JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 11(2), 2022, 153-167
DOI: http://dx.doi.org/10.15408/jp3i.v11i2.24157
http://journal.uinjkt.ac.id/index.php/jp3i

Rasch Analysis of The Indonesian Version of Individual Work Performance Questionnaire (IWPQ)

Weepy Grace Dwiliesanti, Ananta Yudiarso

Faculty of Psychology, Universitas Surabaya, Indonesia

weepy.grace.d@gmail.com

Abstract

The Individual Work Performance Questionnaire was developed by Koopmans et al. (2013). This questionnaire was based on the construct of individual work performance which consists of task performance, contextual performance, and counterproductive work behavior. Widyastuti & Hidayat (2018) adapted the IWPQ into Bahasa Indonesia. The mentioned research used the classical test theory (CTT) approach to validate the instrument. Therefore, the findings were only applicable to the study's sample, as validity and reliability could not be legitimately generalized to other study settings. In comparison, the development of the original IWPQ used Rasch analysis to examine its measurement properties. Rasch analysis is a modern psychometric approach based on item response theory (IRT), which has several advantages over the CTT. This study aimed to validate the psychometric properties of the Indonesian Version of IWPQ using the Rasch model. The psychometric properties discussed in this study include instrument reliability, person and item reliability, unidimensionality, rating scale functioning, and bias detection (Differential Item Functioning). The 213 participants in this research survey were Indonesian citizens aged 18-46 years old (mean = 30.64, SD = 8.55) and were actively working for at least three months at their current job. The result showed that the assumption of the unidimensionality of each sub-scale of IWPQ was fulfilled. The 5-Likert rating scales of this instrument had adequate functionality. The person reliability for all sub-scales ranged from .58 - .80. Meanwhile, the item reliability ranged from .90 - .97. The separations were considered high with a value ranging from 3.04 – 5.77. All items in this instrument functioned well to measure individual work performance except for one item in sub-scale Contextual Performance. This specific item should be revised to achieve a more accurate measurement of the construct. There was one item that was considered biased toward gender in sub-scale Contextual Performance. Also, there was one item that was considered biased toward tenure in sub-scale Counterproductive Work Behavior. These findings had implications for using the Indonesian Version of IWPQ to assess employees' individual work performance and recommendations for future research.

Keywords: individual work performance, individual work performance questionnaire, rasch analysis, work performance

Abstrak

Individual Work Performance Questionnaire (IWPQ) *dikembangkan oleh Koopmans et al. (2013). Kuesioner ini merujuk pada konstruk prestasi kerja individu yang terdiri dari kinerja tugas, kinerja kontekstual, dan perilaku kerja kontraproduktif. Widyastuti & Hidayat (2018) mengadaptasi* IWPQ *ke dalam Bahasa Indonesia. Penelitian tersebut menggunakan pendekatan* classical test theory (CTT) *untuk memvalidasi instrumen. Oleh karena itu, temuan ini hanya dapat diterapkan pada sampel penelitian, karena validitas dan reliabilitas tidak dapat digeneralisasikan secara sah ke latar studi lain. Sebagai perbandingan, pengembangan* IWPQ *asli menggunakan analisis Rasch untuk memeriksa properti pengukurannya. Analisis Rasch adalah pendekatan psikometrik modern berdasarkan* item response theory (IRT), *yang memiliki beberapa keunggulan dibandingkan* CTT. *Penelitian ini*

bertujuan untuk memvalidasi properti psikometri IWPQ Versi Indonesia dengan menggunakan model Rasch. *Sifat psikometrik yang dibahas dalam penelitian ini meliputi* instrument reliability, person and item reliability, unidimensionality, rating scale functioning, *dan* bias detection (Differential Item Functioning). *213 partisipan dalam survei penelitian ini adalah warga negara Indonesia berusia 18-46 tahun* (mean = *30,64*, SD = *8,55*) *dan aktif bekerja setidaknya selama tiga bulan di pekerjaan mereka saat ini. Hasil penelitian menunjukkan bahwa asumsi keunidimensian setiap subskala* IWPQ *terpenuhi. Instrumen ini berskala Likert dengan 5 kategori yang berfungsi dengan baik. Reliabilitas individu untuk semua sub-skala berkisar antara 0.58 – 0.80. Sedangkan reliabilitas soal berkisar antara 0.90 – 0.97. Pemisahan tersebut tergolong tinggi dengan nilai berkisar antara 3.04 – 5.77. Semua item dalam instrumen ini berfungsi dengan baik untuk mengukur prestasi kerja individu kecuali satu item dalam sub-skala Kinerja Kontekstual. Item khusus ini harus direvisi untuk mencapai pengukuran konstruk yang lebih akurat. Ada satu item yang dianggap bias gender dalam sub-skala Kinerja Kontekstual. Juga, ada satu item yang dianggap bias terhadap masa jabatan dalam sub-skala Perilaku Kerja Kontraproduktif. Temuan ini berimplikasi pada penggunaan* IWPQ *Versi Indonesia untuk menilai kinerja individu karyawan dan rekomendasi untuk penelitian di masa mendatang.*

Kata Kunci: analisis rasch, kuesioner prestasi kerja individu, prestasi kerja, prestasi kerja individu

Introduction

Every organization has certain goals. It uses any sources possible and gives the best effort to achieve those goals. Among so many factors, individual work performance is the basic foundation that can predict organizational achievement (Campbell & Wiernik, 2015). Even so, the concept of performance is often misunderstood or used interchangeably with the term productivity. Productivity is the result of input divided by output. It can be said that productivity is a concept that is closely related to the result, while performance is closely related to the process (Rostiana & Lie, 2019).

Individual work performance is a construct regarding the behaviors or actions of an individual that are relevant to organizational goals. According to the result of the study conducted by Koopmans et al. (2011), individual performance consists of three dimensions, including task performance, contextual performance, and counterproductive work behavior. Task performance is also known as proficiency, with which an individual performs central job tasks. It includes work quantity, work quality, and job knowledge. Contextual behavior is defined as individual behaviors that comprehensively support the organizational environment in which the technical core must function. Meanwhile, the definition of counterproductive work behavior is behavior that harms the well-being of the organization, such as being late for work, engaging in off-task behavior, and absenteeism.

A follow-up study about these constructs resulted in an instrument known as Individual Work Performance Questionnaire or IWPQ (Koopmans et al., 2013). IWPQ was developed originally in The Netherlands. To date, there have been a few studies that translate IWPQ into different languages or used it contextually in different countries such as Sweden (Dåderman et al., 2020), Spain (Ramos-Villagrasa et al., 2019) and South Africa (van der Vaart, 2021). In Indonesia, the adaptation of IWPQ was conducted by Widyastuti & Hidayat (2018).

There are several approaches to measuring psychological variables. The Classical Test Theory (CTT) is a relatively popular psychometric theory used in social science disciplines, including psychology. CTT analysis and interpretation can be carried out according to research needs with a number of properties, including descriptive statistics, difficulty level, discriminant index, total item correlation, and item weighting (Bond & Fox, 2015). The effectiveness of CTT in demonstrating the validity and reliability of measuring instruments raised a few criticisms, such as the reliability value of the CTT depended on the sample or the characteristics of the test takers, meaning that if a measuring instrument was used in group A, the reliability results might differ in group B. Reliability was not attached to the instrument or measuring instrument, but it was attached to the score or measurement of the sample. In addition, criticism has also

been raised against CTT for using raw scores in its analysis process. CTT was considered to give less accurate results because it treated raw scores on an ordinal scale in statistical, mathematical calculations, which are actually carried out with interval or ratio data scales (Alagumalai & Curtis, 2005).

Item response theory (IRT) is an approach in measurement theory whose analysis specifically explains the interaction between the person or subject of measurement with the items. One of the most popular IRT models was the Rasch model or the so-called 1-parameter logistic (PL) model. There were also other models, such as 2 PL and 3 PL. Bond & Fox (2015) noted that the Rasch model has an advantage over the other IRT models as it uses the measurement procedures of physical sciences as its reference point. Responding to one of the limitations of CTT regarding the use of raw scores in mathematical calculations, the Rasch model returns the data according to its natural condition in the form of continuum data by accommodating data transformations in logit units (Sumintono & Widhiarso, 2014). Thus, ordinal data from measuring instruments whose original scale distance is not known can be converted into interval data.

The analysis using the Rasch model produces some measurement properties. The accuracy of the item with the model often referred to as infit and outfit, is an indicator of the suitability of items in measuring tools and misconceptions. The results of the analysis that show item misfits can also be seen in the form of a map, widely known as the Wright Map. Reliability values are divided into person and item reliability. DIF or Differential Item Functioning is an indicator to determine whether or not there is an item bias in specific research subject categories, for example, between men and women or between specific age groups (Yu, 2020). Analysis with Rasch modeling through some of these properties can objectively evaluate the accuracy of the instrument in measuring specific attributes or variables.

The development of the original version of IWPQ by Koopmans et al. (2013) was carried out in several stages. Individual performance indicators were obtained through the scientific literature, and existing measurement tools and interviews with experts were used as the basis for constructing the 47 IWPQ 0.1 items. The scale was tested on 1,811 field workers, service workers, and office workers in the Netherlands. Then a factor analysis was carried out, with the results of three dimensions of individual performance. Koopmans et al. (2013) then conducted an analysis using the Rasch model to identify the accuracy of items and individuals. The Rasch analysis resulted in three dimensions, with each dimension perceived as a sub-scale aligned with individual work performance's multidimensional construct. Task performance consisted of 5 items, the contextual performance consisted of 8 items, and counterproductive work behavior consisted of 5 items. Meanwhile, the reliability results were in the range of 0.78 to 0.84.

Widyastuti & Hidayat (2018) adapted the IWPQ to the Indonesian language by testing its content validity, calculating the discriminant index of each item, and estimating the reliability of the measuring instrument using Cronbach's alpha coefficient. The discriminant index on each individual performance dimension was in the range of .447 to .747. The results of Cronbach's alpha coefficient showed good reliability above .8. Based on the classical test theory approach used in that study, the Indonesian version of the IWPQ had met the good psychometric property rules. However, this finding was only applicable to the study's sample as the validity and reliability could not be legitimately generalized to other study settings or samples.

A number of studies on individual performance in various cultural contexts have used the IWPQ as the measurement tool, either in its entirety or the sub-scales which are related to the research topics (Ceschi et al., 2017; Daraba et al., 2021; Metin et al., 2018; van der Lippe & Lippényi, 2020; Varshney & Varshney, 2020). Meanwhile, individual work performance research with Indonesian participants has also been carried out quite a lot, from its relation to personal aspects such as stress (Grasiaswaty, 2020), self-efficacy, and personality (Ramdani et al., 2021) to organizational aspects such as compensation and discipline (Prasetyo et al., 2021), and organizational culture (Srihadi et al., 2019). However, these studies did not use the Indonesian version of the IWPQ adapted by Widyastuti & Hidayat (2018) as a measurement tool.

http://journal.uinjkt.ac.id/index.php/jp3i

Moreover, we found no other studies validated the Indonesian Version of IWPQ using Rasch analysis. Therefore, it was necessary to have an individual work performance measurement tool that is rooted in the original construct and is also proven to be valid and reliable.

Based on this explanation, this study aimed to test the validity and reliability of the Indonesian version of the IWPQ using the Rasch model. If the IWPQ is proven to be valid and reliable, organizations can use it to accurately capture individual performance. Meanwhile, if the test results indicate that there is a need for improvement, then the measuring instrument can be developed to obtain a more suitable and consistent measuring instrument for workers in Indonesia. Validity and reliability tests will provide protection to the public or the scientific community from the use of measuring instruments that are less valid and reliable.

Methods

Participant

The number of samples in the Rasch model is affected by the principle of instrument calibration. When an instrument is calibrated on different samples of similar participants, slightly different results are expected. Therefore, if the sample size is too small, the calibration results will be unstable and less sensitive to describing the actual results. A large sample size will suppress the difference in the calibration result, but it should be noted that cost and time efficiency needs to be further considered. Linacre (1994) suggested that with 99% confidence level, sample size range between 108-243 is sufficient to conduct Rasch analysis. This study was participated by 213 Indonesian workers (145 female, 68 male). All participants were Indonesian citizens aged 18-46 years old (mean = 30.64, SD = 8.55) and were actively working for at least the last three months. They were categorized by the tenure at the last job, three months – 1 year (42 participants), 1 – 3 years (55 participants), 3 – 5 years (30 participants), 5-10 years (36 participants), and more than ten years (50 participants). The nonprobability sampling method was used as the sampling technique. A sampling frame as a requirement of probability sampling was not possible to establish because the size of the population that met the criteria could not be precisely determined.

Instrument

The instrument used in this research was the Indonesian version of the IWPQ adapted by Widyastuti & Hidayat (2018). We managed to grant permission from the mentioned researchers to use this instrument. The instrument consisted of three sub-scales, namely task performance (5 items), contextual performance (8 items), and counterproductive work behavior (5 items). The total number of items in this instrument were 18 items. Table 1 shows the blueprint of each subscale.

Table 1. Blueprint of IWPQ.

Sub-Scale	Item Coding	Item
Task Performance	TP1	*Saya mampu merencanakan pekerjaan sehingga dapat menyelesaikannya tepat waktu*
	TP2	*Saya terus mengingat target kerja yang harus saya capai*
	TP3	*Saya mampu menetapkan prioritas dalam pekerjaan*
	TP4	*Saya dapat menyelesaikan pekerjaan saya secara efisien*
	TP5	*Saya mampu mengatur waktu kerja dengan baik*
Contextual Performance	CP6	*Saya bernisiatif memulai tugas baru setelah tugas sebelumnya selesai*
	CP7	*Saya bersedia menjalankan tugas-tugas yang menantang yang ditawarkan kepada saya*
	CP8	*Saya berusaha memperbarui pengetahuan terkait pekerjaan saya*

	CP9	*Saya berusaha terus memperbarui keterampilan terkait pekerjaan saya*
	CP10	*Saya menemukan solusi kreatif dalam menghadapi masalah baru*
	CP11	*Saya mengambil tanggung jawab tambahan dalam bekerja*
	CP12	*Saya terus mencari tantangan baru dalam pekerjaan saya*
	CP13	*Saya berpartisipasi aktif dalam rapat atau pertemuan*
Counterproductive Work Behavior	CWB14	*Saya mengeluhkan persoalan-persoalan kecil dalam pekerjaan saya*
	CWB15	*Saya cenderung membesar-besarkan masalah di tempat kerja saya*
	CWB16	*Saya cenderung melihat sisi negatif daripada sisi positif di tempat kerja saya*
	CWB17	*Saya membicarakan hal-hal negatif dalam pekerjaan saya dengan rekan-rekan kerja*
	CWB18	*Saya membicarakan hal-hal negatif dalam pekerjaan dengan orang-orang di luar tempat kerja saya*

The item response model in this instrument was a 5-Likert rating scale, which consisted of five answer choices, namely "*jarang*", "*kadang*", "*sering*", "*sangat sering*", and "*selalu*". Respondents were asked to choose a response that was appropriate to their condition for at least the last three months. Instruments were distributed in the form of a Google Form link. The consent to be involved in the research was included in the link before data collection was carried out. Data submitted by respondents were stored in Google Drive, which can only be accessed by the researchers. Advertisements regarding the research, along with the Google Form link, were distributed through social media and were closed when the required number of samples had been met.

Data Analysis

The data analysis using the Rasch model in this study was carried out using WINSTEP® 5.1.0. version. The psychometric properties discussed in this study include instrument reliability, person and item reliability, unidimensionality, rating scale functioning, and bias detection (Differential Item Functioning). The analysis of the Rasch model was carried out on each sub-scale referring to the construct of the individual work performance' theory which was a multidimensional construct. This was in line with the research of Koopmans et al. (2013), which used Rasch analysis per sub-scale when the instrument was developed for the first time.

Results and Discussion

Unidimensionality

The analysis in this study was performed separately for each dimension of individual work performance, following the pattern of the original study (Koopmans et al., 2013). The unidimensionality of each sub-scale was determined by the result of raw variance explained by measure. This was the criteria of Rasch Principal Component Analysis of Residuals (PCAR). According to Holster & Lake (2016), a size of > 40% is sufficient evidence of unidimensionality. The eigenvalue of the first contrast should not be more than 2.0, since a smaller value indicated that the residuals were random noises, not another dimension. The result of this study showed that the Task Performance sub-scale obtained 56.2% of raw variance explained by measure (first contrast = 1.6), meanwhile the second sub-scale, Contextual Performance, had had the size of 49.3% (first contrast = 1.9). The third sub-scale, Contextual Work Behavior, showed the size of 52.9% (first contrast = 1.6). It

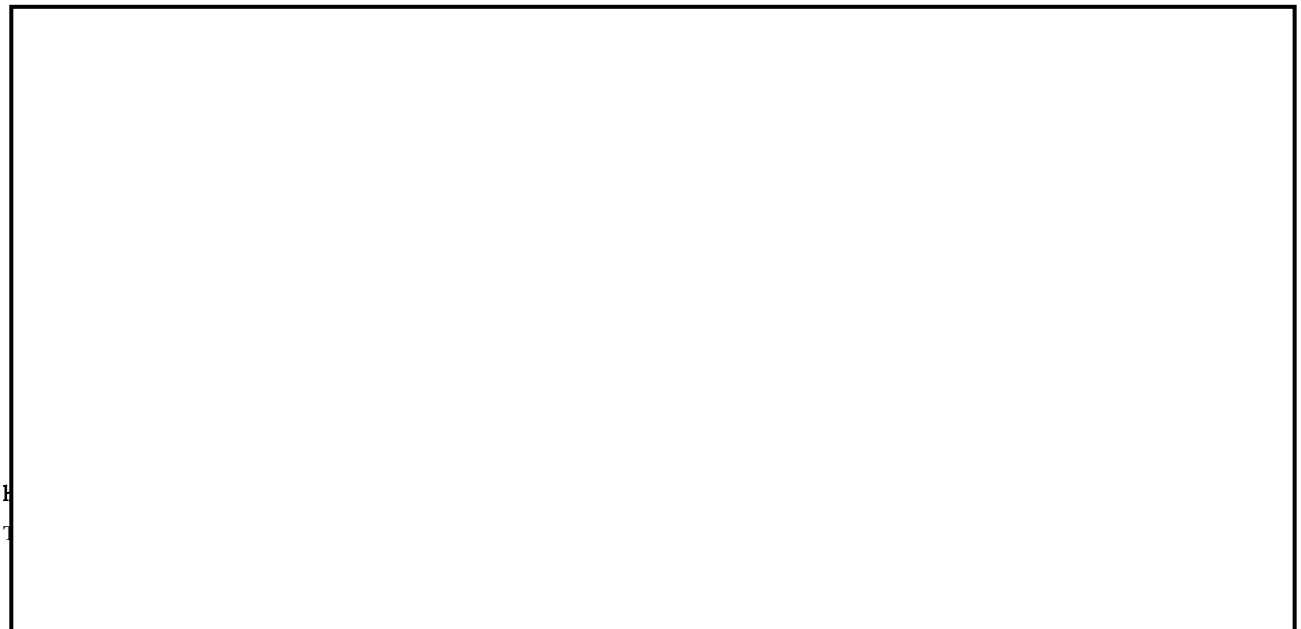
can be said that the assumption of the unidimensionality of each sub-scale of IWPQ had been fulfilled and further analysis could be done.

Rating Scale Diagnostics

IWPQ used a 5-Likert rating scale which consist of "jarang", "kadang", "sering", "sangat sering", and "selalu". A rating scale diagnostic was used to evaluate how the individuals took those choices and interpreted the distance between them. This data served more precise and interpretable measures of the construct because researchers were able to determine the actual distance applicable to respondents when choosing existing options. Each IWPQ sub-scale had a similar diagnostic result that the responses functioned as it should. This conclusion was drawn from Table 2, which shows no category or option has 0 (zero) response frequency in each sub-scale. All sub-scale also showed an increasing threshold from negative to positive in the 5 (five) responses or choices that were used (Linacre, 2012). This strongly indicates that respondents used the response category as well as they should.

Table 2. Rating Scale Diagnostics of IWPQ.

Sub-Scale	Rating Scale	f	%	Average Measure	Infit MNSQ	Outfit MNSQ	Threshold
Task Performance	1	11	1	-2.15	1.04	1.04	NONE
	2	125	12	-.59	1.23	1.12	-3.89
	3	390	37	.74	.90	.91	-1.05
	4	241	23	2.18	.85	.89	1.96
	5	298	28	3.24	1.01	1.03	2.99
Contextual Performance	1	80	5	-1.17	1.05	1.01	NONE
	2	366	21	-.36	1.11	1.14	-2.28
	3	546	32	.31	.84	.81	-.42
	4	371	22	1.10	.85	.81	1.09
	5	341	20	1.85	1.11	1.15	1.61
Counterproductive Work Behavior	1	648	61	-3.69	1.12	1.01	NONE
	2	285	27	-2.02	.97	.76	-2.54
	3	91	9	-.69	.94	.89	-.25
	4	25	2	.16	1.25	1.29	1.11
	5	15	2	1.08	1.51	1.63	1.69



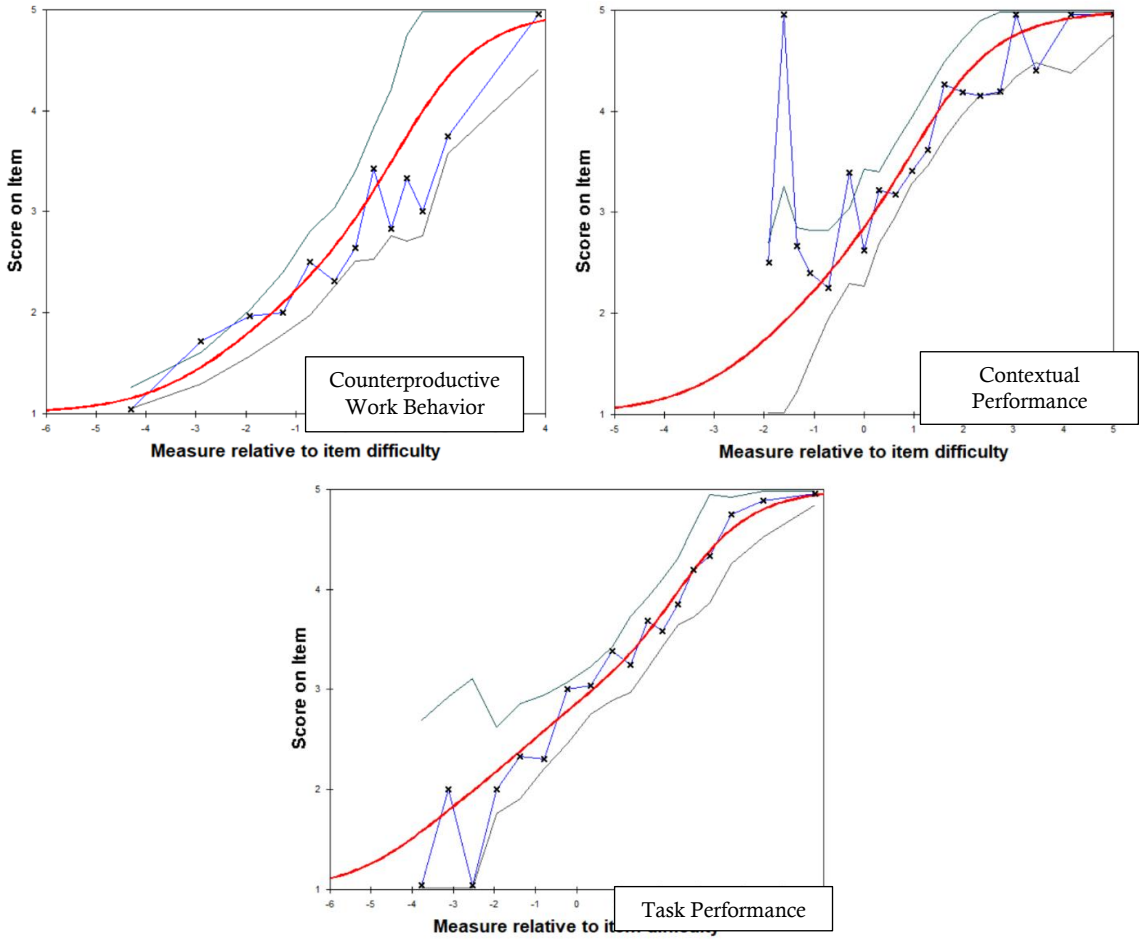


Figure 1. Item Characteristics Curve of IWPQ.

Reliability

The Rasch model estimated the reliability of either the person or the item. The instrument's capability to distinguish respondents regarding the measured variable was called person reliability. Person and item separation reliability (PSR and ISR) were interpreted in the same manner as Cronbach's α, with a minimum value of around 0.80 to be considered reliable. The person separation index (PSI) also showed reliability, although it used the logit scale instead of the raw scores. PSI should be above 3.0 to be considered as high (Linacre, 2012). The result of this study can be seen in Table 3, where person reliability for all sub-scales ranged from .58 - .80. This indicated that subscale 3, Counterproductive Work Behavior, was only fair for distinguishing the person on the measured construct. It means that this sub-scale might not be sensitive enough to distinguish between high and low performers (Bond & Fox, 2015). However, the item reliability ranged from .90 - .97, supported by Cronbach's α that ranged from .82 - .86. The separation was considered high with a value ranging from 3.04 – 5.77. Those findings indicated that the reliability of IWPQ was considered high and that it had good psychometric characteristics, with a particular note for the value of PSR.

Table 3. Instrument Reliability.

	Sub-Scale 1: TP	**Sub-Scale 2: CP**	**Sub-Scale 3: CWB**
N	5	8	5
Person Separation Reliability (>.8)	.79	.80	.58

http://journal.uinjkt.ac.id/index.php/jp3i

Item Separation Reliability (>.8)	.90	.95	.97
Alpha Cronbach (>.8)	.86	.85	.82
Person Separation Index (>2.0)	3.04	4.29	5.77

Item Fit

To determine how well each item measures the construct, the Rasch model tested the item infit, outfit, and point measure correlation. Table 4 shows the result that had been sorted from the items that were difficult to the easier ones. The infit and outfit MNSQ should range between .5 – 1.5 to be considered effective for a measurement. Meanwhile, the point measure correlation should range between .4 – .85 (Fisher, 2007). This study found that the only misfit item in all three sub-scales of IWPQ was item CP6 in sub-scale 2 "*Saya bernisiatif memulai tugas baru setelah tugas sebelumnya selesai*". This item was considered not fit to measure the contextual performance (infit MNSQ = 1.66, outfit MNSQ = 1.69). The point measure correlation value for all three sub-scales was positively correlated and passed the criteria (TP ranged between .75 - .83, CP ranged between .49 - .76, and CWB ranged between .65 - .77). These findings suggest that all items in this instrument function well to measure individual work performance except for item CP6. This specific item should be revised to achieve a more accurate measurement of the construct.

Table 4. Item Calibration of Sub-Scale 1: Task Performance.

Item	Measure	Infit MNSQ	Outfit MNSQ	Point Measure Correlation
Sub-scale 1 Task Performance				
TP2	-.50	1.19	1.21	.75
TP1	.05	1.02	.98	.81
TP3	-.33	1.00	1.02	.78
TP4	.39	.92	.91	.83
TP5	.40	.83	.81	.82
Sub-scale 2 Contextual Performance				
CP6	-.37	1.66	1.69	.49
CP13	.50	1.51	1.52	.59
CP11	.26	1.10	1.11	.68
CP7	-.03	.83	.80	.75
CP12	.57	.79	.79	.76
CP10	.10	.73	.74	.74
CP9	-.50	.72	.67	.75
CP8	-.53	.72	.68	.74
Sub-scale 3 Counterproductive Work Behavior				

CWB15	1.26	1.10	.77	.65
CWB18	.51	1.24	1.07	.69
CWB16	-.07	1.03	.88	.77
CWB17	-.38	1.02	.92	.77
CWB14	-1.33	1.06	1.00	.77

Wright Map

The validity of the construct could be determined by the hierarchy of items that can be observed in a Wright Map. This map showed the difficulty of the item on the right panel and the ability of the person on the left panel. On this map, the easier item is located at the bottom, the item with average difficulty is in the middle (mean, denoted by M on the right side), and the item with greater difficulty is at the top (Yu, 2020). The Wright map of each IWPQ's sub-scale can be seen in Figure 2.

It could be seen that regarding the difficulty level, all items in the sub-scale 1 Task Performance were relatively easy to moderate, shown by only a few persons located under the mean value (M). The mean value for person was 1.87 logit (Standard Deviation = 2.12), much lower than the mean value of the item, which was .00. On the sub-scale 2, Contextual Performance, the most difficult item was CP12 "*Saya terus mencari tantangan baru dalam pekerjaan saya*". The mean value for person was .62 logit (Standard Deviation = 1.25), also much lower than the mean value for an item, which was .00. Lastly, it should be noted that the sub-scale 3, Counterproductive Work Behavior, consists of 5 items with a negative connotation. It was aligned with the construct, which defined counterproductive work behavior as behavior that harms the well-being of an organization. The most difficult item on the subscale 3 was item CWB15 "*Saya cenderung membesar-besarkan masalah di tempat kerja saya*". The person distribution from Figure 2 shows that this sub-scale was relatively hard for the participants as many persons did not choose the extreme responses. The mean value for person was -3.25 logit (Standard Deviation = 1.89), meanwhile, the mean value for the item was .00.

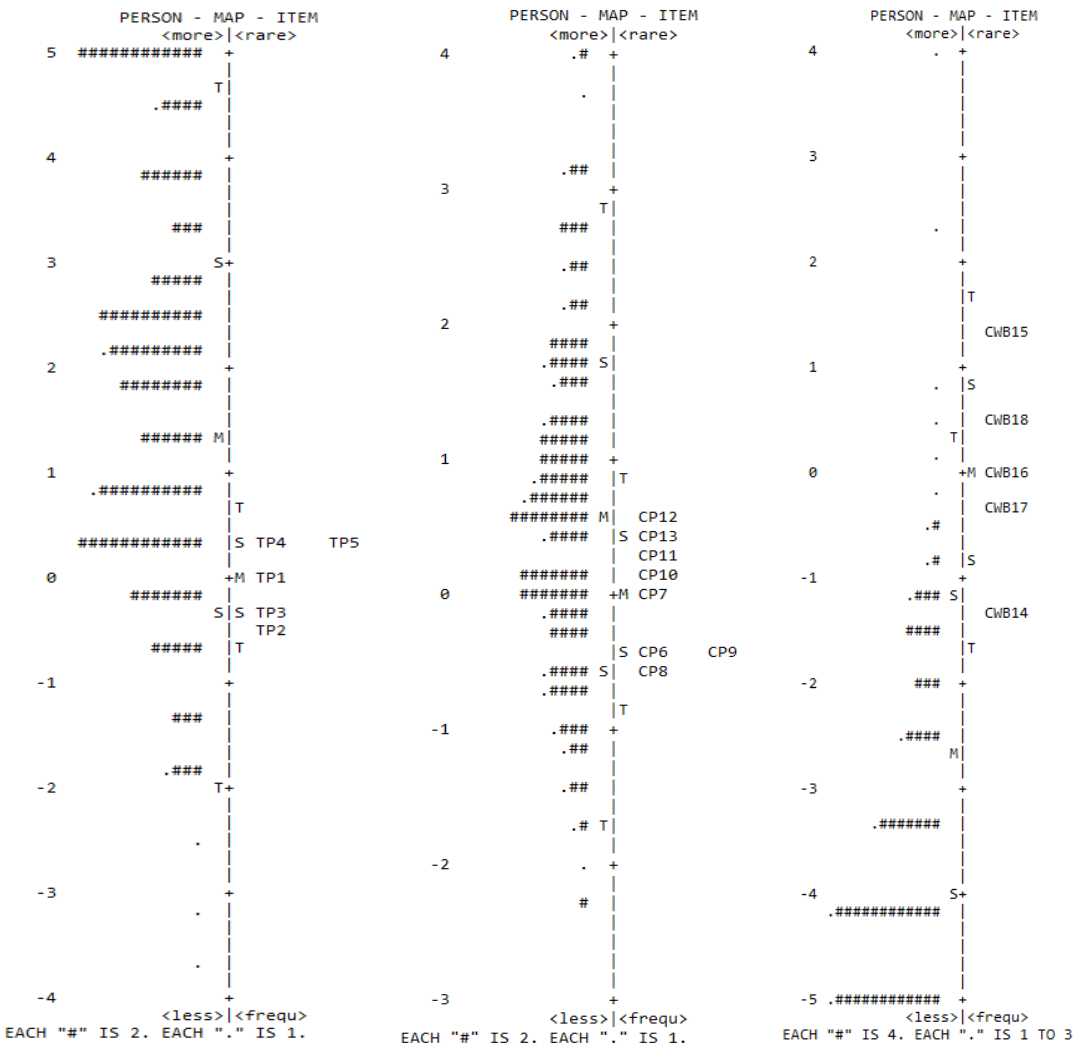


Figure 2. Wright Map of IWPQ.

Differential Item Functioning

Differential Item Functioning (DIF) analysis was used to examine whether subgroups within the sample (divided by gender and tenure) responded differently to the items, despite equal levels of the underlying characteristic being measured. In this study, gender was divided into two sub-groups, male (L) and female (P). Meanwhile, tenure was divided into five sub-groups, group A (3 months-1 years), B (1 – 3 years), C (3 – 5 years), D (5 – 10 years), and E (>10 years). The method used for evaluating DIF in this research was the item-trait chi-square (Linacre, 2007). Significant bias is detected if the probability value of the item is less than .05.

In line with item fit findings, Table 5 shows that item CP6 was considered biased toward gender ($p = .0128$). It means that there could be different interpretation between male and female in understanding item CP6 "*Saya bernisiatif memulai tugas baru setelah tugas sebelumnya selesai*". As seen in Figure 3, women tend to choose a higher rating scale than men on this item. Men were often described as achievement-oriented, whereas women were often seen as benevolent. The finding of this study was congruent with a phenomenon known as "the stereotype backlash effect", which occurs when individuals' behavior deviates from prescriptive stereotypes (Bohlmann & Zacher, 2021). Therefore, it was necessary to be careful in using item CP6 as there was a tendency for women to engage in higher proactive behavior at work. On a more positive

tone, this finding suggested that women empowerment (the popular issue associated with opposing gender stereotypes) might be related to contextual performance.

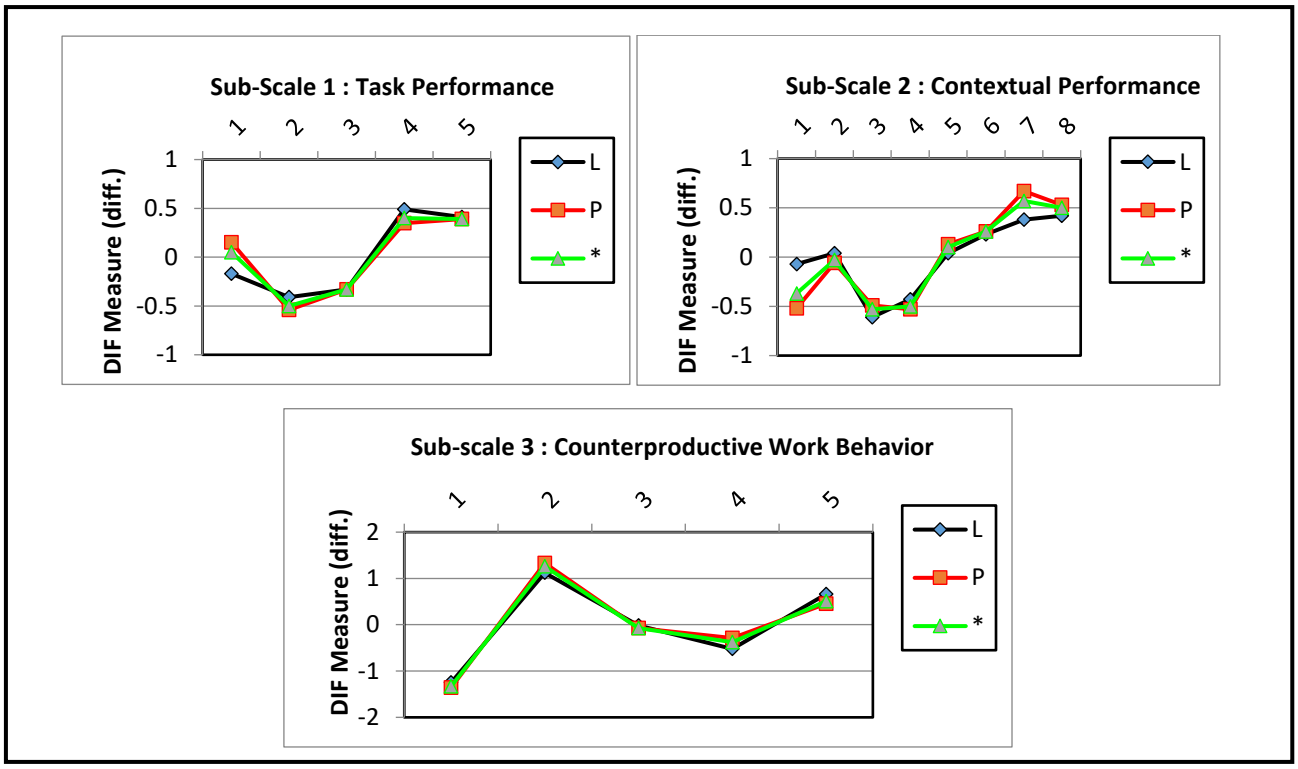


Figure 3. DIF of IWPQ Based on Gender.

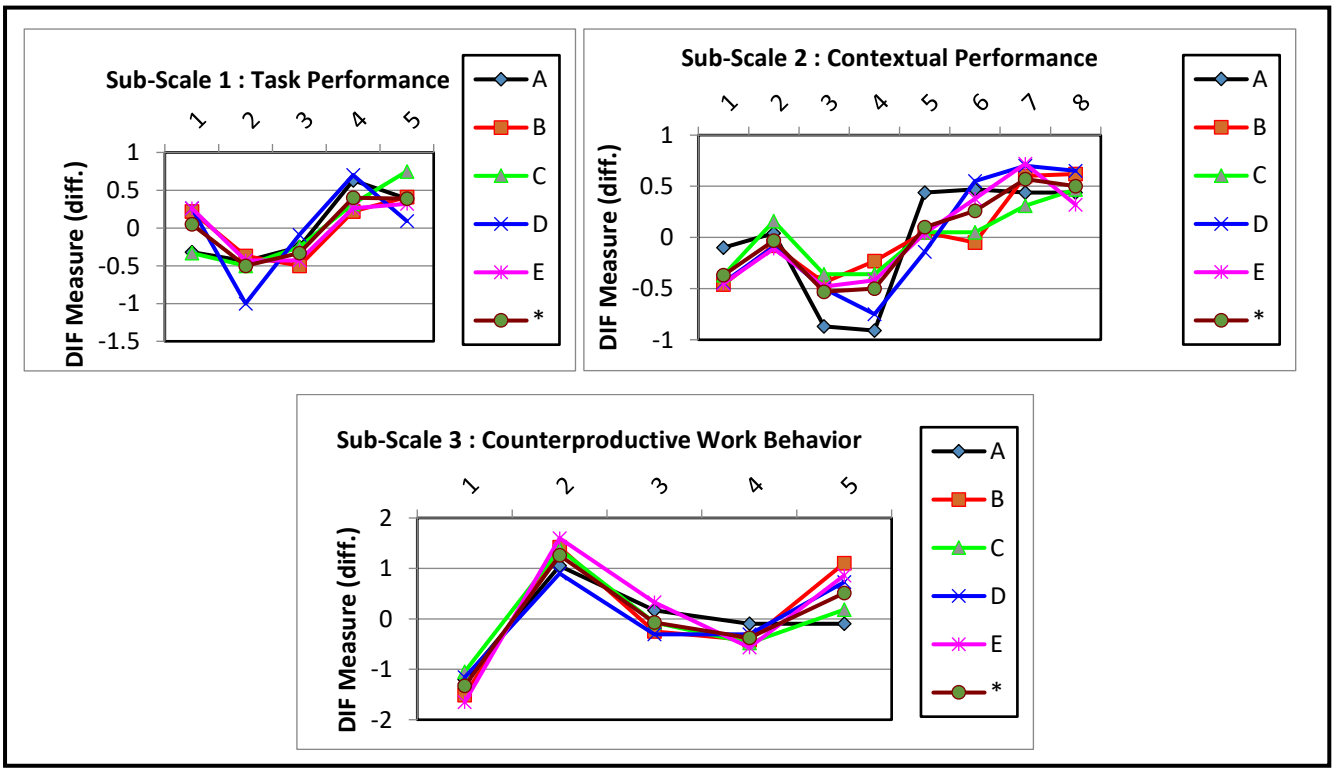


Figure 4. DIF of IWPQ Based on Tenure.

Table 5. DIF of IWPQ Based on Gender.

http://journal.uinjkt.ac.id/index.php/jp3i

Item	Item-Trait Chi-Square Probability	
	Based on Gender	**Based on Tenure**
TP1	.1813	.2147
TP2	.5778	.5602
TP3	1.0	.8034
TP4	.5670	.5782
TP5	.8950	.6419
CP6	.0128*	.6187
CP7	.5738	.8816
CP8	.5181	.4280
CP9	.5765	.0744
CP10	.6007	.3563
CP11	.8378	.0984
CP12	.1080	.6053
CP13	.5223	.6918
CWB14	.1904	.4290
CWB15	.3210	.7717
CWB16	.8428	.6206
CWB17	.3938	.8425
CWB18	.5148	.0489*

*p<0.05

Table 5 shows that bias toward tenure grouping was found in item CWB18 "*Saya membicarakan hal-hal negatif dalam pekerjaan dengan orang-orang di luar tempat kerja saya",* with p = .0489. As seen on Figure 3, individuals who were recently working on their job (group A = 3 months – 1 year) had the lowest tendency to talk about negative things about their employer. Meanwhile, individuals who had been working for 1 – 3 years (group B) had a higher tendency to bad-mouth their employer, closely followed by the group with the longest tenure (> 10 years). The second was pursuant to the previous study conducted by Ng & Feldman (2010), which concluded that organizational tenure was positively related to some counterproductive behaviors. The explanation about group B's DIF on this item might be related to "the hangover period" of employees or the decline of job satisfaction after approximately 1 year of employment. This was based on the assumption that after a "honeymoon period" when individuals started their employment, they might find some aspects of their job or organization they perceived as unsatisfactory. Workers who cannot bear the dissatisfaction leave their organization, while those who might have come to terms with or found ways to cope with it stay (Dobrow & Ganzach, 2014). This explained the decreasing tendency to bad-mouthing in the next longer tenure groups (group C & D).

Conclusion

A few conclusions could be derived from this study. The test of unidimensional assumptions showed that each sub-scale or dimension of individual work performance was unidimensional. The use of five response categories in each sub-scale appeared to be in order, increased from negative to positive which indicated that the response categories functioned as well as it should. The only misfit item in all three sub-scales of IWPQ was item CP6 in the sub-scale Contextual Performance (infit MNSQ = 1.66, outfit MNSQ = 1.69). However, the point measure correlation values for all three sub-scales were positively correlated and passed the criteria. These findings suggest that all the items in this instrument functioned well to measure the construct theory of individual work performance except for item CP6. This specific item could be revised to achieve a more accurate measurement of the construct. Regarding the difficulty

level, all items in the sub-scale 1 Task Performance were relatively easy to moderate, yet on the sub-scale 2 Contextual Performance, the most difficult item was CP12. The person distribution of the sub-scale 3 Counterproductive Work Behavior showed that this sub-scale was relatively hard for the participants as many persons did not choose the extreme responses, but it should be noted that the items of this sub-scale had negative connotations. It aligned with the construct, which defined counterproductive work behavior as behavior that harms the well-being of an organization. The most difficult item on subscale 3 was item CWB15. The next interesting finding in this study was the differential item functioning, which showed there was one item that was considered biased toward gender. Women tend to choose a higher rating scale than men on item CP6. Aligned with previous finding about misfit item, this item might need a revision to increase its accuracy in measuring contextual performance. However, it was interesting to see that this finding was congruent with a phenomenon known as "the stereotype backlash effect", which occurs when individuals' behavior deviates from a prescriptive stereotype. There might be a tendency for women to engage in higher proactive behavior at work, which further suggests that women's empowerment might be related to contextual performance. Other than that, this study also detected bias based on tenure on item CWB18. This finding implied that individuals who had just started their job had the lowest tendency to talk about negative things about their employer. Meanwhile, employees who underwent "the hangover period" or the decline of job satisfaction after approximately one year of employment had the highest tendency to bad-mouth their employer. This finding strengthens the assumption that after a "honeymoon period" when individuals start their employment, they might find some aspects of their job or organization they perceived as unsatisfactory, which leads to counterproductive work behavior in the form of bad-mouthing or negative talk.

The limitation of this study was the data collection process which might not be able to accurately represent the whole population of an employee in Indonesia because of the non-probability sampling technique. However, the Rasch analysis does not solely depend on the sampling involved, thus allowing a generalization of effective measurement properties evaluation of both three sub-scales of individual work performance construct. Even so, further research is needed to explore the validity and reliability of this instrument in more specific populations or demographic characteristics such as industrial sector and age. If such research could be conducted, it should add a more comprehensive understanding of individual work performance measurement in a different context. Overall, the Indonesian version of the Individual Work Performance Questionnaire can be applied to future research and practical application in organizations with considerations, as explained before.

References

Alagumalai, S., & Curtis, D. D. (2005). *Classical test theory bt - applied rasch measurement: A book of exemplars: papers in honour of john p. keeves* (R. Maclean, R. Watanabe, R. Baker, Boediono, Y. C. Cheng, W. Duncan, J. Keeves, Z. Mansheng, C. Power, J. S. Rajput, K. H. Thaman, S. Alagumalai, D. D. Curtis, & N. Hungi (eds.); pp. 1–14). Springer Netherlands. https://doi.org/10.1007/1-4020-3076-2_1

Bohlmann, C., & Zacher, H. (2021). Making things happen (un)expectedly: Interactive effects of age, gender, and motives on evaluations of proactive behavior. *Journal of Business and Psychology*, *36*(4), 609–631. https://doi.org/10.1007/s10869-020-09691-7

Bond, T., & Fox, C. M. (2015). *Applying the asch model: fundamental measurement in the human sciences* (third edition). Routledge. https://doi.org/https://doi.org/10.4324/9781315814698

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 47–74. https://doi.org/10.1146/annurev-orgpsych-032414-111427

Ceschi, A., Fraccaroli, F., Costantini, A., & Sartori, R. (2017). Turning bad into good: How resilience resources protect organizations from demanding work environments. *Journal of Workplace Behavioral Health*, *32*(4), 267–289. https://doi.org/10.1080/15555240.2017.1398659

Dåderman, A. M., Ingelgård, A., & Koopmans, L. (2020). Cross-cultural adaptation, from dutch to swedish language, of the individual work performance questionnaire. *Work (Reading, Mass.)*, *65*(1), 97–109. https://doi.org/10.3233/WOR-193062

Daraba, D., Wirawan, H., Salam, R., & Faisal, M. (2021). Working from home during the corona pandemic: Investigating the role of authentic leadership, psychological capital, and gender on employee performance. *Cogent Business & Management*, *8*(1), 1885573. https://doi.org/10.1080/23311975.2021.1885573

Dobrow, S., & Ganzach, Y. (2014). Job satisfaction over time: A longitudinal study of the differential roles of age and tenure. *Academy of Management Proceedings*, *2014*, 13905. https://doi.org/10.5465/AMBPP.2014.13905abstract

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095.

Grasiaswaty, N. (2020). The role of work stress on individual work performance: Study in civil servants. *Jurnal Manajemen Dan Pemasaran Jasa; Vol 13, No 1 (2020): Maret*. https://doi.org/10.25105/jmpj.v13i1.5051

Holster, T. A., & Lake, J. (2016). Guessing and the rasch model. *Language Assessment Quarterly*, *13*(2), 124–141. https://doi.org/10.1080/15434303.2016.1160096

Koopmans, L., Bernaards, C., Hildebrandt, V., van Buuren, S., van der Beek, A. J., & de Vet, H. C. W. (2013). Development of an individual work performance questionnaire. *International Journal of Productivity and Performance Management*, *62*(1), 6–28. https://doi.org/10.1108/17410401311285273

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet Henrica, C. W., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance: a systematic review. *Journal of Occupational and Environmental Medicine*, *53*(8). https://journals.lww.com/joem/Fulltext/2011/08000/Conceptual_Frameworks_of_Individual_Work.6.aspx

Linacre, J. (2007). *RUMM2020 item-trait chi-square and winsteps dif size*. Retrieved from https://www.rasch.org/rmt/rmt211k.htm

Linacre, J. (2012). *A user's guide to winstep® ministep rasch-model computer programs*. Retrieved from

Metin, U. B., Peeters, M. C. W., & Taris, T. W. (2018). Correlates of procrastination and performance at work: The role of having "good fit." *Journal of Prevention & Intervention in the Community*, *46*(3), 228–244. https://doi.org/10.1080/10852352.2018.1470187

Ng, T. W. H., & Feldman, D. C. (2010). Organizational tenure and job performance. *Journal of Management*, *36*(5), 1220–1250. https://doi.org/10.1177/0149206309359809

Prasetyo, I., Endarti, E. W., Endarto, B., Aliyyah, N., Rusdiyanto, Tjaraka, H., Kalbuana, N., & Rochman, A. S. (2021). Effect of compensation and discipline on employee performance: A Case study indonesia. *Journal of Hunan University Natural Sciences*.

Ramdani, Z., Tae, L. F., Prakoso, B. H., & Luanganggoon, N. (2021). *Personality trait, self-efficacy, and individual work performance on science teachers in indonesia bt - Proceedings of the International Conference on Educational Assessment and Policy (ICEAP 2020)*. 16–21.

https://doi.org/https://doi.org/10.2991/assehr.k.210423.058

Ramos-Villagrasa, P. J., Barrada, J. R., Fernández-Del-Río, E., & Koopmans, L. (2019). Assessing job performance using brief self-report scales: The case of the individual work performance questionnaire. *Revista de Psicologia Del Trabajo y de Las Organizaciones*. https://doi.org/10.5093/jwop2019a21

Rostiana, R., & Lie, D. (2019). Multi-dimensional individual work performance: Predictors and mediators. *GATR Global Journal of Business Social Sciences Review*. https://doi.org/10.35609/gjbssr.2019.7.1(7)

Srihadi, P., Saragih, F., & Nugroho, B. (2019). Effect of organizational culture on individual work performance and organizational performance (Study at pt. kramayudha tiga berlian motors). https://doi.org/10.4108/eai.30-7-2019.2287584

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial*. Trim Komunikata.

van der Lippe, T., & Lippényi, Z. (2020). Co-workers working from home and individual and team performance. *New Technology, Work and Employment*, *35*(1), 60–79. https://doi.org/https://doi.org/10.1111/ntwe.12153

van der Vaart, L. (2021). The performance measurement conundrum: Construct validity of the Individual Work Performance Questionnaire in South Africa . In *South African Journal of Economic and Management Sciences* (Vol. 24, pp. 1–11). scieloza .

Varshney, D., & Varshney, N. K. (2020). Workforce agility and its links to emotional intelligence and workforce performance: A study of small entrepreneurial firms in India. *Global Business and Organizational Excellence*, *39*(5), 35–45. https://doi.org/https://doi.org/10.1002/joe.22012

Widyastuti, T., & Hidayat, R. (2018). Adaptation of individual work performance questionnaire (IWPQ) into bahasa Indonesia. *International Journal of Research Studies in Psychology*. https://doi.org/10.5861/ijrsp.2018.3020

Yu, C. H. (2020). Objective measurement: How rasch modeling can simplify and enhance your assessment. In *Rasch Measurement: Applications in Quantitative Educational Research*. https://doi.org/10.1007/978-981-15-1800-3_4