English language. Participants were recruited among employees of Tufts Medical Center in Boston, MA. In order to promote participation in the pilot-test, an outreach e-mail was send to employees of participating departments, after which an appointment with the researcher (LK) could be made. The pilot-test was approved by the Institutional Review Board of Tufts Medical Center (IRB number 10929).

After signing an informed consent file, participants filled in the American-English IWPQ. "Think aloud" and "probing" techniques [16] were used in order to identify participants' opinion on the comprehensibility, applicability, and completeness of the instructions, items, and answer categories of the translated questionnaire. The duration of the pilot-test was on average 15 minutes, including questionnaire completion. Participants' comments were written down into a report by the researcher (LK). The comments were independently assessed by two researchers (LK and CB), after which a consensus meeting took place. Any discrepancies that remained were discussed with the translators and the other IWPQ developers (VH, HdV, and AvdB), after which consensus was reached on a final American-English questionnaire.

Measurement properties of the pre-final questionnaire

Descriptive statistics of the IWPQ items and scales, and of the socio-demographic characteristics of the participants (gender, age, number of work hours a week, and primary type of occupation) were used to examine the distribution of the data. Internal consistency of the IWPQ scales was determined using Cronbach's alpha [17]. Item-to-scale correlations were calculated to evaluate the fit of the item within the scale. Furthermore, scale scores were examined for floor or ceiling effects (> 15% at the extreme values [16]). Statistical analyses of the data were done in SPSS20.

The content validity of the American-English questionnaire was evaluated by the members of the expert committee throughout the cross-cultural adaptation process, and by the developers of the IWPQ through qualitative analysis of the comments provided by the participants of the pilot-test.

Results

Cross-cultural adaptation

Translation

The forward translation of the IWPQ was conducted and some challenging issues were encountered. All issues were discussed among the two translators and the researcher, until consensus emerged. First, conceptual issues were identified with the instruction. “Behavior at work” was considered too evaluative, and might imply whether or not you obeyed rules like a child. To obtain conceptual equivalence to the original meaning, it was chosen to use “how you conducted yourself at work.” Second, for some questionnaire items, inconsistencies with the verb tense were identified. In Dutch, the simple past (e.g., “started”) and the present perfect (e.g., “have started”) are used interchangeably. It was chosen to consistently use the simple past in the American-English version, because the items refer to a completed action in the past 3 months. Furthermore, there were some idiomatic issues in the translation of items TP1, 3 and 4 (“I was able to” versus “I succeeded in”), TP2 (“keep in mind” versus “keep in sight”), CP8 and 9 (“keep up-to-date” versus “maintain”), CWB14 (“issues” versus “things”) and CWB16-18 (“aspects” versus “sides”). The main reasons for choosing the first option for each item, were similarity to the original Dutch item, generic applicability (suitable for workers in all types of jobs), and appropriateness of a word (decent, proper).

The back-translation was conducted without major difficulties. Issues were discussed among the members of the expert committee, until consensus emerged. First, a conceptual issue was identified with the instruction sentence “how you conducted yourself at work.” Comment was that you cannot “conduct yourself.” To obtain conceptual equivalence to the original meaning, it was chosen to use “how you carried out your work.” Second, there were some linguistic and conceptual issues in the wording of items TP2 (“results I needed to achieve in my work” was considered incorrect use of American-English), TP3 (“distinguish between” was considered double use of words), TP4 (“perform my work” versus “conduct my work”), and CP6 (“I started new tasks on my own initiative” was considered double use of words). These four items were adapted to correct use of American-English, and to obtain conceptual equivalence to the original meaning. Lastly, one translator expressed issues with the answer category “seldom,” and wondered whether this should be “seldomly” or “rarely.” However, consensus emerged that “seldom” is often used is answer scales of American-English questionnaires, and is correct use of American-English. Table 1 shows the pre-final American-English IWPQ.

Table 1. The pre-final American-English IWPQ and descriptive statistics of the items

Instructions: The following questions relate to how you carried out your work during the past 3 months. In order to get an accurate picture of your conduct at work, it is important that you complete the questionnaire as carefully and honestly as possible. If you are uncertain about how to answer a particular question, please give the best possible answer. The questionnaire will take about 5 minutes to complete. The questionnaire is completely anonymous: your answers will not be seen by your supervisor(s) or colleagues.

Task performance (TP) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes)	2 (regularly)	3 (often)	4 (always)		
In the past 3 months...								
TP1	I was able to plan my work so that I finished it on time. *	0 (0)	7 (17.5)	9 (22.5)	13 (32.5)	11 (27.5)	2.70 (1.07)	0.64
TP2	I kept in mind the work result I needed to achieve. *	0 (0)	1 (2.5)	7 (17.5)	12 (30.0)	20 (50.0)	3.28 (0.84)	0.58

Table 1. Continued

Task performance (TP) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes	2 (regularly)	3 (often)	4 (always)		
In the past 3 months…								
TP3	I was able to distinguish main issues from side issues. * §	1 (2.5)	10 (25.0)	15 (37.5)	10 (25.0)	4 (10.0)	3.08 (0.76)	0.46
TP4	I was able to carry out my work well with minimal time and effort. * §	0 (0)	1 (2.5)	7 (17.5)	20 (50.0)	12 (30.0)	2.15 (1.00)	0.55
TP5	I planned my work optimally. §	0 (0)	5 (12.5)	10 (25.0)	16 (40.0)	9 (22.5)	2.73 (0.96)	0.60

Table 1. Continued

Contextual performance (CP) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes)	2 (regularly)	3 (often)	4 (always)		
In the past 3 months...								
CP6	On my own initiative, I started new tasks when my old tasks were completed.	0 (0)	5 (12.5)	8 (20.0)	10 (25.0)	17 (42.5)	2.98 (1.07)	0.24
CP7	I took on challenging tasks when these were available. §	1 (2.5)	2 (5.0)	5 (12.5)	17 (42.5)	15 (37.5)	3.08 (0.97)	0.74
CP8	I worked on keeping my job-related knowledge up-to-date. *	1 (2.5)	2 (5.0)	11 (27.5)	14 (35.0)	12 (30.0)	2.85 (1.00)	0.59
CP9	I worked on keeping my work skills up-to-date. *	0 (0)	3 (7.5)	13 (32.5)	11 (27.5)	13 (32.5)	2.85 (0.98)	0.62

Table 1. Continued

Contextual performance (CP) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes)	2 (regularly)	3 (often)	4 (always)		
In the past 3 months...								
CP10	I came up with creative solutions for new problems.	1 (2.5)	3 (7.5)	11 (27.5)	18 (45.0)	7 (17.5)	2.68 (0.94)	0.52
CP11	I took on extra responsibilities.	1 (2.5)	2 (5.0)	6 (15.0)	23 (57.5)	8 (20.0)	2.88 (0.88)	0.60
CP12	I continually sought new challenges in my work.	0 (0)	4 (10.0)	10 (25.0)	15 (37.5)	11 (27.5)	2.83 (0.96)	0.70
CP13	I actively participated in meetings and/or consultations.	0 (0)	4 (10.0)	5 (12.5)	15 (37.5)	16 (40.0)	3.08 (0.97)	0.43

Table 1. Continued

Counterproductive work behavior (CWB) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes)	2 (regularly)	3 (often)	4 (always)		
In the past 3 months...								
CWB14	I complained about unimportant issues at work. * §	4 (10.0)	20 (50.0)	13 (32.5)	3 (7.5)	0 (0)	1.38 (0.77)	0.70
CWB15	I made problems at work bigger than they were.	15 (37.5)	20 (50.0)	4 (10.0)	1 (2.5)	0 (0)	0.78 (0.73)	0.73
CWB16	I focused on the negative aspects of a situation at work instead of the positive aspects. *	10 (25.0)	19 (47.5)	10 (25.0)	1 (2.5)	0 (0)	1.05 (0.78)	0.72

Table 1. Continued

Counterproductive work behavior (CWB) scale								
		Response, n (%)					Mean (SD)	Item-to-scale correlation
		0 (seldom)	1 (sometimes)	2 (regularly)	3 (often)	4 (always)		
In the past 3 months...								
CWB17	I talked to colleagues about the negative aspects of my work. *	8 (20.0)	18 (45.0)	8 (20.0)	5 (12.5)	1 (2.5)	1.33 (1.02)	0.81
CWB18	I talked to people outside of the organization about the negative aspects of my work. *	10 (25.0)	18 (45.0)	7 (17.5)	4 (10.0)	1 (2.5)	1.20 (1.02)	0.76

*Difficult to translate; §Adjusted after pilot-test.

Pilot-test

The pre-final version of the questionnaire was administered to 40 employees of Tufts Medical Center (n=18 men and n=22 women). On average, participants were 34.5 (9.8) years of age, and worked 45.9 (13.7) hours a week. See Table 2 for an overview of the sample descriptives.

Table 2. Descriptive statistics of the pilot-test sample

	Total (n=40)	**Men (n=18)**	**Women (n=22)**
Age in years, mean (SD)	34.5 (9.8)	35.2 (8.9)	33.8 (10.6)
Working hours/week, mean (SD)	45.9 (13.7)	50.8 (13.6)	41.8 (12.7)
Primary occupation, n (%)			
Research position	19 (47.5)	9 (50.0)	10 (45.5)
ICT position	1 (2.5)	1 (5.6)	0 (0)
Education position	2 (5.0)	0 (0)	2 (9.1)
Clinical position	12 (30.0)	7 (38.9)	5 (22.7)
Administrative position	6 (15.0)	1 (5.6)	5 (22.7)
Highest completed education level, n (%)			
High school degree	3 (7.5)	0 (0)	3 (13.6)
College degree (e.g., Bachelor)	12 (30.0)	4 (22.2)	8 (36.4)
Master degree	9 (22.5)	3 (16.7)	6 (27.3)
MD degree	15 (37.5)	10 (55.6)	5 (22.7)
PhD degree	1 (2.5)	1 (5.6)	0 (0)

Five participants (12.5%) mentioned that the **instructions** were not clear in terms of wording. Two of these participants were unsure what “how you carried out your work” meant, and three of these participants thought “conduct at work” was vague. One participant felt that the use of “conduct at work” had a negative interpretation, while two other participants said that it was not negative per se, but it were just words they would not normally use. During the translation process, conceptual issues had also arisen with these words. After deliberate discussion, the translators reached consensus on “how you carried out your work” as closest to the original meaning. As no better alternative was suggested during the pilot-test, and only a minority of participants reported an issue, it was chosen not to change the instructions.

Ten participants (25%) felt that the distinctions between the **answer categories** were unclear. This mainly concerned the distinction between "regularly" and "often," with eight participants feeling that these categories are almost the same, and could also be placed the other way around. In addition, two participants felt that "seldom" and 'sometimes" were almost the same. One participant felt that "seldom" should be worded as "rarely." Finally, two participants wondered whether everyone would notice the change in answer categories for the CWB scale. Some participants suggested to rename the answer categories to "none of the time – some of the time – half of the time – most of the time – all of the time," or to only name the extreme categories and number the middle categories. Another participant said that no matter how the answer categories are labeled, people will always have trouble distinguishing them, and they will be filled in like a VAS scale. As no clear alternative arose during the pilot-test, and only a minority of participants reported an issue, it was chosen not to change the answer categories in order to retain equivalence to the Dutch version.

Although participants stated that they had no major difficulties in understanding or answering most of the **items**, six items stood out during the pilot-test. Most comments were made on the items in the task performance scale. Twelve participants (30%) were unsure what was meant by "work result" in question **TP2** ("I kept in mind the work result I needed to achieve"). They made suggestions to change "work result" to for example "work goal," "deadlines," or "work outcome." However, as there was no consensus on an alternative amongst the participants, nor amongst the expert committee, the question was not changed. Thirteen participants (32.5%) thought that question **TP3** ("I was able to distinguish main issues from side issues") was oddly phrased. Most of the participants said they would never use the words "main issues and side issues," and were unsure what they meant. Most of the participants suggested replacing these words with "prioritize." Therefore, the question was changed to "I was able to set priorities." Seventeen participants (42.5%) felt that question **TP4** ("I was able to carry out my work well with minimal time and effort") had a negative interpretation. Participants felt you cannot carry out your work well with minimal time and effort. This means that you are lazy, and you take shortcuts. You need to put in time and effort to do your work well. Most participants suggested changing the question to whether you were able carry out your work "efficiently." Therefore, the question was changed to "I was able to carry out my work efficiently." Eleven participants (27.5%) indicated that question **TP5** ("I planned my work optimally") was strangely phrased, although they understood what

was meant. Suggestions for rephrasing this question were diverse, including "I scheduled my work optimally," "I planned my work efficiently," "I managed my time well," "I could get my work done in the best way possible," and "with planning I was able to complete all my work tasks." The translators and authors of the IWPQ agreed that the question should be changed to "I managed my time well." In question **CP7** ("I took on challenging work tasks when these were available"), the word "these" was replaced by "they", based on suggestion from three participants and the translators. Eleven participants (27.5%) commented on question **CWB14** ("I complained about unimportant issues at work"). Seven of these participants tripped over the word "unimportant," and argued that if they complained about it, that meant the issue was not unimportant to them. Four of these participants wanted more specificity as to who and where they should have complained (e.g., to colleagues or to friends, at work or at home). Therefore, it was decided to change the question to "I complained about minor work-related issues at work." In conclusion, a total of five items were changed based on the pilot-test results (marked in Table 1).

Almost all participants (85%) felt that all questions were applicable to their job. Two participants said that question TP5 (planning work optimally) was less relevant to their job, because as doctors, they had little influence on how many patients came in during the day, and which problems were presented. Two participants said CP8 (keeping job-related knowledge up-to-date) was less relevant to their job. Also, one participant said CP9 (keeping work skills up-to-date) was less relevant to their job. Furthermore, five participants reported reservations to answer the CWB questions honestly, because they felt the questions were a bit uncomfortable or intense to answer. Two participants said that the CWB questions were less relevant to them, one because solving problems (negative aspects) was a part of her work, and the other because she was not supposed to complain at work. Based on these few comments, it was not considered necessary to remove any questions from the questionnaire. Due to the generality of the questionnaire, it was considered inevitable that some questions are less relevant to some participants than others.

All participants (100%) stated that the completeness of the questionnaire was good. When asked, 16 participants (40%) had suggestions to expand the questionnaire to include all relevant aspects of their work performance. These suggestions mainly included determinants of individual work performance (e.g., job satisfaction, job tenure, and sleep quality), or indicators of individual work

performance that were previously included, but removed during the development of the questionnaire (e.g., relationship with co-workers and supervisor(s), collaboration with others, access to and use of supplies). Based on the suggestions, it was not considered necessary to add any new questions to the questionnaire. A short questionnaire with content identical to the Dutch version was considered most important.

Measurement properties of the pre-final questionnaire

Descriptive statistics of the IWPQ items can be seen in Table 1, and descriptive statistics of the IWPQ scales can be seen in Table 3. Almost all items showed floor or ceiling effects (>15% at the lowest or highest answer category). At the scale level, the mean score for task performance was 2.79 (SD=0.69), 2.90 (SD=0.65) for contextual performance, and 1.15 (SD=0.73) for counterproductive work behavior. The mean scale scores are comparable to scores in The Netherlands, although the mean scale score for contextual performance was slightly higher than in The Netherlands (2.90 in the USA, versus 2.31 in The Netherlands) [9]. There were no ceiling or floor effects on the scale level. Five percent of the participants showed the highest score (4, “always”) for the task performance scale, and the contextual performance scale. Five percent also showed the lowest score (0, “never”) for the counterproductive work behavior scale.

The Cronbach’s alpha for the task performance, contextual performance, and counterproductive work behavior scales were 0.79, 0.83 and 0.89, respectively (Table 3). The item-to-scale correlations were sufficiently high (r>0.40), except for item CP6 (“On my own initiative, I started new tasks when my old tasks were completed”), which correlated r=0.24 on the contextual performance scale (see Table 1).

Based on the cultural adaptation process, and the comments provided by the participants of the pilot-test, the content validity of the American-English IWPQ was judged to be good. Almost all participants in the pilot-test considered the questions to be applicable and relevant to their job, and all participants felt that the completeness of the questionnaire was good.

Table 3. Descriptive statistics of the pre-final American-English IWPQ scales

	Range (0 - 4)	Mean (SD)	Median	% floor-effects (score 0)	% ceiling-effects (score 4)	Cron-bach's α
Task performance	1.40 - 4	2.79 (0.69)	2.90	0	5	0.79
Contextual performance	1.50 - 4	2.90 (0.65)	2.88	0	5	0.83
Counter-productive work behavior	0 - 3.20	1.15 (0.73)	1.10	5	0	0.89

Discussion

The goal of the current study was to cross-culturally adapt the Individual Work Performance Questionnaire (IWPQ) from the Dutch to the American-English language, and to assess the questionnaire's internal consistency and content validity in the American-English context. The cross-cultural adaptation was systematically performed, resulting in an American-English version of the IWPQ that equals the original version. In general, participants were positive on the comprehensibility of the questionnaire. A few changes were made to optimize the comprehensibility of the questionnaire. Here, the consideration of not changing the wording of a question in order to keep it similar to the original question, versus changing the wording of a question in order to obtain conceptual equivalence to the original question, is important. For example, the answer category labels of the IWPQ were not changed, in order to retain equivalence to the Dutch version, and because no alternative arose that was believed to improve comprehensibility. On the contrary, the wording of task performance items 3 ("*I was able to distinguish main issues from side issues*") and 4 ("*I was able to carry out my work well with minimal time and effort*") was changed in order to improve comprehensibility. In Dutch, it was chosen to give a description of "prioritizing" and "efficiently," as these words are hardly ever used directly. However, based on American participants' suggestions to improve comprehensibility, these items were shortened to more directly ask for "prioritizing" and "working efficiently."

All participants were positive on the completeness of the questionnaire, and almost all participants indicated that all the questions were relevant and

applicable to them. This indicates good content validity of the questionnaire. Thus, there appear to be no cultural differences between The Netherlands and America in measuring the concept of individual work performance, and the indicators used to measure the concept of individual work performance seem to be equivalent over these contexts. Although additional indicators of individual work performance suggested by participants in the pilot-test (e.g., relationship with co-workers and supervisor(s), collaboration with others, access to and use of supplies) might have been included when developing the IWPQ from scratch in America, a short questionnaire with identical content to the Dutch IWPQ was considered most important in the current study. The generalizability of the questions in the Dutch IWPQ was probably promoted by the fact that people from multiple countries (including the USA) were involved in the developmental stages of the IWPQ, for example, during item generation [18].

If the IWPQ items have kept the same meaning after the translation, the American-English questionnaire is expected to retain the same factor structure as in The Netherlands. The sample size in the current pilot-test (n=40) was too small to conduct a confirmatory factor analysis. De Vet et al. [16] recommend a sample size of at least n=100 to perform a reliable factor analysis. However, item-to-scale correlations were examined, and were similar to the item-to-scale correlations in The Netherlands. All items loaded sufficiently high on the expected scales, except for item CP6 ("*On my own initiative, I started new tasks when my old tasks were completed*"). The low loading of this item on the contextual performance scale suggests that this item has a different meaning in the USA than in The Netherlands, either due to the translation, or due to cultural differences. However, no specific comments were made regarding this question during the pilot-testing phase, so the reason for the low loading is unclear. Future research should administer the American-English IWPQ in a larger sample (at least n=100), so that its factor structure can be examined, and the loading of the items on each scale can be examined in more detail.

The measurement properties of the Dutch and American-English IWPQ appear to be similar. The mean item and scale scores appear to be similar in both versions, although the mean scale score for contextual performance was slightly higher for the American-English than Dutch IWPQ. Nevertheless, there were no considerable ceiling or floor effects at the scale level. The internal consistencies of the American-English IWPQ task performance, contextual performance, and counterproductive work behavior scales were 0.79, 0.83, and 0.89, respectively. This

is similar to the Dutch version, where the scale reliabilities are 0.78, 0.85, and 0.79, respectively. The internal consistency of the American-English CWB scale is higher than in The Netherlands.

Limitations

A limitation of the current study was that participants were aware that the questionnaire measured individual work performance, due to the informed consent procedure before the study. Secondly, in the current study, a researcher was sitting next to the participants while they were filling in the questionnaire. Finally, some participants reported reservations to answer the CWB questions honestly, because they felt the questions were a bit uncomfortable or intense to answer. All these factors may have elicited socially desirable answers, and resulted in different scores on the American-English version than the Dutch version of the questionnaire. In general, we recommend leaving out the questionnaire title and scale names when administering the questionnaire, so that participants are less aware they are filling in a questionnaire on individual work performance. We also recommend that participants' answer are always anonymous and are treated confidentially. It should be guaranteed that only group level outcomes will be reported to managers or companies, obtained in large enough groups, so that results can never be traced back to individual participants.

The pilot-test in the current study was conducted in a relatively high-educated sample, with participants primarily working in a pink or white collar job. This may limit generalizability of the results to lower-educated workers, and blue collar workers. Although, in general, the translators were positive on the questionnaire's comprehensibility, applicability, and completeness for lower-educated workers, and blue collar workers, one translator had concerns about the use of the word "priorities" in these groups. Ideally, the comprehensibility, applicability, and completeness of the American-English IWPQ, as well as its internal consistency and content validity, should still be examined in these groups.

Future research

Although the pilot-test results indicate good internal consistency and content validity of the American-English IWPQ, it is only after the cross-cultural translation and adaptation that the real cross-cultural validation takes place [16]. In a larger and more heterogeneous sample, special attention should be paid to the measurement invariance of the questionnaire in the original and the new target population.

Measurement invariance means that a measurement instrument, a scale, or an item, functions in exactly the same way in different populations [16]. This can be examined, for example, using factor analysis or item response theory (IRT) techniques. Future research should perform confirmatory factor analysis in a larger and more heterogeneous sample, and examine if (and if so, why) item CP6 loads insufficiently high on its original dimension. IRT techniques are also a powerful method with which to detect differential item functioning (DIF), by comparing the item characteristic curves of the items in the original version and the translated version [16]. This can give insight into whether the difficulty of an item has changed in the original and translated version. Future research should also further examine the reliability, construct validity, and responsiveness of the American-English IWPQ.

Conclusion

The cross-cultural translation and adaptation of the IWPQ from the Dutch to the American-English language was conducted without major difficulties. The comprehensibility, applicability, and completeness of the translated version of the IWPQ was considered to be good. Also, its internal consistency and content validity appeared to be good. The translated questionnaire can now be used to measure, for example, the effectiveness of workplace interventions on individual work performance in an American-English speaking context. Future research should further examine the measurement invariance, reliability, validity, and responsiveness of the American-English IWPQ in a larger and more heterogeneous sample.

Thank you

We wish to thank Kimi Uegaki, Tammy Rubinstein, Fenna Leijten and Nico Pronk for their help in translating the questionnaire, the EMGO+ Institute for Health and Care Research for providing the travel grant, Tufts Medical Center for their hospitality, and the Tufts Medical Center employees for their participation in the pilot-test.

References

1. European Commission. EUROPE 2020: A strategy for smart, sustainable and inclusive growth. 2010. Available from: http://ec.europa.eu/europe2020/.
2. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, editors. Handbook of industrial and organizational psychology, Vol.1 (2nd ed.). Palo Alto, CA, US: Consulting Psychologists Press; 1990. p. 687-732.
3. Austin JT, Villanova P. The criterion problem: 1917-1992. Journal of Applied Psychology 1992;77(6):836-874.
4. Koopmans L, Bernaards CB, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual frameworks of individual work performance: a systematic review. Journal of Occupational and Environmental Medicine 2011;53(8):856-866.
5. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: a policy-capturing approach. J Appl Psychol 2002;87(1):66-80.
6. Viswesvaran C, Ones DS. Perspectives on Models of Job Performance. International Journal of Selection and Assessment 2000;8(4):216-226.
7. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, editors. Personnel Selection in Organizations San Francisco, CA: Jossey Bass; 1993. p. 71-98.
8. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Development of an Individual Work Performance Questionnaire. International Journal of Productivity and Performance Management 2013;62(1):6-28.
9. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Improving the Individual Work Performance Questionnaire using Rasch Analysis. Journal of Applied Measurement 2014;15(2).
10. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J Appl Psychol 2005;90:1241-1255.
11. Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. Journal of Management 1991;17(3):601-617.

12. Podsakoff PM, MacKenzie SB. A second generation measure of organizational citizenship behavior. Indiana University, Bloomington; 1989.
13. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). Journal of Occupational and Environmental Medicine 2003;45:156-174.
14. Lerner D, Amick BC, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Medical Care 2001;39(1):72-85.
15. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25(24):3186-3191.
16. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. Cambridge University Press; 2011.
17. Cronbach IJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-333.
18. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, van der Beek AJ. Measuring Individual Work Performance: Identifying and Selecting Indicators. Work: Journal of Prevention, Assessment & Rehabilitation 2013;45(3).

Chapter 9

General Discussion

9

Objective of this thesis

The objective of this thesis was to develop and validate a comprehensive, generic, and short questionnaire to measure individual work performance (IWP). This objective was driven by the need to develop and evaluate interventions, procedures and strategies that can maintain, improve, or optimize IWP. An optimal IWP is crucial these days, due to increasing globalization of the economy and corresponding competitiveness between companies all over the world, the economic recession in many countries, and the growing need for sustainable employability. However, an existing lack of consensus in research and practice on how to define, conceptualize and measure IWP impeded valid measurement of the construct of IWP. In this final chapter, reflections on the main results and implications of this thesis are presented.

Overview of the main results

The first step towards a comprehensive, generic, and short measure of IWP was establishing a clear definition and conceptualization of IWP. In Chapter 2, conceptual frameworks of IWP were systematically reviewed from the occupational health, work and organizational psychology, and management and economics literature. The definition of IWP as "behaviors or actions that are relevant to the goals of the organization" [1] was chosen, because it is a broad definition of IWP, and thus, it is usable in multiple research fields, and applicable to a generic working population. Next, the underlying structure of IWP was determined. Numerous conceptual framework of IWP were integrated into one conceptual framework consisting of four broad and generic dimensions. The first dimension, task performance, can be defined as "the proficiency with which individuals perform central job tasks" [1]. The second dimension, contextual performance, can be defined as "behaviors that support the organizational, social and psychological environment in which the technical core must function" [2]. The third dimension, counterproductive work behavior, can be defined as "behavior that harms the well-being of the organization" [3]. Initially, adaptive performance was included as a fourth dimension, and was defined as "an employee's ability to adapt to changes in a work system or work roles" [4]. Later on, however, the adaptive performance dimension was merged with the contextual performance dimension, leading to a conceptual framework consisting of three dimensions.

After the establishment of a clear definition and conceptualization of IWP, none of the existing questionnaires identified in the literature to measure (aspects of) IWP seemed adequate. For example, they did not capture the complete range of

individual work performance, did not include relevant content, or were not generically applicable. Therefore, it was decided to develop a new instrument that could overcome the limitations of existing instruments. In order to develop the Individual Work Performance Questionnaire (IWPQ), first, the operationalization of the IWPQ scales (task performance, contextual performance, and counterproductive work behavior) warranted attention. Indicators that could be used to measure the dimensions were identified from the scientific literature, existing questionnaires, and expert interviews (Chapter 3). From these, the most relevant indicators per dimension were selected by experts from different professional backgrounds. After a pilot-test, the IWPQ was subjected to a field-test in a large sample of blue, pink, and white collar workers (Chapter 4). After this, another improvement round was held in order to improve the targeting of the IWPQ. The questionnaire was tested again in a large sample of blue, pink, and white collar workers, and the final version of the questionnaire – the IWPQ 1.0 – was established (Chapter 5). The internal consistency and construct validity of the IWPQ 1.0 were good (Chapter 6). Conclusions about the responsiveness of the IWPQ cannot yet be drawn, and more research on this characteristic is necessary (Chapter 7). Also, the IWPQ 1.0 was cross-culturally adapted to the American-English language (Chapter 8). The main benefits of the IWPQ are that it measures all relevant dimensions of IWP, it is generically applicable, and it is short.

Methodological issue – generic applicability

A questionnaire to establish the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP, should be generically applicable, because such interventions often take place in varied settings. Thus, the same questionnaire needs to be suitable for different companies (e.g., so that results of an intervention can be compared across companies), and for a company with different types of employees (e.g., cleaners, office workers, managers). However, at the start of this thesis, it remained to be seen whether it was possible to develop such a questionnaire. Generic applicability of the questionnaire developed in this thesis was stimulated by conceptualizing and operationalizing IWP in a multi-disciplinary way, developing and evaluating the questionnaire using Rasch Analysis [5], and testing the questionnaire in a broad sample of blue, pink, and white collar workers.

In its current form, the IWPQ is generically applicable to workers from blue, pink, and white collar sectors. Although the responsiveness of the IWPQ deserves further attention, it is probable that the IWPQ is suitable for longitudinal

comparisons between the same group over time. This characteristic is most important, because the main purpose of the IWPQ is to establish the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP. In its current form, the IWPQ is less suitable for cross-sectional comparisons between different groups (e.g., comparing carpenters and dentists on IWP). Different cut-off points should be used when interpreting scores for workers from different occupational sectors, because the IWPQ includes a few items with Differential Item Functioning (DIF; see Chapters 4 and 5). Items with DIF were allowed in the IWPQ, because in order to detect changes in IWP over time, adequate targeting is more important than DIF-free items. Thus, the IWPQ is generically applicable, but we should be cautious in comparing different occupational groups on IWP.

The IWPQ scales were developed and validated in three broad occupational sectors, namely blue, pink, and white collar workers. As a consequence, we could not examine the reliability and validity of the IWPQ in every specific occupation. Although generic questionnaires pose considerable advantages in research, when examining IWP in a specific job, a job-specific questionnaire may be preferred to capture all the relevant aspects of that job. When such a questionnaire is not available, job-specific questions could be added to the IWPQ. An advantage of doing this is that all relevant dimensions of IWP are included (which often are not in existing questionnaires, as examined in Chapter 3).

Methodological issue – use of self-report

The IWPQ was developed as a self-report questionnaire. Self-reports can be accompanied by several biases, such as *recall bias, social desirability bias, leniency effects*. Due to recall bias, systematic error may be introduced in the answers by inaccuracy or incompleteness of people's recollections of their past behaviors at work [6]. Also, a questionnaire on IWP may elicit socially desirable answers. That is, people tend to respond to an item in a certain way, because they think it is the socially acceptable answer, rather than their true answer [7]. A closely related bias is the leniency effect, that is, people are naturally motivated to present themselves in a favorable, positive light [6]. As a result of social desirability and leniency effects, self-ratings of performance may lead to a higher score than in reality. For example, Van der Heijden and Nijhof [6] found that self-ratings of performance are generally one half to one standard deviation higher than ratings by peers or managers.

In the IWPQ, several precautions were taken to minimize influences of self-report biases. For example, effects of social desirability and leniency were minimized

by leaving out the questionnaire title and scale names when administering the questionnaire, so that negative connotations related to IWP are avoided, and by reassuring participants that their answers are anonymous and treated confidentially. In addition, the center of the answer category scale was shifted, so that the average point was not in the middle but rather to the left of the scale for task and contextual performance, and to the right of the scale for counterproductive work behavior. Doing this will prevent possible floor and ceiling effects that may result from social desirability and leniency biases [8].

Despite possible biases accompanying self-report, the IWPQ was developed as a self-report questionnaire for several reasons. First, in many occupations, objective measures of performance are not easily obtainable [9]. Especially for knowledge work or high complexity jobs, direct measures of countable behaviors or outcomes, such as production quantity or number of errors made, are almost impossible. Second, self-reports have practical advantages such as ease of collection, issues of confidentiality, and less problems with missing data, when compared to objective measures, peer or managerial ratings [10]. Finally, peer or managerial ratings of performance can also be accompanied by several biases, such as *under-sampling bias, halo effects* and *leniency effects*. In comparison to self-ratings, ratings by supervisors are based on a much smaller amount of information, leading to the so-called effect of *under-sampling*. The person who is doing the job possesses the greatest familiarity with the job and their own behavior at work, and because of that, is an appropriate person to fill in the questionnaire [6]. The notion that employees have more opportunity to observe their own behaviors than peers or managers do may be especially true for counterproductive behaviors, because most of these behaviors are intended to be private and, hence, unobservable [11]. A recent study found that self-raters actually reported engaging in more counterproductive behaviors than other-raters reported them engaging in [12]. Self-report methodology is consistently used throughout CWB research, and given limitations in other methods, several studies have concluded that it is the most appropriate method [e.g., 13, 14]. Another bias accompanying peer or supervisor ratings of performance is the *halo effect*, in which the peer's or supervisor's general impression of the employee (for example, liking or disliking the employee) influences the evaluation. As a result, peers and supervisors score the different dimensions of IWP are more similar (i.e., they answer more consistently), and inter-correlations between the dimensions are overestimated [6, 11, 15]. In addition, not only self-ratings, but also peer and supervisor ratings are influenced by *leniency effects*.

Although research has shown that this effect is stronger for self-ratings, it is also present in peer and supervisor ratings [16, 17]. Thus, peers and supervisors tend to see others in a favorable, positive light [6].

Methodological issue – lack of a golden standard

The criterion validity of the IWPQ scales could not be examined, because there was no golden standard available. This makes it difficult to prove that a new measure is indeed measuring what is intended [8]. A perfect golden standard may be an objective measure of IWP, but in many occupations these are hard to obtain [9]. Especially for knowledge work or high complexity jobs, direct measures of countable behaviors or outcomes, such as production quantity or number of errors made, are almost impossible. Other possible golden standards may be peer or managerial ratings. However, peer or supervisor ratings are accompanied by several biases, as described above, and therefore, cannot be considered golden standards of IWP. The solution lies in examining the construct validity of the questionnaire, as was done in the current thesis. Although construct validity is often considered to be less powerful than criterion validation, with strong theories and specific and challenging expectations, it is possible to acquire substantial evidence that the measurement instrument is measuring what it purports to measure [18].

Strengths

IWP is an issue that is researched in multiple research fields, and its conceptualization and operationalization must ideally be considered in a multi-disciplinary way. A strength of the current thesis is that multiple research fields and stakeholders were involved throughout the development process of the IWPQ. First, literature from multiple research fields was used in order to establish a definition and conceptual framework of IWP. Second, the literature, existing questionnaires, and experts from multiple research fields were consulted in order to construct the questionnaire. Third, workers from different occupational sectors were included in the field-testing phase, resulting in a generically applicable questionnaire.

A second strength of the current thesis was that a thorough development and improvement process was applied before the final IWPQ was reached. Often, a major problem with many studies is that insufficient time is allowed for proper field-testing, further adaptation and re-evaluation before the final instrument is used in research and/or practice [18, 19]. As described in more detail in the paragraphs above, in the current thesis, a clear definition and conceptual framework of IWP

were established before developing the questionnaire, consensus was reached on the operationalization of the scales, and a pilot-test, field-test, and improvement round were held. Subsequently, the internal consistency and validity of the final IWPQ scales were tested. This has resulted in a reliable and valid questionnaire that measures IWP in a short, comprehensive and generic way, and that is ready to be used in research and practice.

Another strength of this thesis is that it is the first to develop and evaluate an IWP questionnaire using Rasch analysis [5]. This offered unique insights into the IWPQ scale characteristics. Rasch analysis ensures that key measurement assumptions, such as appropriate category ordening, local independence, and differential item functioning, are tested. In addition, Rasch analysis has particular value in the development of new questionnaires, specifically in guiding item reduction [20]. Furthermore, it ensured that items were suitable for all occupational sectors, and gave insight into whether the items were well distributed over the whole range of the scale (targeting). When there is optimal targeting, one can reliably measure persons at all levels of ability, and discriminate between persons at various ranges on the scale (e.g., discriminate amongst workers with low performance, as well as amongst workers with high performance). When improving the targeting of the IWPQ, Rasch analysis guided the removal of misfitting items and the addition of new items that improved targeting. A final benefit of Rasch analysis is that it provides a statistically proven interval level scale, instead of an ordinal level scale that is formed by the raw scores [21]. This is useful when one wants to measure changes in IWP over time. In an ordinal scale, higher scores indicate higher performance, but the relative distances between the scores are meaningless [22]. For example, it is unclear whether a person whose performance increased from 2 to 3, has made the same amount of improvement as a person whose performance increased from 3 to 4. On the contrary, an interval scale allows straightforward interpretation of the distances between scores, so that a change in score from 2 to 3 is equivalent to a change from 3 to 4. Thus, a benefit of the IWPQ is that its change scores can be estimated and interpreted more accurately than other IWP questionnaires using Rasch analysis, because it has a statistically proven interval scale.

Limitations

A limitation of the present thesis is that the responsiveness of the IWPQ remains unknown. So far, the IWPQ was used in an intervention study that stimulated

physical activity and relaxation amongst office workers, in order to improve their need for recovery and work-related outcomes (such as IWP). In this intervention study, statistically significant changes in tasks performance and contextual performance over time were detected with the IWPQ [23]. However, no firm conclusions could be drawn about the responsiveness of the IWPQ based on the results of that study. Several reasons may have accounted for this, such as a study population with high baseline scores, small changes on many constructs in the study, and unclear guidelines for interpreting responsiveness. Nevertheless, there is good reason to believe that the IWPQ has good responsiveness, because Rasch analysis ensured that items with a high discrimination parameter were included in the IWPQ (see Chapters 4 and 5), the IWPQ showed very good discriminative validity (see Chapter 6), and significant changes in IWP over time were detected in an intervention study [23].

A second limitation is that the IWPQ task performance scale has ceiling effects, and the CWB scale has floor effects. Despite an improvement round to improve the targeting of the scales (see Chapter 5), workers still score relatively high on task performance, and relatively low on CWB. As a consequence, the IWPQ is less able to discriminate workers with high task performance, and less able to discriminate workers with low counterproductive performance. Especially in longitudinal studies, this could be problematic, because workers who at baseline score high on task performance, or low on CWB, cannot show any further improvement (thus, even more task performance, or even less CWB). As previous examination of the IWPQ using Rasch analysis showed that the items of the IWPQ are relatively well-distributed over the scales (see Chapter 5), the ceiling and floor effects are not likely to be a shortcoming of the measurement instrument, but rather, a true characteristic of the population (an actual high/low occurrence of these behaviors in the workplace). Furthermore, social desirability and leniency biases may drive workers to present themselves in a favorable light. When administering the IWPQ, it is therefore important to leave out the questionnaire title and scale names to avoid negative connotations related to IWP, and to guarantee participants' anonymity.

A final limitation is that the IWPQ is unsuitable for individual level use. For individual level use, a minimum reliability value of 0.90, and preferably 0.95, is required. For group level use, a minimum reliability value of 0.70 is required [18, 24, 25]. As the reliability of the IWPQ scales varies from 0.74 for the CWB scale to 0.85 for the contextual performance scale, the IWPQ can currently only be used for cross-

sectional comparisons between, and longitudinal comparisons within, groups. This means that the IWPQ should not be used for comparisons at the individual level, nor for assessments, evaluations, and selection processes at the individual level.

Recommendations for future research on the IWPQ

As evidenced by the current thesis, the development and validation of a questionnaire can easily take years, and even then, it is never quite done. Strictly speaking, one can never state that an instrument is valid, only that it provides valid scores in the specific situation in which it has been tested [18]. As construct validation is an ongoing process, more research should be conducted to create a strong web of evidence to support the validity of the IWPQ. The relationship of the IWPQ scales with other constructs, such as work engagement, job satisfaction, and health, should be examined in settings and populations other than was done in the current thesis. Also, despite the limitations of peer and supervisor ratings, and objective measures of performance, it would be interesting to compare these with self-ratings on the IWPQ scales for validation purposes.

In addition, more research is necessary to draw conclusions about the responsiveness of the IPWQ. Preferably, its responsiveness is examined in intervention studies that directly target IWP, in populations that have low baseline levels on the constructs under examination. Suggestions for such populations could be workers with work-related musculoskeletal health problems, mental health problems, and/or low job satisfaction. An intervention study, which is directly aimed at improving IWP, could obtain greater changes in these populations, making it easier to detect changes in IWP and related constructs. Suggestions for such a study could be an intervention focusing on managerial style, technological improvements at work, and/or job skills training.

For both research and practice, it is important that the scores on the IWPQ scales can be easily interpreted. In Chapter 5, an interpretation is given of *single scores* from “very low” to “very high” performance. However, it remains a question for future research how *change scores* should be interpreted. To do this, the smallest detectable change (SDC) and the minimally important change (MIC) should be determined. The SDC represents the change that can be detected beyond measurement error, and can be determined using the standard error of measurement. To obtain the standard measurement error of the IWPQ scales, test-retest reliability of the scales should be determined in a stable population over a short time interval (e.g., a couple of days). The MIC represents the smallest change

which individuals perceive as important, and should ideally be determined with someone who can assign practical meaning to a change, for example the employees themselves, their colleagues, or their supervisor(s). The MIC could, for example, be expressed in a number of points change, or a percentage change, from the baseline score on the IWPQ scales.

As of yet, the IWPQ is only suitable for group use. Future research could examine whether the IWPQ scales can be adapted for individual use. A strategy could be to add more questions to the scales in order to obtain higher reliability (a minimum value of 0.90 is required for individual use [18, 25]). It remains to be seen whether such questions can be generic across occupations, or whether company- or job-specific questions are more appropriate. When the reliability of the IWPQ scales is high enough for individual use, the IWPQ can be used for cross-sectional comparisons between, and longitudinal comparisons within, individuals. Participants' anonymity and confidentiality of answers should still be guaranteed, so that employees do not have to worry about their answers being seen by their supervisor(s) or colleague(s).

At a group level, tailored feedback could be given to managers or employers who want to monitor, benchmark and improve their own team, department or company, respectively. They could be given an overview of their team's, department's, or company's results, benchmarked against mean scores in similar groups. Future research should determine which advice can be given to managers and employers on how to improve their team, department or company scores on the IWP dimensions.

Finally, IWP may be influenced by the Zeitgeist. In other words, it may not be a stable construct over time. Changes in work, society, and the economy can affect what exactly constitutes performance at work. For example, whereas task performance was considered the main component of IWP in the past, contextual performance (and its variants) have gained more and more currency since the 1960s. Also, adaptive performance behaviors have gained ground into IWP theories and questionnaires, either as a distinct dimension or as part of the contextual performance dimension. Adaptive performance items that have come up in questionnaires because workers need to be increasingly adaptive, versatile, and tolerant, are for example: "*Do you periodically update your skills to accomplish the work or projects you are assigned?*" or *"Do you make friends with people from different countries?"* [26]. In one or two decades from now, the content of the dimensions may have changed, other dimensions may have arisen (e.g.,

environmental sustainability), and/or some existing dimensions may become obsolete. It is therefore important to constantly evaluate and scrutinize what the concept of IWP means, and whether we still measure what we want to measure.

Towards an integration of research fields

Multiple research fields, such as occupational health, work and organizational psychology, and management and economics, are interested in IWP. All have developed their own way of approaching IWP, ways that have influenced why and how measurement instruments are constructed in the first place, and the manner in which they are interpreted. The array of available questionnaires to measure (aspects of) IWP is daunting when searching the literature. Even within one research field, it seems that every study has used a different approach to measure (aspects of) IWP. This proliferation of instruments impedes the interpretation of study results, since findings cannot be compared with each other [8].

In order to advance research on IWP, an integration of the different research fields is proposed. The conceptualization and operationalization of IWP proposed in this thesis provides a good starting point for that, as they were based on all these research fields. We recommend researchers in all research fields to adopt the definition and conceptual framework of IWP proposed in this thesis. Secondly, the IWPQ may be used as the standard instrument to measure IWP in a comprehensive, generic, and short way. Below, it is described what an integration could add to each research field.

In the field of occupational health, the terms absenteeism, presenteeism, and productivity are often used interchangeably with IWP. Consensus on their definitions, and what sets them apart from each other, would bring clarity to the field. Current preventive intervention programs in occupational health, applied to a generic working population, focus on absenteeism and presenteeism as outcome measures, while most of the employees in the generic working population have low rates of absenteeism and presenteeism to start with. As a result, absenteeism and presenteeism instruments suffer from floor and/or ceiling effects, making it very difficult to find further improvements due to interventions. Using the IWPQ as the standard instrument to measure IWP, the field of occupational health will benefit from looking at a wider range of performance behaviors at work. This will enable researchers to show that occupational health interventions are not only suitable for maintaining, improving and optimizing IWP of employees that are frequently absent, or have reduced performance due to health complaints, but also of the larger, mostly

healthy, population of employees. A finding to support the additional value of looking at IWP, is that supervisor-rated performance at work is not that strongly related to the number of days that one is absent from work due to health complaints [27, 28]. This indicates that IWP and absenteeism are two distinct concepts, and both provide unique and valuable information. Furthermore, the field of occupational health could learn from the field of management and economics, by including company-level performance indicators in addition to individual-level indicators such as absenteeism, presenteeism and IWP.

In the field of work and organizational psychology, a lot of research has been done on how to define and measure IWP. A lot of work in the current thesis was drawn from work and organizational psychology. For example, the definition that was adopted in the present thesis originated from work and organizational psychologist Campbell [1]. The long research tradition of work and organizational psychology in the area of IWP has resulted in a lot of different labels and measures for the concepts of task performance, contextual performance, and counterproductive work behavior. The IWPQ could function as the standard measurement instrument for (the dimensions of) IWP, which would increase generalizability and comparability of findings, and avoid problems of existing measurement instruments from work and organizational psychology, such as a lack of content validity of the scales, and overlapping items between scales (leading to unrealistically high correlations between scales). In addition, work and organizational psychology could learn from occupational health, and management and economics, for example by including measures of absenteeism, and indicators of company-level performance, such as employee turnover, customer satisfaction, and financial performance.

In the field of management and economics, performance measurement mostly focuses on the company level, using key performance indicators such as employee turnover, customer satisfaction, and financial performance. This research field could learn from occupational health, and work and organizational psychology, to include IWP, and individual level outcomes such as absenteeism, employee health and well-being, and work engagement, in addition to company level outcomes. This would give insight into the relationship between IWP and company level outcomes – a relationship where there is currently little information on. In addition, including the IWPQ as an outcome measure can determine the effectiveness of company-level interventions, methods and strategies on IWP. This can also give insight into the

possibility that certain strategies might work for some groups of employees but not for others, under which conditions, and why.

Applications for practice

As stated previously, the main purpose of the IWPQ is research-oriented, namely, to establish the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP. Nevertheless, IWP is not only an important outcome measure in research, but also in practice. Words such as performance goals, performance feedback, and performance reviews, are familiar terms in many workplaces. The knowledge gained in the current thesis can be applied in practice in multiple ways. First of all, although the IWPQ is currently not suitable for individual assessments, evaluations, and selection processes, professionals dealing with IWP in the workplace (e.g., managers, human resource managers, and occupational health physicians) can keep in mind the three dimensions of IWP, when observing, evaluating, or improving IWP of employees. Both managers and employees can keep in mind the three dimensions of IWP when going into a performance review and formulating performance goals. For example, they should not only look at whether central job tasks are performed, but also at taking on extra tasks, keeping job knowledge and skills up-to-date, or showing excessive negativity in the workplace. Also, managers should be aware that these dimensions are not necessarily strongly related (i.e., an employee can perform contextual behaviors and counterproductive work behaviors simultaneously).

Second, the knowledge gained in the current thesis could benefit HRM and occupational physicians in companies. Core tasks of HRM include hiring and recruiting new personnel, assessing personnel, and providing training and development programs for personnel. HRM could use the IWPQ to identify groups of employees that perform well on the job (e.g., groups with certain personality traits, or skills), so that this knowledge can be used to hire and recruit new personnel. HRM could use the conceptual framework of IWP proposed in the current thesis (i.e., they should consider task performance, contextual performance, and counterproductive work behavior) when assessing personnel. Furthermore, HRM could use the IWPQ to determine which groups of employees need to improve on which aspect(s) of IWP, and subsequently, which training and development program(s) would be most suitable for those groups. Core tasks of occupational health and safety departments include providing a safe and healthy work environment, guiding sick employees in order to promote return to work, and

monitoring health of employees (e.g., health risk appraisals). Occupational physicians could use the IWPQ to monitor IWP of groups of employees reintegrating to the job after a period of sick leave.

Third, the knowledge gained in the current thesis is interesting for companies. They can include the IWPQ as a key performance indicator, in addition to existing ones such as employee turnover, customer satisfaction, financial performance. This way, the IWPQ can be used to monitor IWP. Also, by combining the IWPQ with other company-level measures, the relationship between IWP and company productivity and competitive ability can be untangled, and effective ways of improving company productivity and competitive ability through improving IWP can be identified.

Trends in work

A large part of the application of the IWPQ lies in the ability of research to answer questions that arise in practice. Current trends in work, such as those described in the General Introduction (globalization, economic recession, and sustainable employability), pose challenges to society that research can help overcome. In order to overcome these challenges, collaboration between research and practice is necessary. Knowledge gained on the predictors and effects of IWP in research can be used to develop interventions, procedures and strategies that are effective in maintaining, improving and optimizing IWP. Such interventions should be developed together with practice (e.g., employees, managers, HRM). Ultimately, this will improve the competitive ability of companies, optimize IWP in times of economic recession, and maintain employability up to an older age. Below, the questions that arise with each trend in work, and the role that research could play in answering them, are described.

Due to globalization, competition between companies from all over the world increases. Therefore, it is essential for companies to improve IWP of their employees, and thereby, their productivity and competitive ability. The IWPQ could be used to answer questions from practice regarding the effects of globalization. For example, globalization is heavily linked with advances in technology [29]. A question that might arise here, is if it is beneficial for IWP if employees are available on their smartphones and laptops 24/7? Does technology aid or hurt employee work engagement, work-home balance, and need for recovery after a workday, and how does this in turn affect IWP? Finally, the IWPQ can be used to gain knowledge on the predictors of IWP, and discover effective ways to improve IWP. For example, how

can technology aid employees, and when so (e.g., faster processes, higher quality) and when not (e.g., less personal communication, dependency on technology)?

Due to the current economic recession, it is important for companies to increase IWP of employees in order to stay afloat. Often, companies see their profits decrease, and in order to survive, they have to cut costs. This is often accomplished by employee reductions, reorganizations, or outsourcing work to cheaper markets. Also, employee salaries can be frozen, and bonuses downsized or eliminated. Such measures are often stressful for employees, lead to increased job insecurity, and subsequently, may have negative effects on IWP. On the other hand, employees might tend to increase their IWP in order to increase their chance to survive possible reorganizations. The IWPQ can be used to monitor the effects of the economic recession on IWP at the group level. It should be noted that, in times of economic recession, it is questionable whether employees will fill out the questionnaire honestly, because they are afraid there may be consequences if they perform poorly (e.g., job loss). In addition to monitoring IWP, research can help answer the question of how companies can increase IWP of employees, despite the economic recession and its possible negative influences on, for example, employee morale, stress, and job security. New Ways of Working is a recent and popular strategy to boost IWP, by giving employees more freedom and flexibility, while at the same time lowering company costs. The IWPQ could be used to determine how New Ways of Working influences IWP. Also, supplemented with additional measures, it could be used to examine which parts of New Ways of Working are especially effective in increasing IWP, and for whom (e.g., effectiveness dependent on age or personality?).

Third, it is increasingly important to promote sustainable employability of older workers in the workforce. The retirement age of workers in Europe will be increased in coming years [30], which means that employees have to work at the same – or even higher – level of IWP, despite possible limitations caused by an older age, such as reduced health. Several studies have found that both negative and positive stereotypes are associated with older workers [e.g., 31-33]. On the one hand, older workers are believed to be less willing to adapt to change, less motivated to learn, and less productive than younger workers. On the other hand, they are perceived to be more reliable, more loyal to the organization, and more experienced and knowledgeable than younger workers. The IWPQ can be used to monitor IWP with age. By including additional measures, predictors of older workers' IWP can be identified. To what extent do health and vitality impact IWP of older workers? Do older employees have difficulty to adapt to and keep up with advances in new

technology, and how does this affect their IWP? Knowledge on the predictors of IWP, and how we can influence them, can tell us how we can maintain IWP up to an older age. What can employees do to stay employable? How can employers facilitate them in staying employable? Are interventions, for example, focusing on a healthy lifestyle and vitality, organizational processes, or New Ways of Working, effective in maintaining, improving and optimizing IWP of older workers? How can technology be used to maintain, improve and optimize IWP of older workers?

A multi-disciplinary challenge

Maintaining, improving and optimizing IWP is a multi-disciplinary challenge, and we have the best chance of competing in a globalizing environment, overcoming the economic recession, and promoting sustainable employability when we approach this challenge in a multi-disciplinary way. With increased integration between research fields, and standardized measurement of IWP, predictors of IWP can be established even more precisely and comprehensively. Different research fields should draw on each other when developing interventions, procedures and strategies to maintain, improve, or optimize IWP, because research has shown that interventions with a multi-disciplinary approach are often most successful [e.g., 34, 35]. Finally, interventions, procedures and strategies should not be developed only by researchers, but in collaboration with practice (e.g., employees, managers, HRM). Cross-over of information between research fields and practice creates 'out of the box' thinking and can result in solutions that normally would not have been thought of.

Conclusion

The objective of this thesis – to develop and validate a comprehensive, generic, and short questionnaire to measure IWP – was achieved. The Individual Work Performance Questionnaire (IWPQ) is based on a conceptual framework consisting of three dimensions, namely, task performance, contextual performance, and counterproductive work behavior. The questionnaire is suitable for generic use (workers in all types of occupations, and workers with and without health problems) and is short with only 18 items. The internal consistency of the IWPQ was good and the construct validity was acceptable. Future research is necessary to determine the responsiveness of the IWPQ. A Dutch and American-English version of the IWPQ are available.

The current thesis contributes towards consensus on the definition and conceptualization of IWP, and hopefully, provides a push towards increased integration between research fields interested in IWP. Furthermore, it provides a reliable and valid instrument to measure IWP in a comprehensive, generic, and short way. Standardized measurement of IWP allows comparability and generalizability of findings, and increases knowledge on predictors and effects of IWP. This knowledge can be used to develop interventions, procedures and strategies to maintain, improve, or optimize IWP, and subsequently, evaluate their effectiveness. So, if you really want to know whether the promise of a financial bonus, being available on your mobile 24/7, or feeling happy, increases performance at work, as the news headings and articles in the General Introduction pose, use the IWPQ! As Lord Kelvin (1883) said: "*Measurement is knowledge*".

References

1. Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, editors. Handbook of industrial and organizational psychology, Vol.1 (2nd ed.). Palo Alto, CA, US: Consulting Psychologists Press; 1990. p. 687-732.
2. Borman WC, Motowidlo SJ. Expanding the criterion domain to include elements of contextual performance. In: Schmitt N, Borman WC, editors. Personnel Selection in Organizations. San Francisco, CA: Jossey Bass; 1993. p. 71-98.
3. Rotundo M, Sackett PR. The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. Journal of Applied Psychology. 2002;87(1):66-80.
4. Griffin MA, Neal A, Parker SK. A new model of work role performance: Positive behavior in uncertain and interdependent contexts. Academy of Management Journal. 2007;50:327-47.
5. Rasch G. Probabalistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
6. Van der Heijden BIJM, Nijhof AHJ. The value of subjectivity: Problems and prospects for 36-degree appraisal systems. The International Journal of Human Resource Management. 2004;15(3):493-511.
7. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology. 2003;88(5):879-903.
8. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development, 4th ed. Oxford University Press; 2008.
9. Jaramillo F, Carrillat FA, Locander WB. A meta-analytic comparison of managerial ratings and self-evaluations. Journal of Personal Selling & Sales Management. 2005;XXV(4):315-28.
10. Schoorman DF, Mayer RC. The value of common perspectives in self-reported appraisals: You get what you ask for. Organizational Research Methods. 2008;11(1):148-59.
11. Dalal RS. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. Journal of Applied Psychology. 2005;90:1241-55.

12. Berry CM, Carpenter NC, Barratt CL. Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. Journal of Applied Psychology. 2012;97(3):613-36.
13. Fox S, Spector PE. A model of work frustration-aggression. Journal of Organizational Behavior. 1999;20(6):915-31.
14. Coyne I, Gentile D, Born M, Ersoy NC, Vakola M. The relationship between productive and counterproductive work behaviour across four European countries. European Journal of Work and Organizational Psychology. 2013;22(4):377-89.
15. Viswesvaran C, Ones DS. Perspectives on models of job performance. International Journal of Selection and Assessment. 2000;8(4):216-26.
16. Holzbach RL. Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology. 1978;63(5):579-88.
17. Thornton GCI. Psychometric properties of self-appraisals of job performance. Personnel Psychology. 1980;33:263-71.
18. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge University Press; 2011.
19. Sudman S, Bradburn NM. Asking questions - A practical guide to questionnaire design. 1st ed. Sudman S and Bradburn NM, editors. San Francisco, USA: Jossey-Bass Publishers; 1982.
20. Tennant A, McKenna SP, Hagell P. Application of rasch analysis in the development and application of quality of life instruments. Value in Health. 2004;7:S22-6.
21. Prieto L, Alonso J, Lamarca R. Classical test theory versus rasch analysis for quality of life questionnaire reduction. Health Quality of Life Outcomes. 2003;1(27).
22. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum Publishers; 2000.
23. Coffeng JK, Hendriksen IJM, Duijts SF, Proper KI, Van Mechelen W, Boot CRL. Effectiveness of a combined social and physical environmental intervention on work-related outcomes in office employees, submitted.
24. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? when should it be applied, and what should one look for in a rasch paper? Arthritis & Rheumatism (Arthritis Care & Research). 2007 12/15;57(8):1358-62.

25. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
26. Pulakos ED, Schmitt N, Dorsey DW, Arad S, Hedge JW, Borman WC. Predicting adaptive performance: Further tests of a model of adaptability. Hum Perform. 2002;15(4):299-324.
27. Bycio P. Job performance and absenteeism: A review and meta-analysis. Human Relations. 1992;45(2):193-220.
28. Morrow PC, McElroy JC, Laczniak KS, Fenton JB. Using absenteeism and performance to predict employee turnover: Early detection through company records. Journal of vocational behavior. 1999;55:358-74.
29. Joshi RM. International business. Oxford University Press, New Delhi and New York; 2009.
30. European Commission. EUROPE 2020: A strategy for smart, sustainable and inclusive growth. Brussels: European Commission; 2010.
31. McGregor J, Gray L. Stereotypes and older workers: The New Zealand experience. Social Policy Journal of New Zealand. 2002(18):163-77.
32. Maurer TJ, Wrenn KA, Weiss EM. Toward understanding and managing stereotypical beliefs about older workers' ability and desire for learning and development. Research in Personnel and Human Resources Management. 2003;22:253-85.
33. Bal AC, Reiss AEB, Rudolph CW, Baltes BB. Examining positive and negative perceptions of older workers: A meta-analysis. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2011;66B(6):687-98.
34. Cancelliere C, Cassidy JD, Ammendolia C, Côté P. Are workplace health promotion programs effective at improving presenteeism in workers? A systematic review and best evidence synthesis of the literature. BMC Public Health. 2011;11:395.
35. Kahn-Marshall JL, Gallant MP. Making healthy behaviors the easy choice for employees: A review of the literature on environmental and policy changes in worksite health promotion. Health Education & Behavior. 2013;39:752-76.

Summary

Introduction

Due to due to increasing economic globalization and corresponding competitiveness between companies all over the world, the economic recession in many countries, and the growing need for sustainable employability, it becomes increasingly important to maintain, improve, and optimize the individual work performance (IWP) of employees (Chapter 1). Different scientific fields have proposed approaches and solutions for doing so. For example, the field of occupational health performs health risk appraisals or develops intervention studies targeting employee health (e.g., by improving working conditions, ergonomics, or a healthy lifestyle). The field of work and organizational psychology has been involved in hiring and recruiting personnel, assessment procedures, and training and development programs. The field of management and economics has primarily focused on the larger work system, including factors such as work processes, technological constraints, and organizational structure.

Despite the importance and popularity of IWP, there is little consensus on how to define and conceptualize this construct. When considering the research on IWP from the different research fields, it seems evident that a clear definition and conceptual framework of IWP is lacking. In accordance, a multitude of instruments exists to measure IWP (or one of its related constructs). This lack of consensus on how to define, conceptualize, and measure IWP is undesirable, because valid measurement is a prerequisite for accurately establishing the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP. Therefore, the objective of this thesis was to develop and validate a comprehensive, generic, and short questionnaire to measure IWP.

Part I. Developmental phase

The first step towards a comprehensive, generic, and short measure of IWP was establishing a clear definition and conceptualization of IWP. Chapter 2 presents a multi-disciplinary, systematic review of the literature on conceptual frameworks of IWP. In this chapter, the definition of IWP as “behaviors or actions that are relevant to the goals of the organization” (Campbell, 1990) was adopted. An integrated, conceptual framework was proposed, in which IWP consists of four broad and generic dimensions, namely, task performance, contextual performance, adaptive performance, and counterproductive work behavior.

This conceptual framework was used as the starting point for the development of the Individual Work Performance Questionnaire (IWPQ). In order to

operationalize the dimensions in this conceptual framework, numerous indicators used to measure IWP were identified in Chapter 3, via the scientific literature, existing questionnaires, and expert interviews. Subsequently, the most relevant indicators per dimension were selected by experts from different professional backgrounds. These indicators were used to construct a first version of the Individual Work Performance Questionnaire (IWPQ).

Part II. Field-testing phase

In Chapter 4, the field-testing of the first version of the Individual Work Performance Questionnaire (the IWPQ 0.1) is described. In order to examine its generic applicability, the IWPQ 0.1 was tested in a large sample of Dutch blue, pink, and white collar workers. The results of the field-test showed that the factor structure of the IWPQ consisted of three dimensions, with the contextual performance and adaptive performance questions loading on one factor. The conceptual framework was changed accordingly, by merging the contextual performance and adaptive performance dimensions. In addition, Rasch analysis (Rasch, 1960) was used to examine the functioning of the IWPQ items in more detail. Only items that showed good fit to the Rasch model, and that were generically applicable, were included in the second version of the IWPQ (the IWPQ 0.2).

For the IWPQ 0.2, it appeared that the targeting of the scales was not yet optimal. For task and contextual performance, there were insufficient items located at the higher range of the scale (i.e. difficult items), while for counterproductive work behavior, there were insufficient items sensitive to the lower range of the scale (i.e. easy items). As a consequence, the IWPQ was less able to discriminate workers with high task and contextual performance, and less able to discriminate workers with low counterproductive performance. In order to improve the targeting of the IWPQ, an improvement round was held, described in Chapter 5. Improved targeting of the task and contextual performance scales was achieved, by adding new items to the scales. The final version of the questionnaire – the IWPQ 1.0 – is presented.

Part III. Validation of the IWPQ

In Chapter 6, two types of construct validity of the IWPQ were examined. First, the relations of the IWPQ with presenteeism and work engagement were examined (convergent validity). The IWPQ scales appeared to correlate lower than expected with absolute presenteeism. These lower than expected correlations may be due to several reasons, such as limitations in administering the presenteeism measure in

the study, or lack of conceptual similarity between IWP and absolute presenteeism. As hypothesized, the IWPQ correlated weakly with relative presenteeism (own IWP compared to that of colleagues), and moderately with work engagement. Second, it was examined whether workers low and high in job satisfaction, and workers low and high in overall health, could be discriminated on IWPQ scores (discriminative validity). As expected, the IWPQ was able to discriminate between these groups. Overall, these results indicate acceptable construct validity of the IWPQ.

In Chapter 7, the responsiveness of the IWPQ was examined in the Be Active & Relax randomized controlled trial. The aim of this trial was to investigate the effectiveness of an intervention to stimulate physical activity and relaxation of office workers on need for recovery. Correlations between changes on the IWPQ and changes on similar constructs (e.g., presenteeism) and distinct constructs (e.g., need for recovery) used in the trial were examined. Although at least 85% of the correlations between change scores were in the expected direction, most of the correlations were weaker than expected. This may indicate low responsiveness of the IWPQ, but can also be explained by several other reasons, such as a relatively healthy, well-functioning study population at baseline, small changes on many constructs in the study, or unclear guidelines on how to interpret correlations between change scores. Thus, no firm conclusions could be drawn about the responsiveness of the IWPQ, and future research regarding this characteristic is recommended.

In order to promote international use of the IWPQ, Chapter 8 presents the cross-cultural adaptation of the IWPQ from the Dutch to the American-English language, using the guidelines of Beaton et al. (2000). This process consisted of five steps: a forward translation by two independent native-American translators, synthesis, back-translation by two independent native-Dutch translators, an expert committee review, and pilot-testing. During the pilot-testing, cognitive interviews with 40 American workers were performed, to examine the comprehensibility, applicability, and completeness of the American-English IWPQ. The translation was conducted without major difficulties. In general, participants were positive on the comprehensibility, applicability and completeness of the questionnaire. Five items were adapted to better suit the American-English language. The pilot-test showed good results concerning the internal consistency of the American-English IWPQ (Cronbach's alphas for the scales between 0.79 and 0.89) and good content validity. The results indicate that the cross-cultural adaptation was successful, and that the measurement properties of the American-English IWPQ were good.

Discussion and conclusion

Chapter 9 provides an overview of the main findings and a discussion of the results. Methodological issues, such as generic applicability, the choice for self-report, and the lack of a golden standard for validation, are discussed. Strengths and limitations of the current thesis, as well as recommendations for future research on the IWPQ, are considered. The current thesis contributes towards consensus on the definition and conceptualization of IWP, and provides an instrument to measure IWP in a standardized way. Hopefully, this provides a push towards integration between research fields interested in IWP. Finally, applications of the knowledge gained in the current thesis are discussed for research and practice.

In conclusion, the objective of this thesis – to develop and validate a comprehensive, generic, and short questionnaire to measure IWP – was achieved. The main benefits of the IWPQ are that it measures all relevant dimensions of IWP, it is generically applicable to workers from different occupational sectors and workers with and without health problems, and it is short with 18 items. Such a measurement instrument is a prerequisite for accurately establishing the effectiveness of interventions, procedures and strategies to maintain, improve, or optimize IWP.

Samenvatting

Inleiding

Door de toenemende globalisering en bijbehorende concurrentiegroei tussen bedrijven over de hele wereld, de economische recessie in veel landen, en de groeiende behoefte aan duurzame inzetbaarheid, wordt het behouden, verbeteren en optimaliseren van individuele werkprestatie (IWP) van werknemers steeds belangrijker (hoofdstuk 1). Verschillende wetenschappelijke velden dragen hiervoor aanpakken en oplossingen aan. Vanuit het veld van de arbeidsepidemiologie worden bijvoorbeeld gezondheidschecks ingezet, en interventies gericht op het bevorderen van de gezondheid van werknemers (door bijvoorbeeld het verbeteren van de arbeidsomstandigheden, ergonomie, of een gezonde leefstijl) ontwikkeld. Het veld van de arbeids- en organisatiepsychologie is betrokken bij het werven en aannemen van personeel, beoordelingsprocedures, en opleidings- en ontwikkelings-programma's. Het veld van management en economie richt zich voornamelijk op het totale werksysteem, en kijkt naar factoren zoals werkprocessen, technologische beperkingen en organisatiestructuur.

Ondanks het belang en de populariteit van IWP, is er weinig consensus over hoe we dit construct moeten definiëren en conceptualiseren. Wanneer men kijkt naar onderzoek uit de verschillende wetenschappelijke velden, lijkt een duidelijke definitie en conceptueel model van IWP te ontbreken. In overeenstemming met het gebrek aan een duidelijke definitie en conceptueel model, is er een veelvoud aan instrumenten om IWP (of daaraan gerelateerde constructen) te meten. Het gebrek aan consensus over hoe we IWP moeten definiëren, conceptualiseren en meten, is onwenselijk, omdat een valide meting een voorwaarde is om de effectiviteit van interventies, procedures en strategieën om IWP te behouden, verbeteren, en optimaliseren accuraat in kaart te brengen. Het doel van dit proefschrift was om een volledig, generiek, en kort meetinstrument voor IWP te ontwikkelen en valideren.

Deel I. Ontwikkelingsfase

De eerste stap richting de ontwikkeling van een volledig, generiek, en kort meetinstrument voor IWP was het vaststellen van een duidelijke definitie en conceptueel model van IWP. Hoofdstuk 2 presenteert een multidisciplinaire, systematische review van de literatuur over conceptuele modellen van IWP. In dit hoofdstuk is gekozen voor de definitie van IWP als "*gedragingen of acties die relevant zijn voor de doelstelling van de organisatie"* (Campbell, 1990). Een geïntegreerd conceptueel model wordt voorgesteld, waarin IWP bestaat uit vier

brede, generieke dimensies: taakprestatie, contextuele prestatie, adaptieve prestatie, en contraproductief werkgedrag.

Dit conceptueel model bood het startpunt voor de ontwikkeling van de Individuele Werkprestatie Vragenlijst (IWPV). Om de dimensies in het conceptueel model te operationaliseren, zijn alle mogelijke indicatoren van IWP geïdentificeerd met behulp van de wetenschappelijke literatuur, bestaande vragenlijsten, en interviews met experts (hoofdstuk 3). Vervolgens werden de meest relevante en generieke indicatoren voor elke dimensie geselecteerd door experts met verschillende achtergronden. De geselecteerde indicatoren zijn gebruikt om de eerste versie van de IWPV te ontwikkelen.

Deel II. Veldtest fase

In hoofdstuk 4 wordt de veldtest van de eerste versie van de IWPV beschreven (de IWPV 0.1). Om de generieke toepasbaarheid te toetsen, is de IWPV 0.1 getest in een brede steekproef van Nederlandse *blue collar* (fysieke arbeiders), *pink collar* (dienstverleners), *en white collar* (kantoorwerkers) werknemers. Uit de resultaten van de veldtest bleek de factor structuur van de IWPV uit drie dimensies te bestaan, waarbij de vragen uit de contextuele prestatie en adaptieve prestatie dimensies op één factor laadden. Op basis hiervan is het conceptueel model aangepast door de contextuele en adaptieve prestatie dimensies samen te voegen. Vervolgens is Rasch analyse (Rasch, 1960) gedaan om gedetailleerder te kijken naar het functioneren van de IWPV vragen. Alleen de vragen die goed in het Rasch model pasten, en generiek toepasbaar waren, zijn geïncludeerd in de tweede versie van de IWPV (de IWPV 0.2).

Op basis van de resultaten uit hoofdstuk 4 bleek de targeting van de IWPV 0.2 nog niet optimaal. Voor de taakprestatie en de contextuele prestatie schaal, waren er onvoldoende vragen die het hogere bereik van de schaal maten (d.w.z. moeilijke vragen), terwijl er voor de contraproductief werkgedrag schaal onvoldoende vragen waren die het lagere bereik van de schaal maten (d.w.z. makkelijke vragen). Hierdoor was de IWPV 0.2 minder goed in staat om werknemers met een hoge taakprestatie en hoge contextuele prestatie van elkaar te onderscheiden, en minder goed in staat om werknemers met laag contraproductief werkgedrag van elkaar te onderscheiden. Om de targeting van de IWPV te verbeteren is een verbeteringsronde gehouden, die wordt beschreven in hoofdstuk 5. Verbeterde targeting van de taakprestatie en contextuele prestatie schaal werd bereikt door nieuwe vragen aan de schalen toe te voegen. De definitieve versie van de vragenlijst – de IWPV 1.0 – is geboren.

Deel III. Validatie van de IWPV

In hoofdstuk 6 worden twee typen construct validiteit van de IWPV onderzocht. Als eerste werd de relatie van de IWPV met presenteïsme en werkbevlogenheid onderzocht (convergente validiteit). De IWPV schalen bleken lager dan verwacht te correleren met absoluut presenteïsme. Deze lager dan verwachte correlaties kunnen verschillende redenen hebben, zoals beperkingen in het afnemen van de presenteïsme vragen in het onderzoek, of gebrek aan conceptuele gelijkheid van IWP en absoluut presenteïsme. Zoals verwacht correleerde de IWPV zwak met relatief presenteïsme (eigen IWP vergeleken met collega's), en matig met werkbevlogenheid. Als tweede werd onderzocht of werknemers met een lage versus hoge werktevredenheid, en een slechte versus goede gezondheid, te onderscheiden waren op de IWPV (discriminatoire validiteit). Zoals verwacht was de IWPV in staat om onderscheid te maken tussen deze groepen. In het geheel genomen is de construct validiteit van de IWPV acceptabel.

In hoofdstuk 7 is de responsiviteit van de IWPV onderzocht in de Be Active & Relax gerandomiseerde gecontroleerde trial. Het doel van deze trial was om de effectiviteit van een interventie gericht op het stimuleren van bewegen en ontspanning van kantoormedewerkers te onderzoeken op herstelbehoefte. Correlaties tussen verschilscores op de IWPV en verschilscores op gelijke constructen (zoals presenteïsme) en niet-gelijke constructen (zoals herstelbehoefte) gebruikt in de trial, zijn onderzocht. Hoewel ten minste 85% van de correlaties in de verwachte richting waren, waren de meeste zwakker dan verwacht. De zwakker dan verwachte correlaties kunnen slechte responsiviteit van de IWPV aanduiden, maar kunnen ook verklaard worden door andere redenen, zoals een relatief gezonde, goed functionerende onderzoekspopulatie, kleine veranderingen op veel constructen in de studie, of onduidelijke richtlijnen over het interpreteren van correlaties tussen verschilscores. Er kunnen nog geen sterke conclusies worden getrokken over de responsiviteit van de IWPV, en meer onderzoek naar dit kenmerk is nodig.

Om internationaal gebruik van de IWPV te stimuleren, wordt in hoofdstuk 8 de cross-culturele aanpassing van de Nederlandse naar de Amerikaans-Engelse taal, volgens de richtlijnen van Beaton et al. (2000), gepresenteerd. Het proces bestond uit vijf stappen: een heenvertaling door twee onafhankelijke vertalers met Amerikaans als moedertaal, consensus, een terugvertaling door twee onafhankelijke vertalers met Nederlands als moedertaal, consensus van de expertgroep, en een pilot-test. Tijdens de pilot-test zijn interviews met 40 Amerikaanse werknemers gehouden om de begrijpelijkheid, toepasbaarheid, en volledigheid van de

Amerikaans-Engelse IWPV te onderzoeken. De vertaling werd uitgevoerd zonder grote problemen. In het algemeen waren deelnemers positief over de begrijpelijkheid, toepasbaarheid, en volledigheid van de vragenlijst. Vijf vragen zijn aangepast om beter aan te sluiten op het Amerikaans-Engels. De pilot-test liet goede resultaten zien wat betreft de interne consistentie van de Amerikaans-Engelse IWPV (Cronbach's alphas voor de schalen tussen 0.79 en 0.89) en goede content validiteit. De resultaten wijzen erop dat de cross-culturele aanpassing succesvol was, en dat de meeteigenschappen van de Amerikaans-Engelse IWPV goed zijn.

Discussie en conclusie

Hoofdstuk 9 geeft een overzicht van de belangrijkste bevindingen en een bespreking van de resultaten. Methodologische kwesties, zoals de generieke toepasbaarheid, de keuze voor zelfrapportage, en het ontbreken van een gouden standaard ter validatie, worden besproken. Sterktes en zwaktes, evenals aanbevelingen voor toekomstig onderzoek naar de IWPV, worden gepresenteerd. Het huidige proefschrift draagt bij aan consensus over de definitie en de conceptualisatie van IWP, en voorziet in een instrument om IWP op een gestandaardiseerde manier te meten. Hopelijk draagt dit bij aan de integratie tussen de verschillende onderzoeksvelden die IWP bestuderen. Ten slotte worden toepassingen van de in het proefschrift ontwikkelde kennis besproken voor onderzoek en praktijk.

Concluderend is het doel van dit proefschrift – om een volledig, generiek, en kort meetinstrument voor IWP te ontwikkelen en valideren – bereikt. De belangrijkste voordelen van de IWPV zijn dat deze alle relevante dimensies van IWP meet, generiek toepasbaar is voor werknemers uit verschillende beroepssectoren en werknemers met en zonder gezondheidsproblemen, en kort met slechts 18 vragen. Een dergelijk instrument is een voorwaarde om de effectiviteit van interventies, procedures en strategieën om IWP te behouden, verbeteren, en optimaliseren accuraat in kaart te brengen.

Dankwoord

Na vier jaar hard werken zit het er op. Het proefschrift is af! Het is waar dat de laatste loodjes het zwaarst vallen, maar dit dankwoord is daar een uitzondering op. Het is een eer om de onderstaande mensen, die de afgelopen jaren allemaal hebben bijgedragen aan mijn proefschrift, te bedanken.

Claire, jij was mijn dagelijkse begeleidster en daar ben ik erg dankbaar voor! Je nam deze taak heel serieus en besteedde altijd veel aandacht aan mijn stukken en vragen. Door jou kritische vragen ("waarom?") hielp je het proefschrift naar een hoger niveau tillen. Je was niet alleen een fijne begeleidster, maar ook een fijne collega! **Vincent**, bedankt voor je begeleiding tijdens mijn proefschrift! In het begin was je er iets vaker dan aan het eind, maar ik zie dit als een compliment dat je er wel vertrouwen in had. Als ik het nodig had, was je er altijd om structuur te geven of me aan het denken te zetten. **Riekie**, wat ben ik blij dat jij een van mijn begeleiders was! Zonder jouw kennis over klinimetrie was dit project niet geslaagd. Jouw snelle reactie op vragen en artikelen, en het feit dat je deur altijd openstond, heeft mij altijd positief verrast! **Allard**, bedankt dat je mij ruim vier jaar geleden hebt aangenomen, ook al kon ik het woord "epidemiologie" toen nog niet vloeiend uitspreken. Je las mijn artikelen altijd met veel detail, en ik kreeg deze altijd terug met jouw karakteristieke asterisken bij elk verkeerd gespeld woord of gemiste komma. Daarnaast heb ik van jou het woord "helikopterview" en wat dit inhoudt geleerd. Bedankt voor je begeleiding!

Alle **collega's, vrienden en familie**, ontzettend bedankt voor jullie interesse in mijn onderzoek, jullie adviezen en discussies over wat werkprestatie nu precies is, hoe je dat kan meten, en wat je daar dan aan hebt. In het bijzonder wil ik de **leden van de klankbordgroep** bedanken, die mij de afgelopen 4 jaar met hun expertise hebben bijgestaan in de ontwikkeling van het theoretische model en van de Individual Work Performance Questionnaire (IWPQ). Daarnaast is het zonder **deelnemers** natuurlijk niet mogelijk om een vragenlijst te ontwikkelen, te testen, en te verbeteren. Ik wil graag iedereen bedanken die in de beginfase de IWPQ hebben ingevuld en van feedback hebben voorzien. Mijn dank gaat uit naar alle anonieme deelnemers die via internetpanels de IWPQ hebben ingevuld, en alle experts die mij in mijn studies hebben voorzien van informatie en advies. Ook wil ik de **vertalers** bedanken, die hebben gewerkt aan de Amerikaanse versie van de IWPQ, en de medewerkers van Tufts Medical Center die hebben geholpen bij het tot stand komen van de definitieve Amerikaanse versie. Tot slot wil ik alle **onderzoekers** bedanken die hun interesse

hebben getoond in de IWPQ en deze zelfs al in onderzoek inzetten. Ik hoop dat er in de toekomst veel gebruik zal worden gemaakt van de IWPQ en dit mooie resultaten oplevert!

Beste **co-auteurs,** bedankt voor de tijd die jullie aan mijn artikelen hebben besteed om deze naar een hoger niveau te tillen! Wilmar, bij mijn allereerste artikel was jij co-auteur en hielp je ons goed aansluiten bij de A&O psychologie. Bedankt voor je hulp en vertrouwen! Stef, bedankt voor je uitgebreide uitleg en hulp bij de Rasch analyses! Jennifer en Cécile, erg leuk ik in zo'n vroeg stadium van de vragenlijst al aan mocht sluiten bij jullie onderzoek, en bedankt voor jullie hulp en input bij het artikel! Debra, thank you for welcoming me at the ICRHPS in Boston and taking me under your wing!

Beste **leden van de leescommissie**, bedankt voor de tijd en aandacht die jullie aan mijn proefschrift hebben besteed. Ik kijk, met gezonde spanning, uit naar de verdediging!

Lieve **TNO collega's**, een eerste werkplek blijft toch altijd bijzonder. Ik heb vier hele fijne jaren met jullie gehad. De interesse in mij en mijn onderzoek, en de gezelligheid in de gang, tijdens lunches, en bij PV uitjes, zal ik missen! Ik wens jullie heel veel geluk bij alles wat jullie doen, en laten we vooral contact houden!

Lieve collega's van de **G/H-0 gang** (Laura, Linda, Jantien, Jennifer, Hanneke, Alwin, Ruben, Myrthe, Esther, Karen, Astrid, Judith, Caroline, en anderen) al snel leerde ik ook jullie kennen en wat was het leuk om samen met allemaal AIO's op een gang te zitten. Waar bij TNO bijna geen AIO's waren, zaten we op onze gang met alleen maar AIO's. Een feest van herkenning. Ook al was ik er niet heel vaak, hebben we elkaar toch goed leren kennen en goede banden opgebouwd!

Naast directe collega's, waren er ook anderen die ik regelmatig tegenkwam. **Suzanne**, ik vond het heel leuk om regelmatig met jou een cursus te volgen, koffie te drinken, of te lunchen. Ik wens je heel veel geluk samen met je vriend in Noorwegen, of waar dan ook! **Gerdien** en **Annemiek**, bedankt voor jullie gezelligheid tijdens onze congressen!

Debbie en Jennifer, wat leuk dat jullie mijn paranimfen zijn! Het avontuur van een proefschrift schrijven beleefden we vaak samen. **Debbie**, wat was het gezellig als kamergenootjes bij TNO! In het begin was het misschien even wennen, maar al snel kwamen we erachter dat we veel gemeenschappelijk hadden. Ik denk dat we veel van elkaar hebben geleerd de afgelopen jaren. Naast het werk was er ook altijd wel iets om het over te hebben: eten, sporten, vriendinnen, noem maar op. Ik ga je missen als collega, maar ik hoop dat we elkaar nog regelmatig blijven zien! **Jennifer**, wij zijn ongeveer gelijk begonnen en zaten vaak in dezelfde fase van het proefschrift. Ook vonden we elkaar al snel op EMGO+ dagen en cursussen. Ik vond het erg leuk dat mijn vragenlijst in jouw onderzoek mee mocht, en we op die manier nog even konden samenwerken. Eén van de hoogtepunten was toch wel het congres in Los Angeles vorig jaar, met alle spanning van het houden van een presentatie, tot een dagje Universal Studios en wandelingen door de achterbuurten van LA!

Mijn nieuwe collega's bij het **NIVEL**, jullie hebben vooral de eindfase, en mijn stress over het opmaken van de cover en de inhoud van het proefschrift meegekregen! Nikki, wat leuk dat wij tegelijkertijd bij het NIVEL zijn gekomen en precies in dezelfde fase zitten. Lisanne, wat leuk dat we na de G/H-0 gang nu weer collega's zijn. Ik kijk ernaar uit om jullie allemaal in de toekomst beter te leren kennen en samen veel mooie projecten met wetenschappelijk en maatschappelijke impact te doen!

Lieve **vrienden**, bedankt voor al jullie interesse in mijn onderzoek de afgelopen jaren, en voor de nodige ontspanning daarvan! De gezellige etentjes, avondjes uit, en weekendjes weg, boden de nodige afleiding van mijn proefschrift. Sander en Olga, wat was het heerlijk om met jullie op Koh Samui te ontspannen voordat ik weer aan de slag moest! Koen, bedankt dat je zo "streng" voor me was tijdens de laatste loodjes. Ondanks de achtergrondgeluiden van The Walking Dead, is het toch allemaal best goed gelukt! ;) Anna, Antje en Patricia, we kennen elkaar al sinds de basisschool, en ik vind het heel bijzonder dat wij al zo lang vriendinnen zijn! Ook al zien we elkaar niet meer zo vaak als vroeger, en zijn er mannen en kinderen bijgekomen, elke keer we elkaar zien is het nog als vanouds. Ik hoop dat we nog heel lang vriendinnen blijven!

Lyn and Graham, your interest in my PhD research has been heartwarming. You not only showed your support in spirit, but occasionally also helped with translating questionnaire items and checking my English writing. Writing the General Discussion was a challenge, but all that Costa coffee got us through. Lyn, I'm so happy that you made such an amazing recovery. And Graham, rest assured that you will not be the only Dr. Taylor in the family for long! ;)

Leave **heit en mem, Sandra en Sjoerd**, bedankt voor jullie steun, vertrouwen en interesse. Zonder jullie had ik dit niet kunnen bereiken. Het moet voor jullie vast wennen zijn geweest toen ik uit ons kleine dorpje naar de grote stad vertrok. Ook al was het voor jullie misschien lastig om een voorstelling te maken van wat ik nou echt deed, jullie waren altijd geïnteresseerd in mijn werk en de eersten die om een exemplaar vroegen als een artikel werd gepubliceerd. Elke keer ik in Hieslum kom, voelt nog altijd als thuiskomen!

Andrew, bedankt voor al je interesse, steun, en aanmoediging de afgelopen jaren! En natuurlijk, voor je hulp en geduld met mij in de afgelopen maanden. Je hebt mij altijd geïnspireerd om het beste uit mezelf te halen, en dit proefschrift is daar zeker een resultaat van. Ik ben blij dat ik jou ook heb kunnen inspireren om je eigen bedrijf te beginnen. Ik ben trots op je en weet zeker dat jij in de toekomst nog veel anderen zult inspireren met jouw enthousiasme en overtuigingskracht! Ik kan me een leven zonder jou niet voorstellen, en kijk uit naar een fantastische toekomst samen!

About the Author

Linda Koopmans was born on December 15th 1985 in Hieslum, a small village in the municipality of Wonseradeel, The Netherlands. After completing VWO at the Marne College in Bolsward in 2004, she studied psychology at Utrecht University. In 2007, she received her Bachelor degree in Social Psychology. In 2009, she received her Master degree in Psychological Health Research. After graduation, Linda shortly worked as a junior researcher at the Netherlands Organisation of Applied Scientific Research (TNO). In December 2009, she started her PhD research at Body@Work, at the Department of Life Style (TNO, Leiden) and the Department of Public and Occupational Health (VU University Medical Center, EMGO+ Institute for Health and Care Research). In the summer of 2013, she conducted part of her PhD research at the Program on Health, Work and Productivity, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA. As of February 2014, Linda works as a Postdoc researcher at the Netherlands Institute for Health Services Research (NIVEL). She will examine the impact of diversity within teams on team functioning, job satisfaction, and quality of health care.

List of Publications

Scientific articles

1. Koopmans L, Bernaards CM, Hildebrandt VH, Schaufeli WB, De Vet HCW, Van der Beek AJ. Conceptual Frameworks of Individual Work Performance – A Systematic Review. Journal of Occupational and Environmental Medicine 2011;53(8):856-866.
2. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Measuring Individual Work Performance - Identifying and Selecting Indicators. Work: A Journal of Prevention, Assessment and Rehabilitation, 2013;45(3).
3. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Development of an Individual Work Performance Questionnaire. International Journal of Productivity and Performance Management 2013;62(1):6-28.
4. Koopmans L, Bernaards CM, Hildebrandt VH, Van Buuren S, Van der Beek AJ, De Vet HCW. Improving the Individual Work Performance Questionnaire using Rasch Analysis. Journal of Applied Measurement 2014;15(2).
5. Koopmans L, Bernaards CM, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Construct Validity of the Individual Work Performance Questionnaire. Journal of Occupational and Environmental Medicine 2014;56(3).
6. Koopmans L, Coffeng JK, Bernaards CM, Boot CRL, Hildebrandt VH, De Vet HCW, Van der Beek AJ. Responsiveness of the Individual Work Performance Questionnaire, submitted.
7. Koopmans L, Bernaards CM, Hildebrandt VH, Lerner D, De Vet HCW, Van der Beek AJ. Cross-Cultural Adaptation of the Individual Work Performance Questionnaire, submitted.

Popular-scientific articles and reports

1. Koopmans L. Individual Work Performance Questionnaire – Manual. Nov 2013.
2. Koopmans L, Bernaards CM, Hildebrandt VH, Van der Beek AJ, De Vet HCW. De Individuele Werkprestatie Vragenlijst (IWPV): interne consistentie, construct validiteit en normering, ingediend.

3. Koopmans L, Bernaards CM, Hildebrandt VH, Van der Beek AJ, De Vet HCW. Meten is werkprestatie weten. Praktijkblad Preventie, Nov 2011.
4. Koopmans L, Bernaards CM, Hildebrandt VH, Van der Beek AJ, De Vet HCW. Werkprestatie: wat is het en hoe meet je het? Tijdschrift voor Ergonomie, Oct 2011.
5. Koopmans L, Engbers L, Gelinck R, Hildebrandt VH. Vitaal belang: waardevolle rol voor leefstijlbeleid ouderen. Arbo, Apr 2010.
6. Koopmans L. Leefstijl interventies voor oudere werknemers. TNO-rapport, KvL/GB 2010.002. Leiden, Jan 2010.
7. Koopmans L, Engbers L, Gelinck R, Hildebrandt VH. Vitale oudere werknemer krijgt steeds meer aandacht – Belangrijke rol weggelegd voor leefstijlbeleid. TNO-rapport, KvL/GB 2009.126. Leiden, Dec 2009.

Presentations

1. Guest lecturer Organizational Psychology at Radboud University Nijmegen, The Netherlands (2014).
2. Northeastern University, Boston, MA, USA (2013).
3. Conference on Epidemiology in Occupational Health (EPICOH), Utrecht, The Netherlands (2013).
4. European Association of Work and Organizational Psychology (EAWOP), Münster, Germany (2013).
5. Work, Stress and Health conference, Los Angeles, CA, USA (2013).
6. ArboNed, Utrecht, The Netherlands (2012).
7. Key-note speaker at Office Ergonomics Research Committee, Boston, MA, USA (2012).
8. European Association Occupational Health Psychology (EAOHP), Zurich, Switzerland (2012).
9. European Association of Work and Organizational Psychology (EAWOP), Maastricht, The Netherlands (2011).
10. Body@Work retreat 'Environment and Healthy Living' (2010).