Journal of Management Info (JMI)
ISSN:2313-3376
www.readersinsight.net/jmi

Validating the Development of Instrument for Measuring Nurses' Performance Scale

Suhaila Haron*[1], Aini Suzana Ariffin[2], Durrishah Idrus[3]

[1,2,3]*Perdana Centre, Universiti Teknologi Malaysia, (UTM), Kuala Lumpur, Malaysia*

** Corresponding author: suhaila@gmail.com*

Abstract

Measuring and evaluating nurses' performance are vital to identify areas for improvement in maintaining quality of service delivery and ensuring sustainability of current practices. This study attempts to examine the content validity of the nurses' performance scale. It is also aimed to achieve acceptable criteria for content validity of this instrument. Construct and content domain of nurses' performance were identified followed by items generation and instrument formation. Subsequently, assessments of content validity were performed based on content validity Index (CVI), Inter-rater agreement percentage (IRA%) and modified Kappa statistic. Two level of judgement were performed using the lay expert panel and research expert panel. Criteria were established based on these indices as basis for item reduction process. Pilot study was conducted on 50 respondents to assess the internal consistency using Cronbach's alpha value of finalized NPQ instrument. 71 items are yielded during developmental stage of instrument to measure four dimensions of nurses' performance. Assessment of content validity based on lay and research expert judgement resulting in elimination of 27 items (38%). Computed modified Kappa statistic further supplemented that the remaining 44 items as 'excellent'. As for conclusion, NPQ instrument has attain acceptable criteria of content validity assessment utilized in this study and therefore proved its potential for further research.

ARTICLE INFORMATION

Received: 10 Dec 2018
Revised: 10 Jan 2019
Accepted: 31 Jan 2019

DOI: 10.31580/jmi.v6i1.495

Keywords: *Inter Rater Reliability, Content Validity, Nurses' Performance Scale, Expert Review*

INTRODUCTION

Measuring the performance of hospital nursing care is crucial to facilitate policy makers in identifying organizational needs and subsequently determine appropriate strategies and initiatives to enhance quality of care in hospitals. Effective tools in measuring nurses' performance will enable health care stakeholders to better understand and monitor the degree to which nursing care influences patient safety and health care quality (Needleman, Kurtzman, & Kizer, 2007). Several studies have highlighted that validity of methods for measuring nurses performance needs to be better define (Mert & Ekici, 2015; Rowe, De Savigny, Lanata, & Victora, 2005). Failure to understand biasness in performance measurement tools can lead to erroneous conclusions about the adequacy of performance and may result to mismatch in selections of intervention/s to improve performance. Policy makers must assess strategy options (including single interventions and combinations) appriopriate for both short- and long-term measures (ie over 5 years), cost to implement such strategies as well as soft/hard infrastucture requirement. Subsequently, to measure the effectiveness of such strategies, policy makers must be able to identify the correct indicators and determinant to ensure continous improvement in deliveries of services in hospitals (Rowe et al., 2005). Sustaining organizational and practice changes are among key challenges in healthcare and was idenfied as main barrier for maitaining beneficial innovations over the long period of time (Fleiszer, Semenic, Ritchie, Richer, & Denis, 2015). Continuous monitoring of team dynamic is also vital in maintaining the operational sustainability in patient care and service delivery (Agarwal et al., 2012).

Performance measures and indicators are useful numerical information that quantify input, output, outcomes that are affected and/or influenced by nursing personnel (Needleman et al., 2007). Few literatures have attempted to provide universal definitions of 'nurse job performance". The World Health Organization (WHO) suggested four underlying dimensions of health worker performance namely availability, responsiveness, competence and productivity (WHO, 2006, 2010). One noted that nurses job performance is regarded as "the way nurses performed their job in serving for patients or others, and its process of servings"(Mehmet, 2013). Al-Makhaita described nurses performance as a multi-layered phenomenon with dynamic level and influencers including workload, work satisfaction, personal competencies, individual characteristics, achievement' recognitions, social support, communication and feedbacks, leadership behavior and organizational climate (Al-Makhaita, Sabra, & Hafez, 2014). Yakusheva and Weiss elaborated nurses' performance as "the capacity of an individual nurse to carry out and accomplish job"(Yakusheva & Weiss, 2017). Other defined nurses' performance as "the willingness to come to work regularly, work diligently and be flexible and willing to carry out the necessary tasks"(Dagne, Beyene, & Berhanu, 2015). It has been suggested that these ambiguous definitions of nurses' performance were resulted from multi-dimensional characteristics of individual performance concept (Sonnentag, Volmer, & Spychala, 2008; Viswesvaran & Ones, 2000) as well as various roles held by nurses between specialty and work environments (Smith, 2012). This study attempts to measure and conceptualized nurses job performance based on four underlying dimensions namely availability, responsiveness, competence and productivity as proposed by WHO.

LITERATURE REVIEW

Assessment of valid and reliable instruments has been acknowledged as vital process in studying complex construct in research (Mackenzie, Podsakoff, & Podsakoff, 2011). Pre-testing a survey questionnaire is important in order to reduce ambiguities in the question (Ali Memon, Ting, Ramayah, Chuah, & Cheah, 2017) as well as to reduce biasness caused by the instrument (Mackenzie & Podsakoff, 2012). Various quantitative measures has been established by previous scholars to assess the validity and reliability of an instrument as summarized in Table 1 (de Vet, Terwee, Knol, & Bouter, 2006; DeVon et al., 2007; Guyatt, Walter, & Norman, 1987; Lawshe, 1975; Lynn, 1986). Content validity indicated that the items were representative and relevance to the attribute under study. To ensure representativeness of an instrument, researcher should include the largest pool of potential items as possible during the early stages of instrument development, which is to be reduced based on content review by experts (Netemeyer, Bearden, & Sharma, 2003). On the other hand, items' reliability refers to the ability of an instrument to consistently measure an attribute (DeVon et al., 2007).

Table 1: Summary of indices been used to quantify both items' content validity and inter-rater reliability during pre-testing.

(See Appendix – A)

Despite the emphasize on the importance of preliminary evident of instrument validity have been highlighted in prior nursing researches (Horgas, Yoon, Nichols, & Marsiske, 2008; D. F. Polit & Beck, 2006; Rattray & Jones, 2007), these fragment were constantly reported in such superficial and transient manner (Zamanzadeh, T and Nemati, 2014). Furthermore, the use of multiple term in characterizing nurses' performance such as productivity, outcomes, effectiveness, efficiency and quality has result to poor conceptualization in measuring nurses' performance (Dubois, D'Amour, Pomey, Girard, & Brault, 2013). With this cognizant, this study aims to assess the content validity and inter-rater reliability of nurses' performance questionnaire (NPQ) using 2-stages of pre-testing process of lay-experts and research experts review. At the same time, this study attempts to achieve acceptable criteria for content validity of this instrument and to report the process.

METHODOLOGY/MATERIALS

Our approach is based on two-stage processes namely Development Stage and Content Judgement Stage (Figure 1). Development Stage consists of identification of construct and content domain, item generation and adaptation, instrument formation and items' translation to target language. Judgement Stage consist of 2-steps judgement from lay experts and research experts. Further elaboration on the research process are discussed in the section below.

Phase 1: Development Stage of Instrument

The first stage of scale development and validation comprises of defining the conceptual domain of the constructs. Construct is defined as variable that is abstract in nature in which scientist "purposely construct" or put together (from their own imaginations). This construct does not exist as an observable or concrete measure (Nunnally and Bernstein, 1994 as cited in Mackenzie et al., 2011). This stage comprises of identification of construct's intention as well as differentiation between two or more interchangeably-used-constructs (Mackenzie et al., 2011).

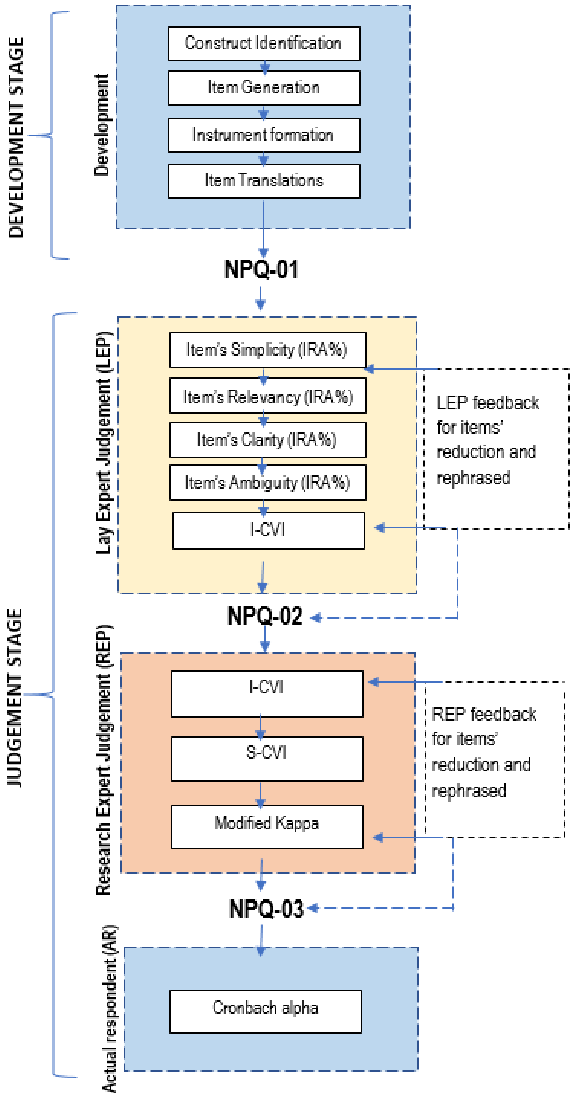

Figure 1: Steps in content validity process

Content domain can be gathered though literature review or qualitative assessment of focus group or respondent (Podsakoff, MacKenzie, & Podsakoff, 2016). In this study, content domain related to nurses' performance were gathered through literature reviews. Three main parallel streams of conceptualizing and measuring nurses' performance were identified by prior studies namely nursing activities, patient safety and quality assurance movement (Dubois et al., 2013). This study employed a four-dimensional concept of nurses' performance namely competence, responsiveness, availability and productivity as proposed by Dieleman (2006). Researcher has summarized content domain for nurses' performance and proposed conceptual definitions of these domain as in Table 2.

It is also suggested that existing construct from prior studies can also be adapted to fit the assessment's intention, the target population, reliability, validity and other considerations (Bogaert et al., 2017; Epstein, Osborne, Elsworth, Beaton, & Guillemin, 2015; Ferrer et al., 1996; Olsson, Forsberg, & Bjersa, 2016; Scott, Mannion, Davies, & Marshall, 2003). Advantages of adaptation and adoption of existing instruments include the fact that these instruments have been assessed for validity and reliability. Comparison were made between findings promotes knowledge development for instrument measurement (Kitchenham & Pfleeger, 2002).

Next, the items were translated into Bahasa Melayu version. The Bahasa Melayu version of the survey questions is more understandable to nurses as the Bahasa is commonly used in most of Malaysian public hospitals. Instrument were formed consisting bilingual version for each item (English and Bahasa Melayu) labelled as NPQ-01. NPQ-01 will be judge based on lay expert committee approach. This will provide clearer version of translated instrument as any mistake can be readily identified by the expert panels (Cha, Kim, & Erlen, 2007).

Table 2: Summary of content domain in nurses' performance
(See Appendix – B)

Phase 2: Content Judgements

In this phase, content domain and items' representativeness judgment were carried out by two different expert groups.

LAY EXPERTS JUDGEMENT

30 Lay expert panels (LEP) were invited from potential research subjects (Ali, Tretiakov, & Whiddett, 2014; McElroy & Esterhuizen, 2017). Inclusion criteria for Lay expert panels were (a) currently working as a nurse in Malaysia public hospitals; (b) having minimum of 5 years working experience in acute care; (c) being willing and able to participate (Meyer & Booker, 2001). This procedure helps researchers to identify accurate information about the potential problem regarding equivalence of translated measures and wording comprehensiveness based on level of similarities of the study conditions.(Cha et al., 2007). Lay expert panels were briefed on the research' operational definitions, construct definition and items being used to measures these constructs. Then, they were asked to rate NPQ-01 using 4-points-Likert scales based on 4 criteria of judgement (see Table 3).

Table 3: Lay expert panel judgement criteria for each item (Adapted from Yagmale,2003)

Scale	Clarity	Relevancy	Simplicity	Ambiguity
1	Not clear	Not relevant	Not simple	Doubtful
2	Item need some revision	Item need some revision	Item need some revision	Item need some revision
3	Clear but need minor revision	Relevant but need minor revision	Simple but need minor revision	No doubt but need minor revision
4	Very clear	Very relevant	Very simple	Meaning is clear

Items reliability was computed using Inter-rater agreement percentage (IRA) indices of these 4 criteria. Percentage of rater that agree with the clarity, relevancy, simplicity and ambiguity of each item (rating 3 and 4) as compared to total number of lay experts. IRA indices, exhibits the extent to which different rater assign the same precise value for each item being rated (Gisev, Bell, & Chen, 2013). The minimum IRA among lay expert is 80% (Topf, 1986).

The content validity for individual item were computed based on item-level content validity indices (I-CVI). The I-CVI expresses the proportion of agreement on the relevancy of each item, which is between 0 and 1 (D. F. Polit & Beck, 2006). I-CVI value can be obtain by dividing the number of those judging the item as relevant or clear (rating 3 or 4) to the number of lay experts. Various preposition of I-CVI value guidelines has been established by previous scholars, Sousa (2011) and Lynn (1986) proposed of >0.78, Yagmale (2003) suggested value of >0.75, Polit & Beck (2006) with value of more than 1 to be retained. However, more detail preposition of I-CVI interpretation is suggested by Zamanzadeh (2014) with I-CVI is higher than 79%, the item is considered appropriate, item with I-CVI value between 70% and 79% needs revision and item with I-CVI less than 70% should be eliminated. This study adopted I-CVI value as proposed by Zamanzadeh (2014).

Content validity of the overall scale (S-CVI) were computed by average of I-CVI for each subconstruct and referred as S-CVI/Ave. S-CVI/Ave is acknowledge as more liberal interpretation for scale-level validity index (D. F. Polit, Beck, & Owen, 2007). This step is used to support the conceptual semantic of the translated instrument via quantitative indices of IRA and I-CVI. These indices can be used as evidence of abstract concepts can be link into observable and measurable indicators (Wynd, Schmidt, & Schaefer, 2003). Furthermore, the structure of sentences used in the item can be improve based on the lay experts' suggestion for NPQ-01. Reduced questionnaire based on these criteria were labelled as NPQ-02.

RESEARCH EXPERT JUDGEMENTS

The second step for instrument judgement stage is the assessment of research experts' panel (REP). REP are professionals who have research experience or work in the field validity including academic experts. 5 research experts were invited to assess the content validity of NPQ-02 as suggested by Yahgmaie (2003). Inclusion criteria for RES were; (a) being able and willing to provide their expert opinion on the area of construct; (b) two expert have experience in fiend of content validation and questionnaire development; and (c) two experts have experience in nursing and acute healthcare management (McElroy & Esterhuizen, 2017). REP we requested to assessed NPQ-02 on the item relevancy to the construct and comprehensiveness of the questionnaire based on 4-points Likert scale. Again, CVI and S-CVI were computed.

Kappa coefficient is then computed as additional information beyond proportion of agreement by removing random chance agreement (Beckstead, 2009; Wynd et al., 2003). Kappa statistic is acknowledged as an important supplement to CVI (Zamanzadeh, T and Nemati, 2014). To calculate modified Kappa statistic, the probability of chance agreement was first calculated for each item by the following formula:

$PC = [N!/A!(N - A)!] * 0.5N$.

N = number of experts in a panel

A = number of panelists who agree that the item is relevant.

Kappa was computed by entering the numerical values of probability of chance agreement (PC) and CVI of each item (I-CVI) in the following formula:

$K = (\text{I-CVI} - PC)/(1 - PC)$

Evaluation criteria for Kappa is the values above 0.74, between 0.60 and 0.74, and the ones between 0.40 and 0.59 are considered as excellent, good, and fair, respectively.

Phase 3: Pilot Study

The pilot study were conducted on 50 nurses from the Malaysian public hospitals. Hair et al. (2006), Kumar et al. (2013), and Zikmund et al. (2013) agree that reliability is used to measure the internal consistency of the constructed questionnaire. Internal consistency describes the extent to which all the items in a test measure the same construct and hence connected the inter-relatedness of the items within test (Netemeyer et al., 2003). The Cronbach's alpha was applied in order to test the reliability of the data involved in the present study. Cronbach's alpha indices is the most frequently employed estimate of a multiple-item scale's reliability in organisational research (Cho & Kim, 2015). It is recommend that an acceptable number for Cronbach's alpha is between 0.7 to 0.95 (Nunnally & Bernstein, 1994; Tavakol & Dennick, 2011).

RESULTS AND FINDINGS

Results of Phase 1: Designing nurse performance questionnaire (NPQ-01)

Development of nurse performance questionnaire (NPQ) was performed through literature review identifying content domain within four main dimensions of performance including availability, responsiveness, competence and productivity. Each of these content domain was defined theoretically by extensive literature review as presented in Table 2. There are 58 items obtained from literatures of related instruments combined with 13 additional items are generated and proposed by the researcher (see Table 4). A total of 71 items were finalized to measure the constructs of nurse perceived performance consist of four dimensions namely availability, responsiveness, competence and productivity. Finalized instrument is labelled as NPQ-01.

Table 4: Summary of Nurse Performance Questionnaire (NPQ-01)

Dimension	Sources	No of items in NPQ-01
Availability	Lutwama, 2011 (3 items) Proposed by researcher (9 items)	12
Responsiveness	Manojlovich & Sidani, 2008 (1 item), Lutwama,2011 (1 item), Meretoja & Koponen, 2012 (6 items) Kassa et al., 2014 (2 items), proposed by researcher (2 items)	12
Competence	Nurse Competence Scale (NCS), Meretoja & Koponen, 2012	35
Productivity	North & Hughes, 2012 (3 items), Ciconelli et al., 2006 (1 item) Leach & Mayo, 2013, Kalisch, Lee, & Salas, 2012(6 items), Proposed by researcher (2 items)	12
Total		**71**

Results of Phase 2: Instrument Judgement

Lay Experts' Judgement

In the first round of judgement, 28 LEP managed to attend the session hosted by researcher. The remaining 2 did not attend due to overlapping schedule. These experts are nurses experienced in acute curative care in public hospitals with at least 5 years working experience (see section 3.2.1). The conceptual definitions of the construct, its dimensions and items established for each dimension were briefed to the experts at the beginning of the session. In the first round of judgment, the lay experts were requested to judge by scoring 1 to 4 on the relevancy, clarity, simplicity and ambiguity of instrument items according to Yaghmale (2003) for content validity index. Then, they were also asked to comment on the structure and content of sentences in each items and proposed modification of the item if needed.

Table 5: Inter-Rater Agreement (IRA) based on Lay Expert Judgement

(See Appendix – C)

Table 6: Summary for content validity index and Modified Kappa computed based on REP Judgement

Variables Name	Subconstruct	No. of items in NPQ-02	I-CVI >0.79	SCVI-Ave	Modified Kappa_AVE	No. of items in NPQ-03
Perceived performance	Availability	10	10	1	1	10
	Responsiveness	5	5	1	1	5
	Competence	22	21	0.99	1	21
	Productivity	8	8	1	1	8
		45	44 (97%)			44

71 items of NPQ-01 were tested for lay expert judgement. 24 (33.8%) items had inter-rater agreement percentage (IRA%) lower than 80% on item's clarity, relevancy, simplicity and ambiguity. Content validity index (CVI) were also computed based on item's relevancy judgement of LEP. 45 items scores item-level content validity index (I-CVI) of more than 0.79, which is considered as "appropriate" according to Zamanzadeh (2014). 16 items with I-CVI value between 0.70 to 0.78 were reviewed and compared with IRA judgement results. As a result, 24 items (33.3%) were eliminated from NPQ-01. Scale-level Content Validity Index (S-CVI) were also computed prior and after removal of the items. Table 5 shows summary of inter-rater reliability and content validity computed for each dimension of NPQ-01. After elimination and amendment of items based on LEP judgement, the newly revised instrument is labelled as NPQ-02 for the research experts panel judgement.

Research Experts' Judgement

In the second round of judgement, 5 research experts were invited to assess the relevancy of individual items to the construct (see research expert criteria in section 3.2.2). REP was provided information on the research objectives, conceptual framework, as well as operational definition for each construct. REP was asked to rate the relevancy of each items based on 4-point Likert Scale (refer Table 3) on the 45 remaining items of NPQ-02. All five experts responded on the overall judgements of the instrument. Item-level Content validity index (I-CVI) were computed for each item by dividing the number of those judging the item as relevant by the number of content experts (N=4) as one of the REP not responded to relevancy score judgement form. In this round, among the 45 instrument items, only 1 item with a CVI score lower than 0.79 were eliminated. Modification of items wording was performed according to the recommendations made by REP. 44 items (97.8%) scored I-CVI value of 1 indicated that 100% agreement among REP on the relevancy of these items to measure the construct. Modified Kappa were also computed, and average value of this index were computed for scale-level assessment. Polit et al. states that after controlling items by calculating modified Kappa statistic, each item with I-CVI equal or higher than 0.78 would be considered excellent. Table 6 shows the summary of I-CVI, S-CVI and modified kappa computed for 45 remaining items of NPQ-02.

Results of Stage 4: Pilot

Inspection of reliability of the items and constructs were done using Cronbach's alpha values. As shown in table 7, Cronbach's alpha value for the scale range of 0.701 to 0.967. Overall, Cronbach's alpha for NPQ-03 scale is 0.967. Thus, coefficient of the revised instrument satisfied the acceptable level of 0.70 (Nunnally & Bernstein, 1994; Tavakol & Dennick, 2011). Inspection of Cronbach's alpha if item deleted were also been done. In particular, deletion of item DVA1-2, DVA2-1, DVA2-2, DVA2-3, DVB1-3 and DVC6-2 will increase the alpha value of availability, responsiveness and competence construct by average of 0.33%. Provided that this value did not marginally increase the reliability value of the scale, therefore there was no statistical reason to drop these items (Cho & Kim, 2015). This also proved that further analysis of latern variable modelling procedures are required. It is highlighted that dispensing item from a scale component to maximally increase coefficient alpha may in fact entail considerable loss in criterion validity of the scale (Raykov, 2008).

Table 7: Reliability Analysis of NPQ-03

Scale	Subscales	Code	Scale Cronbach's alpha	Cronbach's Alpha Based on Standardized Items	Cronbach's Alpha if Item Deleted
Availability	Staff Ratio	DVA-1-1	.701	.703	.623
		DVA-1-2			.699
		DVA-1-3			.672
	Absent rate	DVA-2-1	.799	.803	.762
		DVA-2-2			.658
		DVA-2-3			.749
	Waiting time	DVA-3-1	.941	.941	.934
		DVA-3-2			.897
		DVA-3-3			.938

		DVA-3-5			.921
Responsiveness	Empathy	DVB-1-1	.844	.865	.762
		DVB-1-3			.762
	Receptiveness	DVB-3-1	.924	.933	.833
		DVB-3-2			.886
		DVB-3-3			.935
Competence	teaching/ Coaching	DVC-1-1	.857	.860	.774
		DVC-1-3			.742
		DVC-1-4			.867
	Diagnostic Function	DVC-2-4	.751	.757	.609
		DVC-2-5			.609
	Situation Management	DVC-3-1	.933	.935	.899
		DVC-3-2			.907
		DVC-3-3			.900
		DVC-3-5			.944
	Therapeutic intervention	DVC-4-1	.928	.928	.865
		DVC-4-3			.865
	Quality Assurance	DVC-5-3	.918	.920	.851
		DVC-5-4			.851
	Work Roles	DVC-6-2	.967	.970	.971
		DVC-6-4			.961
		DVC-6-5			.960
		DVC-6-7			.959
		DVC-6-8			.962
		DVC-6-9			.961
		DVC-6-10			.959
		DVC-6-11			.965
Productivity	Effectiveness	DVD-1-2	.849	.855	.776
		DVD-1-3			.751
		DVD-3-2			.831
	Efficiency	DVD-3-3	.759	.784	.903
		DVD-3-5			.590
		DVD-3-6			.543
	Presenteeism	DVD-2-1	.843	.843	.728
		DVD-2-2			.728

DISCUSSION AND CONCLUSION

Present paper demonstrates quantitative indices being used for content validity and reliability of Nurse Performance Questionnaire during design and development stage of the scale. These indices have evidently provided systematic criteria for items' reduction processes comprises two-step judgement process. Some limitations of content validity studies should be noted. First, experts' feedback is subjective; thus, the study is subject to bias that may exist among the experts. Secondly, quantification of content validity alone may results in collapse response category during computation of the index (Beckstead, 2009). Thus, the utilization of multiple content validity indices in this study provides multifaceted criteria for item's reduction process. Finally, limitation for NPQ may appear if content domain is not well identified. Experts were also asked to suggest other items for the instrument, which may improve the quality of each instrument. Subsequent analysis should be directed and shall include construct validity through factor analysis, reliability evaluation and criterion-related validity.

Acknowledgement:

The authors would like to extend our appreciation to the nurses and research experts who had participated in the study. The authors also thank the Ministry of Health and the Director General of Health for their support and permission to conduct this study. The protocol of this study was approved by Medical Research and Committee of National Medical Research Register (NMRR) Malaysia: NMRR-16-2831-32428(IIR).

References:

Abd Manaf, N. H., Abdullah, A. H. A., Abu Bakar, A., Ali, R., Bidin, N., Ismail, W. I. W., … Wan Ismail, W. I. (2011). "Hospital waiting time: the forgotten premise of healthcare service delivery? *International Journal of Health Care Quality Assurance*, *24*(7), 506–522. https://doi.org/10.1108/09526861111160553

Agarwal, H. S., Saville, B. R., Slayton, J. M., Donahue, B. S., Daves, S., Christian, K. G., … Harris, Z. L. (2012). Standardized postoperative handover process improves outcomes in the intensive care unit: A model for operational sustainability and improved team performance. *Critical Care Medicine*, *40*(7), 2109–2115. https://doi.org/10.1097/CCM.0b013e3182514bab

Al-Makhaita, H. M., Sabra, A. A., & Hafez, A. S. (2014). Job performance among nurses working in two different health care levels, Eastern Saudi Arabia: A comparative study. *International Journal of Medical Science and Public Health*, *3*(7), 832–837.

Ali Memon, M., Ting, H., Ramayah, T., Chuah, F., & Cheah, J.-H. (2017). Editorial - A Review of the Methodological Misconceptions and Guidelines Related to the Application of Structural Equation Modelling. *Journal of Applied Structural Equation Modeling*, *1*(1).

Ali, N., Tretiakov, A., & Whiddett, D. (2014). A Content Validity Study for a Knowledge Management Systems Success Model in Healthcare. *Jitta*, *15*(2), 21–36.

Beckstead, J. W. (2009). Content validity is naught. *International Journal of Nursing Studies*, *46*(9), 1274–1283. https://doi.org/10.1016/j.ijnurstu.2009.04.014

Blazun, H., Kokol, P., & Vosner, J. (2015). Survey on specific nursing competences: Students' perceptions. *Nurse Education in Practice*, *15*(5), 359–365. https://doi.org/10.1016/j.nepr.2015.02.002

Bogaert, P. Van, Peremans, L., Heusden, D. Van, Verspuy, M., Kureckova, V., Van De Cruys, Z., & Franck, E. (2017). Predictors of burnout, work engagement and nurse reported job outcomes and quality of care: a mixed method study. *BMC Nursing*, *16*(5), 1–14. https://doi.org/10.1186/s12912-016-0200-4

Bramley, L., & Matiti, M. (2014). How does it really feel to be in my shoes? Patients' experiences of compassion within nursing care and their perceptions of developing compassionate nurses. *Journal of Clinical Nursing*, *23*(19–20), 2790–2799. https://doi.org/10.1111/jocn.12537

Brooten, D., Youngblut, J. M., & Youngblut, J. M. (2006). Nurse dose as a concept.(patient nurses). *Journal of Nursing Scholarship*, *38*(1), 94.

Cha, E. S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing*, *58*(4), 386–395. https://doi.org/10.1111/j.1365-2648.2007.04242.x

Chan, E. A., Jones, A., Fung, S., & Wu, S. C. (2012). Nurses' perception of time availability in patient communication in Hong Kong. *Journal of Clinical Nursing*, *21*(7–8), 1168–1177. https://doi.org/10.1111/j.1365-2702.2011.03841.x

Chan, J. (2014). *Modelling The gynecologic Oncology Workforce Using System Dynamics*. University of Toronto.

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha. *Organizational Research Methods*, *18*(2), 207–230.

https://doi.org/10.1177/1094428114555994

Ciconelli, R. M., de Soárez, P. C., Kowalski, C. C. G., & Ferraz, M. B. (2006). The Brazilian Portuguese version of the Work Productivity and Activity Impairment - General Health (WPAI-GH) Questionnaire. *Sao Paulo Medical Journal*, *124*(6), 325–332. https://doi.org/10.1590/S1516-31802006000600005

Cimiotti, J. P., Aiken, L. H., Sloane, D. M., & Wu, E. S. (2012). Nurse staffing, burnout, and health care-associated infection. *American Journal of Infection Control*, *40*(6), 486–490. https://doi.org/10.1038/jid.2014.371

Coatsworth, K., Hurley, J., & Miller-Rosser, K. (2015). A phenomenological study of student nurses volunteering in Nepal: Have their experiences altered their understanding of nursing? *Collegian*. https://doi.org/10.1016/j.colegn.2016.07.003

Dagne, T., Beyene, W., & Berhanu, N. (2015). Motivation and Factors Affecting it… Motivation and Factors Affecting It among Health Professionals in the Public Hospitals, Central Ethiopia. *Ethopian Journal of Health Science*, *25*(3), 231–242. https://doi.org/10.4314/ejhs.v25i3.6

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

DeVon, H. a., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., … Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability [Electronic Version]. *Journal of Nursing Scholarship*, *39*(2), 155–164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Dubois, C.-A., D'Amour, D., Pomey, M.-P., Girard, F., & Brault, I. (2013). Conceptualizing performance of nursing care as a prerequisite for better measurement: a systematic and interpretive review. *BMC Nursing*, *12*(1), 7. https://doi.org/10.1186/1472-6955-12-7

Epstein, J., Osborne, R. H., Elsworth, G. R., Beaton, D. E., & Guillemin, F. (2015). Cross-cultural adaptation of the Health Education Impact Questionnaire: Experimental study showed expert committee, not back-translation, added value. *Journal of Clinical Epidemiology*, *68*(4), 360–369. https://doi.org/10.1016/j.jclinepi.2013.07.013

Ferrer, M., Alonso, J., Prieto, L., Plaza, V., Monsó, E., Marrades, R., … Antó, J. M. (1996). Validity and reliability of the St George's respiratory questionnaire after adaptation to a different language and culture: The Spanish example. *European Respiratory Journal*, *9*(6), 1160–1166. https://doi.org/10.1183/09031936.96.09061160

Fleiszer, A. R., Semenic, S. E., Ritchie, J. A., Richer, M. C., & Denis, J. L. (2015). An organizational perspective on the long-term sustainability of a nursing best practice guidelines program: A case study. *BMC Health Services Research*, *15*(1), 1–16. https://doi.org/10.1186/s12913-015-1192-6

Flinkman, M., Leino-Kilpi, H., Numminen, O., Jeon, Y., Kuokkanen, L., & Meretoja, R. (2017). Nurse Competence Scale: a systematic and psychometric review. *Journal of Advanced Nursing*, *73*(5), 1035–1050. https://doi.org/10.1111/jan.13183

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring Change Over Time-Aseessing the Usefulness of Evaluative Instruments. *Journal of Chronic Diseases*, *40*(2), 171–178.

Horgas, A. L., Yoon, S. L., Nichols, A. L., & Marsiske, M. (2008). Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations. *Research in Nursing & Health*, *31*(4), 341–354. https://doi.org/10.1002/nur

Huicho, L., Dieleman, M., Campbell, J., Codjia, L., Balabanova, D., Dussault, G., & Dolea, C. (2010). Increasing access to health workers in underserved areas: A conceptual framework for measuring results. *Bulletin of the World Health Organization*, *88*(5), 357–363. https://doi.org/10.2471/BLT.09.070920

Jaakkimainen, L., Glazier, R., Barnsley, J., Salkeld, E., Lu, H., & Tu, K. (2014). Waiting to see the specialist: patient and provider characteristics of wait times from primary to specialty care. *BMC Family Practice*, *15*(1), 16. https://doi.org/10.1186/1471-2296-15-16

Johns, G. (2010). Presenteeism in the workplace: a review and research agenda. *Journal of Organizational Behavior*, *31*, 519–542. https://doi.org/10.1002/job.630

Kalisch, J. B., Lee, H., & Salas, E. (2012). The development and testing of nursing teamwork survey. In F. D. Polit & C. T. Beck (Eds.), *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice* (9th ed., pp. 338–351). Lippincott Williams & Wilkins.

Kanchanachitra, C., Lindelow, M., Johnston, T., Hanvoravongchai, P., Lorenzo, F. M., Huong, N. L., … Dela Rosa, J. F. (2011). Human resources for health in southeast Asia: Shortages, distributional challenges, and international trade in health services. *The Lancet*, *377*(9767), 769–781. https://doi.org/10.1016/S0140-6736(10)62035-1

Kassa, H., Murugan, R., Zedwu, F., Hailu, M., & Woldeyohannes, D. (2014). Assessment of knowledge , attitude and practice and associated factors towards palliative care among nurses working in selected hospitals , Addis Ababa , Ethiopia. *BMC Palliative Care*, *13*(4), 1–11. https://doi.org/10.1186/1472-684X-13-6

Kitchenham, B., & Pfleeger, S. L. (2002). Principles of survey research part 3: Constructing a survey instrument. *ACM SIGSOFT Software Engineering Notes*, *27*(2), 20. https://doi.org/10.1145/638574.638580

Kneafsey, R., Clifford, C., & Greenfield, S. (2013). What is the nursing team involvement in maintaining and promoting the mobility of older adults in hospital? A grounded theory study. *International Journal of Nursing Studies*, *50*(12), 1617–1629. https://doi.org/10.1016/j.ijnurstu.2013.04.007

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/J.1744-6570.1975.TB01393.X

Leach, L. S., & Mayo, A. M. (2013). Rapid Response Teams : qualitative analysis of their effectiveness. *American Journal of Critical Care*, *22*(3), 198–210.

Lemieux-Charles, L., & McGuire, W. L. (2006). What Do We Know about Health Care Team Effectiveness? A Review of the Literature. *Medical Care Research and Review*, *63*(3), 263–300. https://doi.org/10.1177/1077558706287003

Lutwama, G. (2011). *The performance of health workers in decentralised services in Uganda*. Retrieved from http://umkn-dsp01.unisa.ac.za/handle/10500/4866

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Reserach*, *35*(6), 382–385. https://doi.org/http://dx.doi.org/10.1097/00006199-198611000-00017

Mackenzie, S. B., & Podsakoff, P. M. (2012). Commentary on " Common Method Bias : Nature , Causes , and Procedural Remedies ". *Journal of Retailing*, *88*(January), 542–555. https://doi.org/10.1016/j.jretai.2012.08.001

Mackenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and Behavioral Research : Integrating New and Existing Techniques. *MIS Quarterly*, *35*(2), 293–334.

Makai, P., Cramm, J. M., van Grotel, M., & Nieboer, A. P. (2014). Labor productivity, perceived effectiveness, and sustainability of innovative projects. *Journal for Healthcare Quality : Official Publication of the National Association for Healthcare Quality*, *36*(2), 14–24. https://doi.org/10.1111/j.1945-1474.2012.00209.x

Manojlovich, M., & Sidani, S. (2008). Nurse dose: What's in a concept? *Research in Nursing and Health*, *31*(4), 310–319. https://doi.org/10.1002/nur.20265

McElroy, C., & Esterhuizen, P. (2017). Compassionate communication in acute healthcare: establishing the face and content validity of a questionnaire. *Journal of Research in Nursing*, *22*(1–2), 72–88. https://doi.org/10.1177/1744987116678903

Mehmet, T. (2013). Organizational variables on nurses' job performance in Turkey: Nursing Assesments. *Iranian Journal of Public Health*, *42*(3), 261–271. https://doi.org/10.2337/db06-1182.J.-W.Y.

Meretoja, R., Isoaho, H., & Leino-Kilpi, H. (2004). Nurse Competence Scale: development and psychometric testing. *Journal of Advanced Nursing*, *47*(2), 124–133. https://doi.org/10.1111/j.1365-2648.2004.03071.x

Meretoja, R., & Koponen, L. (2012). A systematic model to compare nurses' optimal and actual competencies in the clinical setting. *Journal of Advanced Nursing*, *68*(2), 414–422. https://doi.org/10.1111/j.1365-2648.2011.05754.x

Mert, T., & Ekici, D. (2015). Development of an Assessment Model for Evaluating the Performance of Nursing Services. *International Journal of Hospital Research*, *4*(1), 9–14.

Meyer, M. A., & Booker, J. M. (2001). *Eliciting and Analysing Expert Judgement A practical Guide*. American Statistical Association and the Society for Industrial and Applied Mathematics.

Needleman, J., Kurtzman, E. T., & Kizer, K. W. (2007). Performance measurement of nursing care : State of the science and the current consensus. *Medical Care Research and Review*, *64*(2), 10S–43S. https://doi.org/10.1177/1077558707299260

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling Procedures Issues and Applications*. Thousand Oak, London: SAGE Publications. https://doi.org/10.4135/9781412985772

North, N., & Hughes, F. (2012). A systems perspective on nursing productivity. *Journal of Health Organization and Management*, *26*(2), 192–214. https://doi.org/10.1108/14777261211230772

Numminen, O., Leino-Kilpi, H., Isoaho, H., & Meretoja, R. (2015). Newly Graduated Nurses' Competence and Individual and Organizational Factors: A Multivariate Analysis. *Journal of Nursing Scholarship*, *47*(5), 446–457. https://doi.org/10.1111/jnu.12153

Numminen, O., Meretoja, R., Isoaho, H., & Leino-Kilpi, H. (2013). Professional competence of practising nurses. *Journal of Clinical Nursing*, *22*(9–10), 1411–1423. https://doi.org/10.1111/j.1365-2702.2012.04334.x

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric Theory* (3rd Editio). McGraw-Hill, Inc.

Olsson, C., Forsberg, A., & Bjersa, K. (2016). Safety climate and readiness for implementation of evidence and person centered practice - A national study of registered nurses in general surgical care at Swedish university hospitals. *BMC Nursing*, *15*(1), 54. https://doi.org/10.1186/s12912-016-0174-2

Pillay, M. S., Johari, R., Hazilah, M., Asaari, A., Azman, B., Salikin, F., … Ismefariana, I. W. (2011). Hospital waiting time : the forgotten premise of healthcare service delivery ? *International Journal of Healthcare*, *24*(7), 506–522. https://doi.org/10.1108/09526861111160553

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences. *Organizational Research Methods*, *19*(2), 159–203. https://doi.org/10.1177/1094428115624965

Polit, D. F., & Beck, C. T. (2006). The content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Research in Nursing & Health*, *29*, 489–497. https://doi.org/10.1002/nur

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations. *Research in Nursing & Health*, *30*, 459–467. https://doi.org/10.1002/nur.20199

Rattray, J., & Jones, M. C. (2007). Essential elements of questionnaire design and development. *Journal of Clinical Nursing*, *16*(2), 234–243. https://doi.org/10.1111/j.1365-2702.2006.01573.x

Raykov, T. (2008). "Alpha if Deleted" and Loss in Criterion Validity Appeared in. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 275–285.

Rowe, A. K., De Savigny, D., Lanata, C. F., & Victora, C. G. (2005). How can we achieve and maintain high-quality performance of health workers in low-resource settings? *Lancet*, *366*(9490), 1026–1035. https://doi.org/10.1016/S0140-6736(05)67028-6

Sand-Jecklin, K., & Sherman, J. (2014). A quantitative assessment of patient and nurse outcomes of bedside nursing report implementation. *Journal of Clinical Nursing*, *23*(19–20), 2854–2863. https://doi.org/10.1111/jocn.12575

Scott, T., Mannion, R., Davies, H., & Marshall, M. (2003). Methods The Quantitative Measurement of Organizational Culture in Health Care : A Review of the Available Instruments. *Health Services Research*, *38*(3), 923–945.

Sheaff, R., Pickard, S., & Smith, K. (2002). Public service responsiveness to users' demands and needs: theory, practice and primary healthcare in England. *Public Administration*, *80*(3), 435–452. https://doi.org/10.1111/1467-9299.00312

Smith, S. (2012). Nurse Competence : A Concept Analysis. *International Journal of Nursing Knowledge*, *23*(3), 172–182.

Sonnentag, S., Volmer, J., & Spychala, A. (2008). Job Performance. *The SAGE Handbook of Organizational Behavior*, *1*, 427–447. https://doi.org/10.4135/9781849200448

Stolt, M., Charalambous, A., Radwin, L., Adam, C., Katajisto, J., Lemonidou, C., … Suhonen, R. (2016). Measuring trust in nurses ??? Psychometric properties of the Trust in Nurses Scale in four countries. *European Journal of Oncology Nursing*, *25*, 46–54. https://doi.org/10.1016/j.ejon.2016.09.006

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Topf, M. (1986). Three estimates of interrater reliability for nominal data. *Methodology Corner*, *35*(4), 253–255.

Umann, J. ., Guido, L. A. ., & Grazziano, E. S. . (2012). Presenteeism in hospital nurses [Presenteísmo em enfermeiros hospitalares]. *Revista Latino-Americana de Enfermagem*, *20*(1), 159–166. https://doi.org/10.1590/S0104-11692012000100021

Vanessa, A., Rodrigues, D., Vituri, D. W., Louren, C., Terezinha, M., Vannuchi, O., & Tiago, W. (2012). Nursing responsiveness the client ' s view. *Rev Esc Enferm USP*, *46*(6), 1446–1452.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on Models of Job Performance. *International Journal of Selection and Assessment*, *8*(4), 216–226. https://doi.org/10.1111/1468-2389.00151

WHO. (2006). *Working together For Health, The WHO Health Report 2006*. *World Health* (Vol. 19). https://doi.org/10.1186/1471-2458-5-67

WHO. (2010). *Increasing Access To Health Workers In Remote And Rural Areas Through Improved Retention: Global Policy Recommendations*. *WHO*. WHO Press.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*(5), 508–518. https://doi.org/10.1177/0193945903252998

Yakusheva, O., & Weiss, M. (2017). Rankings matter: nurse graduates from higher-ranked institutions have higher productivity. *BMC Health Services Research*, *17*(1), 134. https://doi.org/10.1186/s12913-017-2074-x

Zamanzadeh, T and Nemati, N. (2014). Details of content validity and objectifying it in instrument development. *Nursing Practice Today*, *1*(3), 163–171.

APPENDIX – A

Table 1: Summary of indices been used to quantify both items' content validity and inter-rater reliability during pre-testing.

Measure	Indices	Definition	Sources
Content Validity	1. Content Validity Index (CVI)	-Degree to which an instrument has an appropriate sample of items for construct being measured	Lynn (1986)
	2. Content Validity Ratio (CVR)		Lawshe (1975)
	3. Item-level CVI (I-CVI)	-the ratio of expert indicating item as "necessary" to the construct	Polit (2006, 2007)
	4. Scale-level CVI (S-CVI)	-Proportion of content experts giving item a relevance rating of 3 or 4	
	a. S-CVI/UA	-Content Validity of the overall scale	
	b. S-CVI/AVE	-Proportion of items on a scale that achieves a relevance rating of 3 or 4 by all the experts	
		- Average of the I-CVIs for all items on the scale	
Inter-Rater Reliability	1. Inter-Rater Agreement (IRA)	-the degree to which scores/ rating are identical	Gisev et al (2013)
	2. Inter-Rater Reliability (IRR)		
	3. Modified kappa	-the extent to which raters can consistently distinguish between different items on a measurement scale.	
		- the proportion of agreement remaining after chance agreement is removed	
	4. Cronbach alpha		Wynd (2003)

APPENDIX – B

Table 2: Summary of content domain in nurses' performance

Dimension	Subconstruct	Sources
1. Availability The degree of nurse perception that they have sufficient supply of nurses with optimal attendance to perform job activity and availability of patient time	Staff ratio: perception of having enough ratio of staff needed throughout group, a ratio on number of staffs as compared to other group of staff are sufficient	Huicho et al., (2010); Kanchanachitra et al., (2011); Lutwama, (2011); Cimiotti, Aiken, Sloane, & Wu, (2012)
	2. Absence Rate: perception on rate of staff being absence from work due to diseases or attend to work but perform their activity non-productively/ inherent to their function	Lutwama (2011); Numminen, Meretoja, Isoaho, & Leino-Kilpi, (2013); Kassa et al., (2014); Manojlovich & Sidani, (2008)
	3. Waiting time: Timely provision of service and impression of patient waiting time	Pillay et al., (2011); Chan et al., (2012); Chan, (2014); Abd Manaf et al., (2011); Jaakkimainen et al., (2014)
2. Responsiveness The extent of willingness or readiness of nurses to response to the needs and demands of patients	Empathy nurses' ability to customize themselves accordingly and willingness to help	Chan et al. (2012); Vanessa et al.,(2012); Coatsworth, Hurley, & Miller-Rosser, (2015); Bramley & Matiti, (2014); Numminen, Meretoja, Isoaho, & Leino-Kilpi, (2013)
	Acceptability: the extent of patients to have access to service provided by nurses	Vanessa et al. (2012); Sheaff, Pickard, & Smith, (2002); Brooten, Youngblut, & Youngblut, (2006)
	Receptiveness: an act of welcoming, an action of approximation, a "being with" and "being around", i.e., an attitude of inclusion	Stolt et al., (2016), Brooten et al. (2006), Vanessa et al. (2012), Numminen, Meretoja, Isoaho, & Leino-Kilpi, (2013)
3. Competence The degree of nurses' perception to the possession of required skill, knowledge, qualification, or capacity	NCS categories 1 = Teaching/coaching, 2 = Diagnostic functions, 3 = Managing situations, 4 =Therapeutic interventions, 5 = Ensuring quality, 6 = Work role	Numminen, Meretoja, Isoaho, & Leino-Kilpi, (2013); Numminen, Leino-Kilpi, Isoaho, & Meretoja, (2015), Meretoja, Isoaho, & Leino-Kilpi, (2004); Blazun, Kokol, & Vosner, (2015); Smith, (2012); Flinkman et al., (2017)
4. Productivity The degree of nurses' perception that they perform their job efficiently and providing effective services with reduced waste of staff time	Efficiency maximum output of nursing work as compared to inputs based on perceived adequacy of staffing	North & Hughes, (2012); Sand-Jecklin & Sherman, (2014)
	Presenteeism the condition in which nurses attend to work but perform activities/ functions in a non-productive way	Umann, Guido, & Grazziano, (2012); Johns, (2010);
	Effectiveness coordination between a collection of individuals who are interdependent in their task who shared collective responsibility for outcomes	Leach & Mayo, (2013), Lemieux-Charles & McGuire, (2006); Makai, Cramm, van Grotel, & Nieboer, (2014); Kneafsey, Clifford, & Greenfield, (2013)

APPENDIX – C

Table 5: Inter-Rater Agreement (IRA) based on Lay Expert Judgement

Subconstruct	Total item from LR (NPQ-01)	Items' Clarity *	Items' Relevancy *	Items' simplicity *	Items' ambiguity *	I-CVI >0.79	Item eliminated (%)	S-CVI/ Ave before item removed	Total item retained (%)	S-CVI/Ave after item removed
Availability	**12**	**11**	**10**	**10**	**11**	**10**	**2 (16.7%)**	**0.881**	**10 (83.3%)**	**0.893**
Responsiveness	12	6	4	3	3	4	5 (41.7%)	0.747	7 (58.3%)	0.813
Competence	35	23	23	18	22	24	13 (37.1%)	0.816	22 (62.9%)	0.856
Productivity	12	9	5	9	7	7	4 (33.3%)	0.815	8 (66.7%)	0.869
Total	71	49	42	40	43	45 (63.3%)	24 (33.3%)		47 (66.2%)	

Notes:
***Refers to number of items with IRA scores more than 80% degree of agreement among LEP.**
If the I-CVI is higher than 0.79, the item will be appropriate. If it is between 0.70 and 0.79, it needs revision.