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].

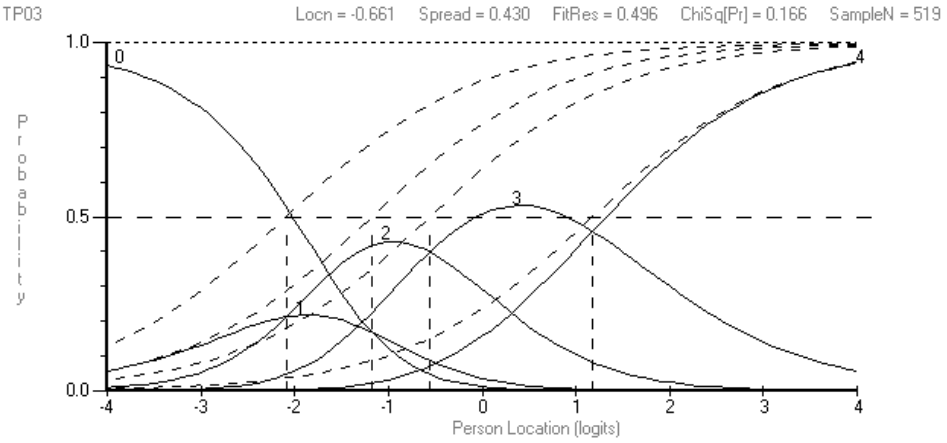

Figure 1. The category probability curves showing disordered thresholds for task performance item 3, for pink collar workers. The latent dichotomous responses (dotted lines) represent the observed responses for each answer category. The category characteristic curves (solid lines) represent the probability 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.

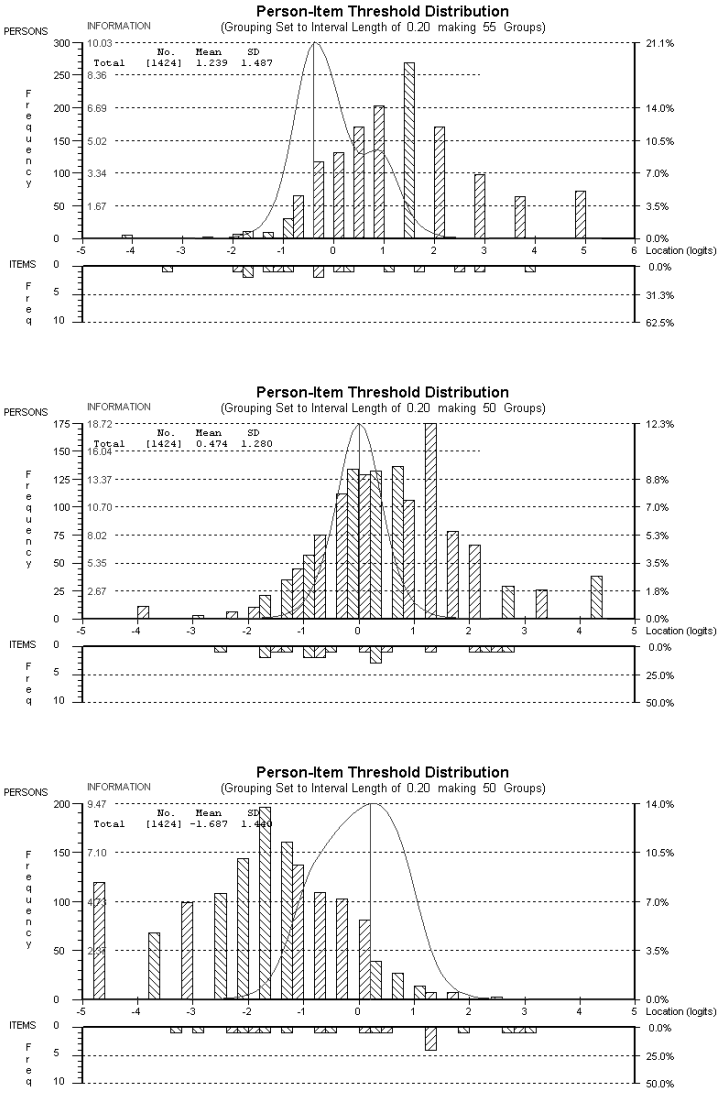

Figure 2. From top to bottom: person-item threshold maps representing the targeting of the IWPQ 0.2 task performance, contextual performance, and counterproductive work behavior scale, respectively. The top distribution in each map shows the persons, and the bottom distribution shows the item thresholds.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

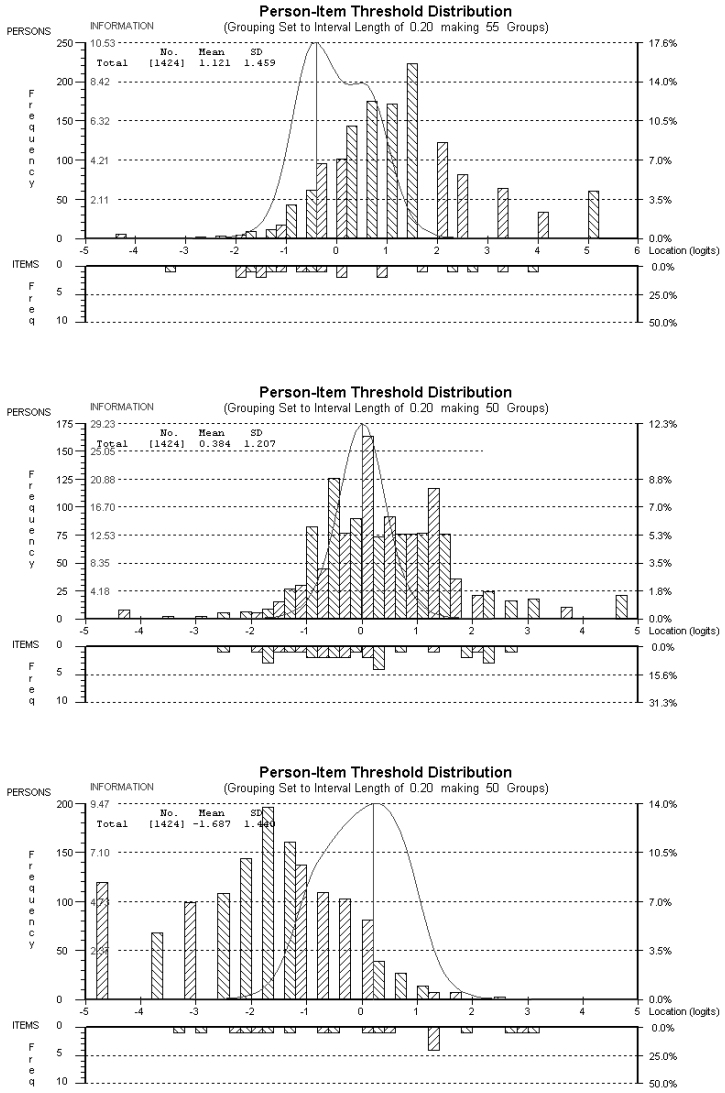

Figure 3. From top to bottom: person-item threshold maps representing the targeting of the IWPQ 1.0 task performance, contextual performance, and counterproductive work behavior scale, respectively. The top distribution in each map shows the persons, and the bottom distribution shows the item thresholds. 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.