Received 19 March 2012
Revised 19 July 2012
Accepted 2 August 2012

Development of an individual work performance questionnaire

Linda Koopmans
Life Style, Body@Work, TNO, Leiden, The Netherlands and Department of Public and Occupational Health, VU University Medical Center, Amsterdam, The Netherlands
Claire Bernaards and Vincent Hildebrandt
Life Style, TNO, Leiden, The Netherlands
Stef van Buuren
Life Style, TNO, Leiden, The Netherlands and Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
Allard J. van der Beek
Department of Public and Occupational Health, VU University Medical Center, Amsterdam, The Netherlands, and
Henrica C.W. de Vet
Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands

Abstract

Purpose – The purpose of the current study is 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.

Design/methodology/approach – 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.

Findings – 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.

Research limitations/implications – 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.

Originality/value – This study shows that, using Rasch analysis, a generic and short questionnaire can be used to measure individual work performance.

Keywords Performance, Performance measurement, Individual work performance, Job performance, Measurement, Rasch analysis, Performance management, The Netherlands

Paper type Research paper

International Journal of Productivity and Performance Management
Vol. 62 No. 1, 2013
pp. 6-28

1741-0401
DOI 10.1108/17410401311285273

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. Rotundo and Sackett, 2002; Dalal, 2005). After all, a clear definition and theoretical framework of IWP is a prerequisite for valid measurement of the construct.

IWP was defined by Campbell (1990, p. 704) as "behaviors or actions that are relevant to the goals of the organization." Thus, IWP focusses 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 (Rotundo and Sackett, 2002). 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 (Campbell, 1990, pp. 708-9). Behaviors used to describe task performance often include work quantity and quality, job skills, and job knowledge (e.g. Rotundo and Sackett, 2002; Campbell, 1990).

Although it has long been recognized that IWP is a multidimensional construct (Campbell, 1990; Austin and Villanova, 1992), only more recently has the role of employee behaviors beyond task performance received full attention (e.g. Rotundo and Sackett, 2002; Dalal, 2005; Borman and Motowidlo, 1993). It is now generally agreed upon that, in addition to task performance, the IWP domain consists of contextual performance and counterproductive work behavior (CWB) (e.g. Rotundo and Sackett, 2002; Viswesvaran and Ones, 2000). Contextual performance can be defined as behaviors that support the organizational, social, and psychological environment in which the technical core must function (Borman and Motowidlo, 1993, p. 73). Behaviors used to describe contextual performance include, for example demonstrating effort, facilitating peer and team performance, cooperating, and communicating (Rotundo and Sackett, 2002; Campbell, 1990). CWB can be defined as behavior that harms the well-being of the organization (Rotundo and Sackett, 2002, p. 69). It includes behaviors such as absenteeism, off-task behavior, theft, and substance abuse (Koopmans *et al.*, 2011).

A recent review by Koopmans *et al.* (2011) has identified the new and upcoming dimension of adaptive performance in IWP frameworks (e.g. Pulakos *et al.*, 2000; Sinclair and Tucker, 2006; Griffin *et al.*, 2007). This dimension focusses 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 (Griffin *et al.*, 2007, p. 331).

Numerous scales have been developed to measure the dimensions of IWP. For example, Williams and Anderson (1991) 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 (1989) or Van Scotter and Motowidlo (1996). The former focusses on measuring altruism, conscientiousness, sportsmanship, courtesy, and civic virtue. The latter focusses on measuring interpersonal facilitation and job dedication. Scales used to assess CWB were developed by, for example, Bennett and Robinson (2000) or Spector *et al.* (2006). The former authors focus on measuring organizational 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.* (2000). It measures, for example, whether employees are able to solve problems creatively, to deal with uncertain or unpredictable work situations, and to learn 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 (1991) was based on a sample of employees with a technical/professional background, and the contextual performance scales by Podsakoff and MacKenzie (1989) and Van Scotter and Motowidlo (1996) were based on a sample of petrochemical employees and US 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 (Dalal, 2005). That is, items overlapping in content can be found in scales measuring different dimensions. This is especially the case for contextual performance and CWB 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 CWB 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. Methods discusses the developmental process of the IWPQ. It describes the field testing in a generic population and the analysis of the resultant data. Results 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

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 (Koopmans *et al.*, 2011). For each dimension, one scale

was developed. The operationalization of the scales was based on a study by Koopmans *et al.* (submitted). 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 among 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 in Koopmans *et al.* (submitted), 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 II). 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 eight (e.g.: "I have demonstrated flexibility"), and CWB of ten (e.g.: "I complained about unimportant matters at work").

Pilot-testing

A pilot study among 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 four weeks to three 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 four 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 three months and a five-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 (Dalal, 2005). Frequency ratings require individuals to recall and mentally calculate how often one engaged in each behavior (Schwarz and Oyserman, 2001), 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;

Van der Heijden and Nijhof, 2004), 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 (Streiner and Norman, 2008). For these reasons, the remaining task, contextual, and adaptive behaviors (items 7-38) were rated from "seldom," "sometimes," "frequently," "often," to "always." As the counterproductive behaviors (items 39-49) were expected to produce floor rather than ceiling effects, the center of this scale was shifted to the right, ranging from "never," "seldom," "sometimes," "frequently," and "often."

Field testing

The IWPQ was tested in a study among 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 five-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 1-10 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. In all, 14 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 percent at the extreme values; De Vet *et al.*, 2011). 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 data set 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 (Rasch, 1960), 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 (Tennant *et al.*, 2004; Tennant and Conaghan, 2007). 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 ordening (does the category ordening of polytomous items work as expected), and differential item functioning (differential item functioning (DIF); do subgroups in the sample respond differently to items). Analyses were conducted using RUMM2020 software (Andrich *et al.*, 2003).

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: χ^2-fit, item fit residuals, and person fit residuals. The χ^2-fit statistic is an item-trait interaction score, reflecting the property of invariance across the trait. Generally, a non-significant χ^2-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 (Traub, 1983; Lundgren Nilsson and Tennant, 2011). 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 (Andrich and Styles, 2009). 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 (Tennant and Conaghan, 2007).

Reliability. The person separation index (PSI) estimates the internal consistency of a scale. PSI is similar to Cronbach's α (Cronbach, 1951), 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 (Tennant and Conaghan, 2007).

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 (De Vet *et al.*, 2011).

Category ordening. In addition to good model fit, the data have 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 (Tennant and Conaghan, 2007).

DIF. Rasch analysis assumes that a scale functions consistently, irrespective of subgroups within the sample being assessed. 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 (Tennant and Conaghan, 2007).

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 correlations0.40. Local independence is often used to give an indication of unidimensionality of a scale (Tennant and Conaghan, 2007).

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 (Tennant and Conaghan, 2007). 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

In total, 1,181 Dutch workers filled in the 47-item IWPQ in June 2011. Participants were all employed, and aged 18-65+ years. Almost half of the participants (49.5 percent) were females. 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 I presents further participant characteristics.

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

	Total sample (%)	Occupational sector: 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 (years)				
18-24	6	5	9	2
25-34	17	13	16	23
35-44	27	28	25	29
45-54	31	31	32	30
55-64	18	22	18	16
65+	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

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

slightly less understandable than pink ($M = 3.3$, SD = 0.6) and white collar workers ($M = 3.3$, SD = 0.7), $F(2, 1{,}178) = 4.037$, $p < 0.05$. 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, $F(2, 1{,}178) = 2.071$, $p > 0.05$.

Conceptual framework

In all, 38 of the 47 items showed ceiling effects, i.e. more than 15 percent of the responses at the high end of the scale. Especially CWB items (recoded) showed ceiling effects, ranging up to 96.6 percent of the scores at the extreme value. None of the items showed very low (> 0.20) or very high (> 0.90) inter-item correlations. In total, 14 items had a "not applicable" category, which was used by 14 percent of the respondents, on average.

For each occupational sector, the inter-item correlations were appropriate for factor analysis, with KMO 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, 7-9, 11, 12, and contextual performance items 1, 2, and 5 (see Table II). 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 items 3, 4, and 6 were retained for white collar workers. For all occupational sectors, the contextual performance scale consisted of contextual performance items 7-10, 12-14, and adaptive performance items 1-8. 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 1-10. For pink collar workers, this scale was split into two factors: a minor CWB factor (Items 1-5), and a serious CWB factor (Items 6-10).

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 II), all the scales showed good model fit (Table III, 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).

Table II.
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	Occupational sector		
			M (SD)	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 three months?	"Insufficient"-"very good"	3.0 (0.7)	3.1 (0.7)[a]	3.0 (0.7)[b]	3.0 (0.7)[b]
TP2	Compared to last year, I judge the quality of my work in the past three months to be…	"Much worse"-"much better"	2.4 (0.6)	2.3 (0.5)[a]	2.4 (0.7)[b]	2.4 (0.7)[a]
TP3	How often was the quality of your work below what it should have been in the past three months?	"Never"-"often"	1.0 (0.7)	0.9 (0.7)[b]	1.1 (0.8)[b]	1.1 (0.7)[b]
TP4	How do you rate the quantity of your own work in the past three months?	"Insufficient"-"very good"	3.0 (0.9)	3.0 (0.8)[a]	3.0 (0.8)[b]	2.9 (0.9)[a]
TP5	Compared to last year, I judge the quantity of my work in the last three months to be…	"Much worse"-"much better"	2.4 (0.8)	2.4 (0.7)[a]	2.4 (0.8)[a]	2.5 (0.8)[a]
TP6	How often was the quantity of your work less than it should have been in the past three months?	"Never"-"often"	1.0 (0.9)	0.9 (0.9)[a]	1.0 (0.9)[a]	1.0 (0.9)[b]
TP7[c]	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)
TP8	I worked towards the end result of my work		3.2 (0.8)	3.3 (0.8)[b]	3.3 (0.8)[b]	3.1 (0.9)[b]
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)[a]	0.8 (1.0)[a]	0.9 (0.9)[a]

(*continued*)

Table II.

Items		Rating scale (0-4)	Total sample *M* (SD)	Occupational sector: Blue collar *M* (SD)	Pink collar *M* (SD)	White collar *M* (SD)
TP11[c]	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[c]	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)[a]	0.9 (1.0)[b]	1.4 (1.0)[b]
Dimension: contextual performance						
CP1	I was able to meet my appointments	"Seldom"-"always"	3.2 (0.7)	3.2 (0.7)[d]	3.3 (0.7)[b]	3.2 (0.7)[d]
CP2	I was able to fulfill my responsibilities		3.4 (0.7)	3.4 (0.7)[b]	3.4 (0.7) [3]	3.3 (0.7)[b]
CP3	Collaboration with others went well		3.2 (0.7)	3.2 (0.7)[a]	3.2 (0.8)[a]	3.1 (0.7)[d]
CP4	Others understood me well, when I told them something		3.0 (0.7)	3.1 (0.7)[d]	3.1 (0.7)[a]	3.0 (0.7)[b]
CP5	I understood others well, when they told me something		3.1 (0.7)	3.1 (0.7)[b]	3.1 (0.7)[b]	3.0 (0.6)[b]
CP6	Communication with others led to the desired result		2.9 (0.8)	2.9 (0.9)[d]	3.0 (0.8)[a]	2.9 (0.8)[d]
CP7	I came up with creative ideas at work		2.3 (1.0)	2.2 (1.1)[b]	2.2 (1.1)[b]	2.4 (0.9)[b]
CP8	I took the initiative when there was a problem to be solved		2.5 (1.0)	2.5 (1.0)[b]	2.5 (1.0)[b]	2.7 (0.9)[b]
CP9	I took the initiative when something had to be organized		2.1 (1.1)	2.0 (1.2)[b]	2.1 (1.1)[d]	2.2 (1.0)[d]
CP10[c]	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)[a]	1.6 (1.1)[a]	1.9 (1.1)
CP12	I was open to criticism of my work		2.6 (1.0)	2.4 (1.2)[d]	2.6 (1.1)[b]	2.6 (1.0)[d]
CP13	I tried to learn from the feedback I got from others on my work		2.4 (1.1)	2.4 (1.1)[d]	2.6 (1.0)[b]	2.7 (1.0)[d]
CP14[c]	I took on challenging work tasks, when available		2.4 (1.1)	2.4 (1.1)	2.4 (1.1)	2.5 (1.0)

(*continued*)

Table II.

Items		Rating scale (0-4)	Total sample	Occupational sector		
			M (SD)	Blue collar *M* (SD)	Pink collar *M* (SD)	White collar *M* (SD)
CP15	I think customers/clients/patients were satisfied with my work		3.1 (0.6)	3.2 (0.6)[a]	3.2 (0.6)[a]	3.0 (0.6)[b]
CP16	I took into account the wishes of the customer/client/patient in my work		3.4 (0.7)	3.4 (0.7)[a]	3.5 (0.7)[a]	3.1 (0.8)[a]
Dimension: adaptive performance						
AP1[c]	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[c]	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)	3.1 (0.9)[d]	3.2 (0.8)[b]	3.0 (0.8)[b]
AP4	I was able to cope well with difficult situations and setbacks at work		2.6 (1.0)	2.7 (1.0)[b]	2.6 (1.0)[b]	2.5 (0.9)[d]
AP5	I recovered fast, after difficult situations or setbacks at work		2.7 (0.9)	2.8 (1.0)[d]	2.7 (0.9)[b]	2.7 (0.9)[d]
AP6[c]	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)
AP7	I was able to cope well with uncertain and unpredictable situations at work		2.6 (0.9)	2.6 (1.0)[b]	2.7 (0.9)[d]	2.6 (0.9)[b]
AP8	I easily adjusted to changes in my work		2.8 (0.9)	2.9 (1.0)[b]	2.7 (0.9)[d]	2.8 (0.9)[b]
Dimension: counterproductive work behavior						
CWB1[c]	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[c]	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[c]	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)

(*continued*)

Items	Rating scale (0-4)	Total sample	Occupational sector		
			Blue collar	Pink collar	White collar
		M (SD)	*M* (SD)	*M* (SD)	*M* (SD)
CWB4[c]	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)
CWB5[c]	I spoke with people from outside the organization about the negative aspects of my work	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)[b]	0.2 (0.6)[b]	0.2 (0.6)[b]
CWB7	I purposely left my work so that someone else had to finish it	0.1 (0.5)	0.2 (0.5)[b]	0.1 (0.4)[b]	0.1 (0.5)[b]
CWB8	I behaved rudely towards someone at work	0.3 (0.6)	0.3 (0.6)[b]	0.2 (0.5)[b]	0.3 (0.6)[b]
CWB9	I quarrelled with my colleagues, manager, or customers	0.3 (0.6)	0.3 (0.6)[b]	0.3 (0.6)[b]	0.3 (0.6)[b]
CWB10	I purposely made mistakes	0.1 (0.3)	0.1 (0.3)[b]	0.0 (0.3)[b]	0.1 (0.3)[b]

Notes: [a]Items removed from the scale based on factor analysis; [b]items removed from the scale to improve model fit; [c]items that were included in the generic scales; [d]items removed from the scale because it was job-specific

Table II.

Analysis number, description		Item fit residual (mean ± SD)	Person fit residual (mean ± SD)	Item-trait total χ^2 χ^2 (df)	p	PSI
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
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

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

For all occupational sectors, the counterproductive scale included displaying excessive negativity (CWB1-3), 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.

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 II). 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 (CWB1-3), 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 χ^2-fit statistics. However, when setting the sample size at 200 (Andrich and Styles, 2009), the χ^2-fit statistics became non-significant, indicating good model fit (analyses 14-16). Additionally, when testing the generic scales separately per occupational sector, the χ^2-fit statistics indicated good model fit (analyses not shown). This indicated that the previously significant χ^2-fit statistic was caused by the power of the test, and that the data do 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 (Streiner and Norman, 2008; Tennant and Conaghan, 2007).

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

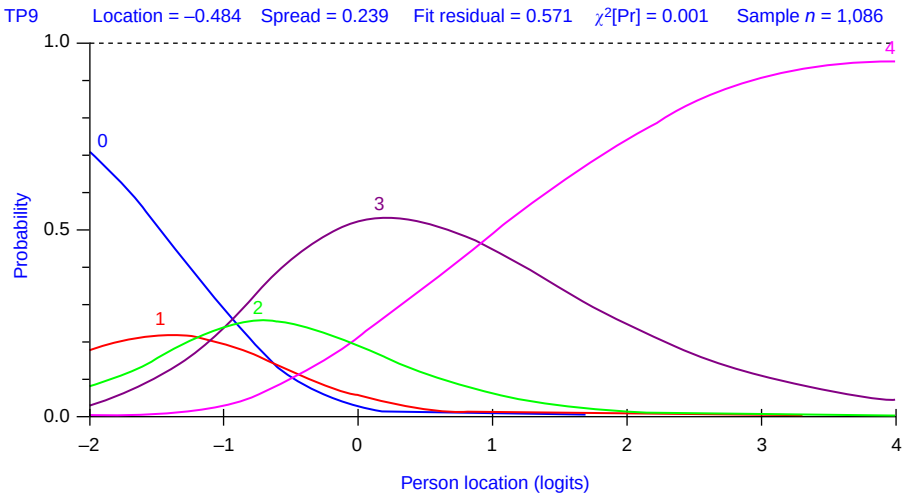

Figure 1.
Category probability curve showing disordered thresholds for result-oriented working (TP9)

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 (Westers and Kelderman, 1991). However, DIF tests are sensitive, and DIFs found in large samples may be statistically significant, but of little practical relevance (De Vet *et al.*, 2011). 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 correlations0.40. 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

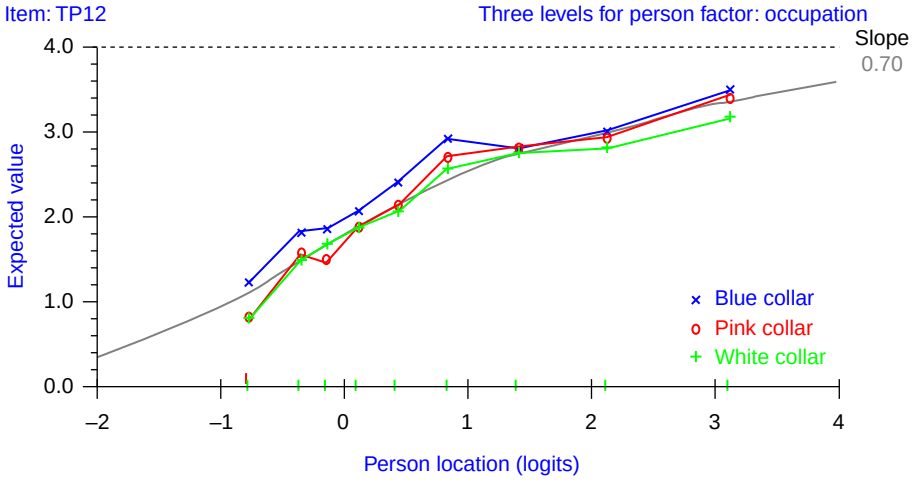

Figure 2.
Item characteristic curves showing DIF between occupational sectors for working efficiently (TP12)

dependencies. This can be illustrated by the following example: in a two-item scale, a sum score 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 (RUMM Laboratory, 2011). 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 (RUMM Laboratory, 2011). 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.

Discussion

Conceptual framework

The IWPQ 0.1 was based on a four-dimensional conceptual framework (Koopmans *et al.*, 2011). 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 CWB. Although several studies have argued for adaptive performance as a separate dimension of IWP (e.g. Pulakos *et al.*, 2000; Sinclair and Tucker, 2006; Griffin *et al.*, 2007), 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 (Koopmans *et al.*, 2011), 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.

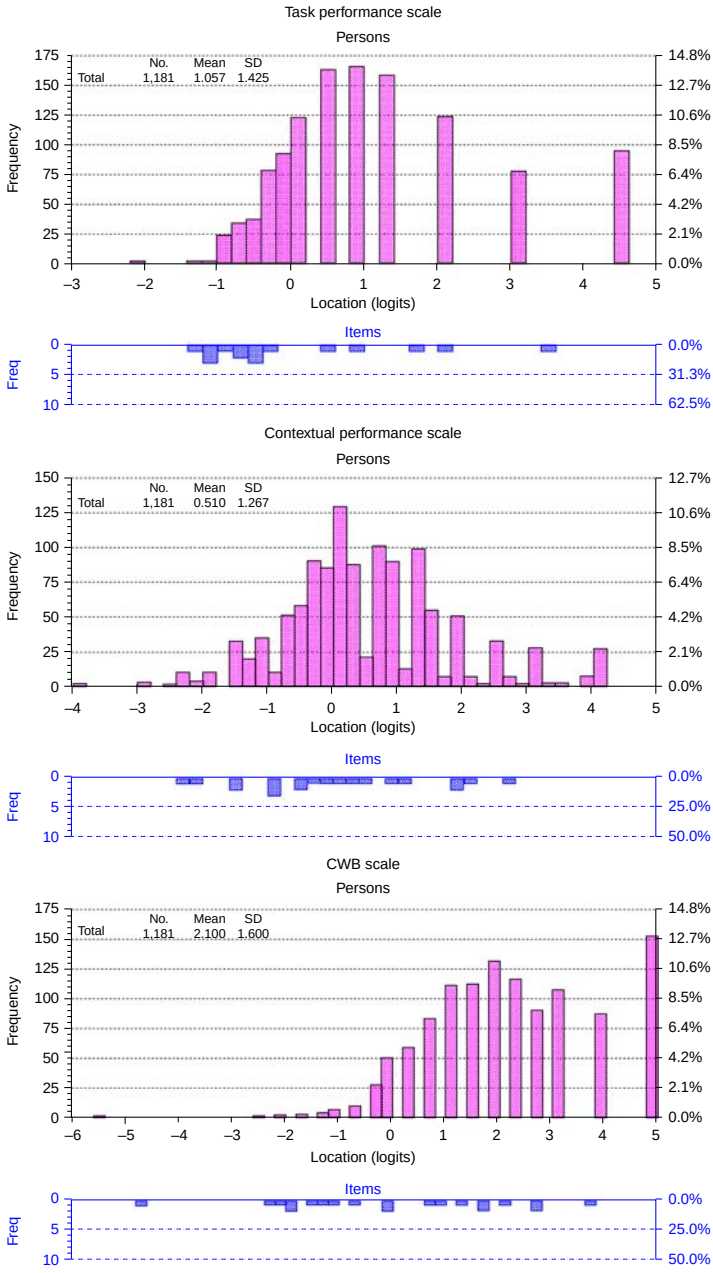

Figure 3.
The person-item threshold maps showing the distribution of persons and items for the generic task performance, contextual performance, and CWB scales

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. Rotundo and Sackett, 2002; Koopmans *et al.*, submitted). 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 focussed 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 IWPQs 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) (Penney *et al.*, 2011).

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 (De Vet *et al.*, 2011), 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.* (2005) 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 0.35 (Harris and Schaubroeck, 1988) and 0.19 (Jaramillo *et al.*, 2005). Second, self-ratings are known to show leniency effects (Van der Heijden and Nijhof, 2004). 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 (Van der Heijden and Nijhof, 2004).

Nevertheless, self-report scales were chosen for several reasons. First, in many occupations, objective measures of performance are not easily obtainable (Jaramillo *et al.*, 2005). 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 (Van der Heijden and Nijhof, 2004). 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 (Dalal, 2005). A recent study by Berry *et al.* (2012) found that self-reports of CWB 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 (Dalal, 2005; Viswesvaran *et al.*, 2005). 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 (Schoorman and Mayer, 2008).

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 (Koopmans *et al.*, 2011, submitted). 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 (De Vet *et al.*, 2011).

Another strength of the present study is that it is the first to develop or evaluate an IWPQ 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 (Tennant *et al.*, 2004). 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 (Tennant *et al.*, 2004).

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 (Thornton, 1980; Spector *et al.*, 2010). 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 among 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 (De Vet *et al.*, 2011). This will show whether the 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 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 toward establishing the predictors and effects of IWP even more accurately and completely.

References

Andrich, D. and Styles, I.M. (2009), "Distractors with information in multiple choice items: a rationale based on the Rasch model", in Smith, E. and Stone, G. (Eds), *Criterion Referenced Testing: Using Rasch Measurement Models*, JAM Press, Maple Grove, MN, pp. 24-70.

Andrich, D., Lyne, A., Sheridan, B. and Luo, G. (2003), *RUMM 2020*, RUMM Laboratory, Perth.

Austin, J.T. and Villanova, P. (1992), "The criterion problem: 1917-1992", *Journal of Applied Psychology*, Vol. 77 No. 6, pp. 836-74.

Bennett, R.J. and Robinson, S.L. (2000), "Development of a measure of workplace deviance", *Journal of Applied Psychology*, Vol. 85 No. 3, pp. 349-60.

Berry, C.M., Carpenter, N.C. and Barratt, C.L. (2012), "Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison", *Journal of Applied Psychology*, Vol. 97 No. 3, pp. 613-36.

Borman, W.C. and Motowidlo, S.J. (1993), "Expanding the criterion domain to include elements of contextual performance", in Schmitt, N. and Borman, W.C. (Eds), *Personnel Selection in Organizations*, Jossey Bass, San Francisco, CA, pp. 71-98.

Campbell, J.P. (1990), "Modeling the performance prediction problem in industrial and organizational psychology", in Dunnette, M.D. and Hough, L.M. (Eds), *Handbook of Industrial and Organizational Psychology*, Consulting Psychologists Press, Palo Alto, CA, pp. 687-732.

Cronbach, I.J. (1951), "Coefficient alpha and the internal structure of tests", *Psychometrika*, Vol. 16 No. 3, pp. 297-333.

Dalal, R.S. (2005), "A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior", *Journal of Applied Psychology*, Vol. 90 No. 6, pp. 1241-55.

De Vet, H.C.W., Terwee, C.B., Mokkink, L.B. and Knol, D.L. (2011), *Measurement in Medicine*, Cambridge University Press, New York, NY.

Griffin, M.A., Neal, A. and Parker, S.K. (2007), "A new model of work role performance: positive behavior in uncertain and interdependent contexts", *Academy of Management Journal*, Vol. 50 No. 2, pp. 327-47.

Harris, M.M. and Schaubroeck, J. (1988), "A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings", *Personnel Psychology*, Vol. 41 No. 1, pp. 43-62.

Jaramillo, F., Carrillat, F.A. and Locander, W.B. (2005), "A meta-analytic comparison of managerial ratings and self-evaluations", *Journal of Personal Selling & Sales Management*, Vol. XXV No. 4, pp. 315-28.

Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., De Vet, H.C.W. and Van der Beek, A.J. (submitted), "Measuring individual work performance – identifying and selecting indicators".

Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., Schaufeli, W.B., De Vet, H.C.W. and Van der Beek, A.J. (2011), "Conceptual frameworks of individual work performance – a systematic review", *Journal of Occupational and Environmental Medicine*, Vol. 53 No. 8, pp. 856-66.

Lundgren Nilsson, A. and Tennant, A. (2011), "Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited", *Journal of Rehabilitation Medicine*, Vol. 43 No. 10, pp. 884-91.

Penney, L.M., David, E. and Witt, L.A. (2011), "A review of personality and performance: identifying boundaries, contingencies, and future research directions", *Human Resource Management Review*, Vol. 21 No. 4, pp. 297-310.

Podsakoff, P.M. and MacKenzie, S.B. (1989), *A Second Generation Measure of Organizational Citizenship Behavior*, Indiana University, Bloomington, IN.

Pulakos, E.D., Arad, S., Donovan, M.A. and Plamondon, K.E. (2000), "Adaptability in the workplace: development of a taxonomy of adaptive performance", *Journal of Applied Psychology*, Vol. 85 No. 4, pp. 612-24.

Rasch, G. (1960), *Probabalistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press, Chicago, IL.

Rotundo, M. and Sackett, P.R. (2002), "The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: a policy-capturing approach", *Journal of Applied Psychology*, Vol. 87 No. 1, pp. 66-80.

RUMM Laboratory (2011), "Factor analysis and negative PCA values", available at: www.rummlab.com.au/faq12.html (accessed March 15, 2012).

Schoorman, D.F. and Mayer, R.C. (2008), "The value of common perspectives in self-reported appraisals: you get what you ask for", *Organizational Research Methods*, Vol. 11 No. 1, pp. 148-59.

Schwarz, N. and Oyserman, D. (2001), "Asking questions about behavior: cognition, communication, and questionnaire construction", *American Journal of Evaluation*, Vol. 22 No. 2, pp. 127-60.

Sinclair, R.R. and Tucker, J.S. (2006), "Stress-care: an integrated model of individual differences in soldier performance under stress", in Britt, T.W., Castro, C.A. and Adler, A.B. (Eds), *Military Life: The Psychology of Serving in Peace and Combat (Vol. 1): Military Performance*, Praeger Security International, Westport, CT, pp. 202-31.

Spector, P.E., Bauer, J.A. and Fox, S. (2010), "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*, Vol. 97 No. 4, pp. 781-90.

Spector, P.E., Fox, S., Penney, L.M., Bruursema, K., Goh, A. and Kessler, S. (2006), "The dimensionality of counterproductivity: are all counterproductive behaviors created equal?", *Journal of Vocational Behavior*, Vol. 68 No. 3, pp. 446-60.

Streiner, D.L. and Norman, G.R. (2008), *Health Measurement Scales: A Practical Guide to their Development*, 4th ed., Oxford University Press, New York, NY.

Tennant, A. and Conaghan, P.G. (2007), "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)*, Vol. 57 No. 8, pp. 1358-62.

Tennant, A., McKenna, S.P. and Hagell, P. (2004), "Application of Rasch analysis in the development and application of quality of life instruments", *Value in Health*, Vol. 7 No. S1, pp. S22-S26.

Thornton, G.C. (1980), "Psychometric properties of self-appraisals of job performance", *Personnel Psychology*, Vol. 33 No. 2, pp. 263-71.

Traub, R.E. (1983), "A priori considerations in choosing an item response model", in Hambleton, R.K. (Ed.), *Applications of Item Response Theory*, Educational Research Institute of British Columbia, Vancouver, BC, pp. 57-70.

Van der Heijden, B.I.J.M. and Nijhof, A.H.J. (2004), "The value of subjectivity: problems and prospects for 36-degree appraisal systems", *The International Journal of Human Resource Management*, Vol. 15 No. 3, pp. 493-511.

Van Scotter, J.R. and Motowidlo, S.J. (1996), "Interpersonal facilitation and job dedication as separate facets of contextual performance", *Journal of Applied Psychology*, Vol. 815, pp. 525-31.

Viswesvaran, C. and Ones, D.S. (2000), "Perspectives on models of job performance", *International Journal of Selection and Assessment*, Vol. 8 No. 4, pp. 216-26.

Viswesvaran, C., Schmidt, F.L. and Ones, D.S. (2005), "Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences", *Journal of Applied Psychology*, Vol. 90 No. 1, pp. 108-31.

Westers, P. and Kelderman, H. (1991), "Examining differential item functioning due to item difficulty and alternate attractiveness", *Psychometrika*, Vol. 57 No. 1, pp. 107-18.

Williams, L.J. and Anderson, S.E. (1991), "Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors", *Journal of Management*, Vol. 17 No. 3, pp. 601-17.

Corresponding author
Linda Koopmans can be contacted at: linda.koopmans@tno.nl