Review Article

Development and Psychometric Evaluation of Scales: A Survey of Published Articles

Foroozan Atashzadeh-Shoorideh, PhD[1]; Farideh Yaghmaei, PhD[2*]

[1] PhD in Nursing, Nursing Management Department, Nursing & Midwifery School, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2] Department of Nursing, Zanjan Branch, Islamic Azad University, Zanjan, Iran.

Abstract

Background and purpose: *Using valid and reliable instruments is an important way for collecting data in qualitative researches. This paper is a report of a study conducted to examine the extent of psychometric properties of the scales in research papers published in Journal of Advanced Nursing.*
Methods: *In this study, the Journal of Advanced Nursing was chosen for systematic review. All articles which were published during 2007-2009 in this journal were collected and articles related to instrument development were selected. Each article was completely reviewed to identify the methods of instrument validation and reliability.*
Results: *From 980 articles published in Journal of Advanced Nursing during 2007-2009, 41 (4.18%) articles were about research methodology. In these, 12 articles (29.27%) were related to developing an instrument. In this study, review of 12 articles that published in Journal of Advanced Nursing, 2007-2009, showed that some of the articles did not measure psychometric properties properly, thus some of the developed scales need to measure other types of necessary validity. In addition, reliability testing needs to be performed on each instrument used in a study before other statistical analysis are performed. From 12 articles, all of the articles measured and reported Cronbach's alpha, but four of them did not measure test-retest.*
Conclusions: *Although researchers put a great emphasis on methodology and statistical analysis, they pay less attention to the psychometric properties of their new instruments. The authors of this article hope to draw the attention of researcher to the importance of measuring psychometric properties of new instruments.*

Keywords: *PSYCHOMETRIC, SCALES, CRITICAL REVIEW*
Journal of Medical Education Summer 2015; 14(4):174-205

Introduction

The credibility of results from a study is totally dependent on identifying, measuring, and collecting the right variables. Instruments are used to measure variables directly from subjects (1) and research instruments refer to questionnaires or inventories on which, data from a research project can be entered and stored for later analysis. An important part in the process of developing a questionnaire is to ensure its validity and reliability (2).
Using a valid and reliable instrument is an integral part of any research. Since interpretation of results depends on the validity of instruments used in studies, researchers should be sure about it (3). Validity is a significant and complicated issue which is considered by authors as well as readers (4). Types of validity includes: face validity, content validity, construct validity (factor analysis, validity by convergent validity, divergent validity, discriminating analysis) criterion validity (concurrent

∗Corresponding author: Farideh Yaghmaei, Associate professor, Department of Nursing, Zanjan Branch, Islamic Azad University, Zanjan, Iran.
Email: farideh.yaghmaei@iauz.ac.ir

validity and predictive validity), and successive verifications (5).

Measuring and reporting content validity of instruments is very important (6). Some authors in their articles have reported the process of measuring content validity frequently, while others did not. This type of validity can also help to ensure construct validity and give confidence to the readers and researchers about instruments. Content validity is used to measure the variables of interest. It is also known as content related validity, intrinsic validity, relevance validity, representative validity and logical or sampling validity (7-9). Therefore, content validity measures the comprehensiveness and representativeness of the content of a scale (10, 11).

Construct validity of an instrument is the theoretical frame or feature of a concept that the instrument measures such as intelligence, sorrow, or prejudice. Construct validity can be calculated by different methods including contrasted groups, convergent and divergent analysis or discriminate and factor analysis (12).

The criterion validity indicates to what degree the subject's performance on the measurement instrument and subject's actual behavior are related. Two forms of criterion-related validity are concurrent and predictive. Concurrent validity refers to an instrument's ability to distinguish among people who differ in their present status on the same criterion (13). Predictive validity refers to an instrument's ability to differentiate between people's performances or behaviors on the same future criterion (12).

Reliability refers to the consistency with which participants of similar characteristics and outlook understand and respond to the questions (2). The most common method of testing a scale's reliability is Cronbach's Alpha coefficient (14), and to determine the stability of the instrument, a test-retest must be carried out (15, 16). The internal consistency may be a necessary condition for homogeneity or unidimensionality of a scale and Cronbach's alpha should be 0.70 or higher (14, 17, 18).

Test-retest can be used to determine the stability of the instrument (15, 16). It is accomplished by administrating an instrument, waiting a reasonable period of time, and then re-administrating the instrument. The best correlation coefficient between the two sets of item scores is 0.70 or higher (1, 16).

Since strong measurement strategy is critical for proper research (1, 19), this study was conducted to evaluate the process of measuring validity and reliability of 12 development instruments papers published in Journal of Advanced Nursing (JAN) during 2007-2009.

Methods

In this study, the "Journal of Advanced Nursing" was chosen for review. All articles published during 2007-2009 in this journal were collected and articles related to instrument development were included. Each article was completely reviewed to identify the methods of instrument validation and reliability.

Results

From 980 articles published in Journal of Advanced Nursing during 2007-2009, 41 (4.18%) articles were about research methodology. In these research methodology papers, 12 articles (29.27%) were related to developing a instrument. Table 1 shows the features of the articles. None of 12 articles mentioned their psychometric properties absolutely (Table 1).

Discussion

Appropriate instruments have a significant influence on validity of a study. Invalid and unreliable instruments may show incorrect results and using findings is doubtful. In

Table 1. Instruments' characteristics of published articles in Journal of Advanced Nursing, 2007-2009.

Author /s	Instrument/s	Type of validity	Criticism of validity	Type of reliability	Criticism of reliability
Ushiro R (2009) (20)	The psychometric properties of the Nurse–Physician Collaboration Scale (NPCS)	***Content validity,*** in this article was measured by revising the content and wording based on the responses made by the physicians and nurses. ***Factor analysis:*** with exploratory factor analysis was (CFI) <0.8 and RMSEA >0.08 for the single-factor model, and CFI <0.9 and RMSEA <0.08 for the three-factor model. ***Concurrent validity*** was measured by relationships between nurses' responses to the Nurse–Physician Collaboration Scale (NPCS) and the Intergroup Conflict Scale. There were statistically significant negative correlations for all three factors (r = _0.20 to _0.236, P < 0.01). Among the relationship between physicians' responses to the Nurse–Physician Collaboration Scale (NPCS) and the Intergroup Conflict Scale, there were statistically significant small negative correlations for shared patient's information, (r = _0.165, P < 0.01) and	***Content validity*** is an initial step in establishing validity, but the best method in this regard is Content Validity Index (14), that didn't measure in this study. In addition, the number of person for measuring content validity should be between 15-20 (9) that did not mention in this study. ***Factor analysis*** with exploratory factor analysis was measured and reported. It is acceptable but cut-off value for factor loadings wasn't reported. ***Concurrent validity*** was reported but the ranges of correlations for item-totals and inter-item were low. The concurrent validity value must be ranging from 0 to +1 (4). ***Convergent validity*** was reported but these ranges were low. The convergent validity value must be ranging	***Cronbach's alpha*** coefficients and test–retest reliability coefficients were measured. Cronbach's alpha coefficient for the physicians' responses to the Nurse–Physician Collaboration Scale (NPCS) were 0.911 for shared patient information, 0.926 for joint participation in the cure/care decision-making process and 0.842 for cooperativeness. When Cronbach's alpha coefficient of the item-total correlations were compared with those obtained when an item had been eliminated, no items was found lower than coefficient value.	The alpha coefficients of 0.70 and above indicate that these scales are internally consistent (16). All results for test–retest reliability were satisfactory, except for the physician responses regarding sharing of patient information (0.629). However, other α values were 0.70 – 0.92, which confirms the stability of the scales. The test–retest correlation coefficients for nurses were mentioned and it is acceptable.

cooperativeness. (r = _0.152, P < 0.01). ***Convergent validity*** was done with the Team Characteristic Scale and with both the nurses' responses (r = 0.360–0.523, P < 0.01) and physicians' responses (r = 0.435–0.639, P < 0.01) to the Nurse–Physician Collaboration Scale (NPCS). The used scale in this study for convergent validity did not validate or didn't report its validity and reliability.	from 0 to +1 (4, 5). In addition; the psychometric of the used scale for convergent validity did not mentioned.	The item-total correlation values were high, ranging from 0.502 to 0.801. The item-total correlation values were high, ranging from 0.423 to 0.787. The ***test–retest*** (The interval between the first and the second test was 2–3 weeks) correlation coefficients for nurses were 0.710 (P<0.01) for sharing of patient information, 0.658 (P<0.01) for joint participation in the cure/care decision-making process, and 0.676 (P < 0.01) for cooperativeness. The test–retest correlation coefficients for physicians were 0.624 (P < 0.01) for sharing patient information, 0.798 (P < 0.01)

				for joint participation in the cure/care decision-making process and 0.774 (P < 0.01) for cooperativeness.	
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Chang H-J et al (2009)(2)	Chinese version of the Positive and Negative Suicide Ideation (PANSI) Inventory	***Content validity***, in this article was not measured. ***Factor analysis*** was examined by using both with exploratory factor analysis and confirmatory factor analysis (CFA) and all item-total coefficients ranged from 0.42 to 0.71. The results indicated that the two factor oblique model had the best fit. The confirmatory factor analysis using the two factor model yielded the following results: CFI = 0.950, RMSEA=0.078. ***Convergent validity*** was demonstrated by statistically significant and positive correlations between total scores on the positive and negative suicide ideation-negative suicide ideation (PANSI-NSI) and the Children's Depression Inventory (CDI) (r=0.61), the positive and negative suicide ideation positive ideation (PANSI-PI) and the Cognitive Triad for Children	***Content validity*** or face validity is an initial step in establishing validity (6) that was not measured in this study. ***Factor analysis*** with exploratory factor analysis and confirmatory factor analysis was measured and reported. It is acceptable but, cut-off value for factor loadings wasn't reported. ***Convergent validity*** was reported and these ranges were moderate level. The convergent validity value must be ranging from 0 to +1. If the convergent measures are closely related, the validity of each instrument is strengthened (Burns and Grove 2007). ***Divergent validity*** was reported but the ranges of correlations were	Cronbach's alpha coefficients and test–retest reliability coefficients were measured. The ***Cronbach's alpha Coefficients*** were 0.86 and 0.94 for the total scores on the positive and negative suicide ideation positive ideation (PANSI-PI) and the positive and negative suicide ideation-negative suicide ideation (PANSI-NSI) respectively. The ***test–retest*** (The interval between the first and the second test was 4 weeks) was carried out. Intra-class correlation coefficients were	Internal consistency based on the suggested criterion level indicating adequate internal consistency for a coefficient's α of 0.70 or above (14).

Inventory (CTI-C) (r = 0.65), the positive and negative suicide ideation positive ideation (PANSI-PI) and the self-control schedule (SCS) (r = 0.46). ***Divergent validity*** was demonstrated by statistically significant and negative correlations between the total Scores on the positive and negative suicide ideation positive ideation (PANSI-PI), the Children's Depression Inventory (CDI) (r=-0.52), the negative suicide ideation-negative suicide ideation (PANSI-NSI), the Cognitive Triad for Children Inventory (CTI-C) (r=-0.52), and the negative suicide ideation-negative suicide ideation (PANSI-NSI) and the self-control schedule (SCS) (r= -0.30). All correlations were statistically significant at the P<0.01 level. ***Predictive Validity*** was measured one year after first-wave study with the Chinese Version of the Positive and Negative Suicide Ideation Inventory (PANSI-C). Logistic regression analysis showed that the total score on the negative suicide	moderate. The divergent validity value must be ranging from -1 to 0. If the convergent measure of instrument is negatively correlated with other measures, validity for each of the instrument is strengthened (Burns and Grove 2007). The process of predictive validity and the score of this study is acceptable.	0.82 and 0.70 for the total scores on the positive and negative suicide ideation positive ideation (PANSI-PI) and positive and negative suicide ideation-negative suicide ideation (PANSI-NSI). All correlations were statistically significant at the P<0.05 level.	

		ideation-negative suicide ideation (PANSI-NSI) in the first-wave study statistically significantly predicted the attempted- suicide behaviour after 1 year (coefficient = 0.095, P<0.001; CI = 1.05–1.15). The overall classification rate was good, at 89.4%. The total score of the positive and negative suicide ideation positive ideation (PANSI-PI) in the first-wave study also statistically significantly predicted the attempted suicide behaviour after 1 year (coefficient = _0.084, P<0.05, CI = 0.86–0.99).			
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Eizenberg MM et al (2009) (22)	Moral Distress Questionnaire for Clinical Nurses	***Content validity,*** in this article was not measured. ***Factor analysis*** was examined by using exploratory factor analysis and all item-total coefficients ranged from 0.56 to 0.90. The results indicated that the three factors yielded. The authors didn't report CFI and other results of factor analysis. But they mentioned cut-off value. ***Discriminate validity***: In addition, to provide additional evidence for the construct validity of the	***Content validity*** is an initial step in establishing validity (6, 16), that didn't measure in this study. Measuring and reporting of content validity in questionnaire developing is necessary and important (16). It is recommended to determine content validity before construct validity. ***Factor analysis*** with exploratory factor analysis was measured and reported. It is	Internal consistency was measured by using ***Cronbach's alpha***. For the three factors the internal consistency is above 0.79 (for three factors are 0.851, 0.791 and 0.804). Stability was examined by use	The alpha coefficients of 0.70 and above indicate that these scales are internally consistent (15, 16). The test–retest correlation coefficients were mentioned but it is low (1). It is recommended to increase the items in second version of this questionnaire.

		questionnaire, a comparison was made between two groups (hospital nurses and community clinic nurses), as it was assumed that differences would be observed in pressure resulting from different moral dilemmas. To examine these differences, t-tests for independent samples were conducted. A statistically significant difference was found between means for two of the three factors relationships and time (For relationship t=2.171 and for time t=2.208). These differences provide further evidence for the discriminant validity of the questionnaire.	necessary reporting of their results but the authors didn't report CFI and other results of factor analysis (23).	of ***test-retest reliability*** (The interval between the first and the second test was 1 month). The correlation between the two measurements was 0.624 (P<0.001), 0.385 (P<0.05) and 0.535 (P<0.01) respectively for the three factors.	
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Liu M et al (2009) (24)	Competency Inventory for Registered Nurses in Macao	***Content validity***, in this article and Content Validity Index (CVI) was reported based on the other studies. ***Factor analysis*** with exploratory factor analysis was (CFI) <0.8 and RMSEA >0.08 for the single-factor model, and CFI <0.9 and RMSEA <0.08 for the three-factor model. Confirmatory factor analysis was employed to test the construct validity of the	***Content validity*** is an initial step in establishing validity (6), and it supports construct validity (3) that didn't measure in this study. Measuring and reporting of content validity in questionnaire developing is necessary and important (16). It is recommended to determine content validity and Content	Internal consistency reliability and stability were estimated by Cronbach's alpha and paired t-test, respectively. Internal consistency ***Cronbach's alpha*** was 0.90 for the overall scale and 0.71–0.90 for	Measuring reliability is reported and is acceptable. The stability indicates a high degree of stability over a period of time and satisfactory degree of homogeneity (8).

		instrument. The factor loading value across 55 items ranged from 0.310 to 0.725. A cut-off value of 0.3 for factor loadings was applied as this is considered to indicate statistical significance.	Validity Index (CVI) in every questionnaire developing (6, 16). ***Factor analysis*** with exploratory factor analysis and confirmatory factor analysis was measured and reported.	subscales. Internal consistency was 0.74. The interval between the first and the second test didn't reported.	The best interval time between first and second test in test-retest is 2-4 weeks (5, 16). It is recommended to report of interval between two tests.
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Zisberg A, Young HM & Schepp K (2009) (25)	Scale of Older Adults' Routine (SOAR)	***Content validity***: In this study, items were generated on the basis of a literature review and then systematically tested for content validity. Then, the instrument's content validity was rated on the basis of the instrument's item relevance to older adult routine in the pilot sample. The relevance, clarity, simplicity based on Content Validity Index (CVI) items weren't reported. ***Convergent validity:*** In order to test the convergent validity of SOAR, the mean deviation scores on the subscale level correlated with the functional indicators (ADL and IADL). The ADL score was found to be negatively correlated with the consistency of time spent (mean deviation score for	***Content validity:*** Measuring and reporting of content validity in questionnaire developing is necessary and important (16, 19). ***Convergent validity*** was reported and these ranges were moderate level. In this study, the authors reported the convergent validity between _1 to 0. But the convergent validity value must be ranging from 0 to +1. If the convergent measures are closely related, the validity of each instrument is strengthened (5).	Intra-class correlation coefficient statistics were used to test reliability at the item level of the continuous scores as well as subscale scores. Across all types of scores, 21 (50%) consistently presented moderate to high test–retest reliability (ICC >0.41). Six items (14.3%) presented poor reliability on all four scores (ICC <0.40). These items were shopping, passive transportation,	ICC scores should be considered as reliability indices in four groups of estimate levels: high (ICC >0.80), substantial (0.60 <ICC<0.80), moderate (0.41< ICC <0.60) and poor to fair (ICC <0.40) (26). Kappa coefficient is almost perfect (27). The test–retest correlation coefficients were mentioned but, it is low (1). Reliability for

		duration) on each basic and rest activity (r = -0.41, -0.34; P < 0.01 respectively), as well as with the consistency of total time spent on basic and rest activities (mean deviation score for total duration, r = 00.56, -0.33; P < 0.01 respectively).		and medical treatment, attending concerts/movies/sports events, participating in group activities and taking care of an older person. On the subscale level, over 73% of the scores showed high to substantial reliability and none showed poor reliability. Kappa coefficients was done for nominal variables and it was over 0.75 (item % of agreement = 88.4%–100%). Only16.6% had kappa coefficients in the low range (j < 0.40). Test–retest reliability for subscales is 0.46 to 0.85. The interval between the first and the second test didn't report.	overall scale wasn't mentioned.
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**	
Pelande	Child Care	***Content validity***: In this	***Content validity***: A	Internal	A correlation	

r T, Leino-Kilpi H & Katajisto J (2009) (28)	Quality at Hospital (CCQH) instrument	study, following a literature review and interviews/ drawings by hospitalized children (n=40), the items were designed and an expert panel (n=7) assessed the instrument's ***content validity***. To judge the validity of the items and subcategories on a scale from one to four for relevance and clarity; to indicate whether or not (yes/no) a subcategory belonged to a particular main category; whether or not the subcategory measured quality and whether or not there was any overlap between the different subcategories. The least relevant subcategories were 0.38 and 0.67, so these items deleted. The least clarity of subcategories was 0.65 and 0.69, whereas the level of agreement for all other subcategories was over 0.90. Level of agreement among nurses was over 0.95 for all subcategories measuring quality, except for appearance (0.37), sense of humour (0.69) and humanity (0.93). In the nurses' assessments, the subcategories of humanity (0.31), caring and	scale-level CVI of %75 or higher is acceptable. The reporting of content validity index must be based on percent (3, 16). ***Factor analysis*** didn't report obviously. The process of it should be clear.	consistency by using ***Cronbach's alpha*** was 0.373–0.812 for subscales, but for the overall scale didn't report. The alpha values showed a tendency to increase during the course of the instrument development for all the main categories: in nursing characteristics from 0.383 to 0.557, nursing activities from 0.763 to 0.809, and nursing environment from 0.584 to 0.761. Item-to-total correlations were calculated for the various subcategories in nursing activities and environment and for the main category of nurse characteristics. Item-to-total correlations ranged from 0.062	Coefficient between 0.80 and 0.90 is desirable, but 0.70 is acceptable for new instruments (29). Correlations for item-totals and inter-item were reported. Combining certain subcategories or increasing more items, especially in the subcategory, can improve the reliability (30).

		communication (0.31), and education (0.31) showed the greatest overlap with other subcategories. The ***factor analysis*** of CCQH was assessed by using principal component analysis to measure the level of congruence of empirical results with the main categories of nursing activities and environment. No principal component analysis was carried out for the main category of nurse characteristics.		to 0.611. The lowest item-to-total correlations were obtained for the subcategories of physical care and treatment, and entertainment. The items 'takes account of child's food preferences' and 'provides relief for pain' were the most problematic. These items were, however, not deleted from the instrument as their contents are crucial in this context.	
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Carlson C (2008) (31)	Carlson's Prior Conditions Instruments (CPCIs), to assess the four theoretically-derived prior conditions of Previous practice, felt needs/problems, innovativeness	***Content validity*** was done by reviewing literature and theoretical definition and was supported through review by experts. The average of CVI scores for relevancy of all items within each instrument were 0.79 to 1.0 (The average of CVI scores of all items within each instrument were 1.0 for the Previous Practice Instrument, 0.79 for the Felt Needs/Problems	The reporting of content validity in this study is acceptable. The reporting of content validity index must be based on percent (3, 16). Rattray and Jones suggest that a KMO greater than 0.5 supports a factor analysis, and that anything less than 0.5 is probably not amenable to useful factor analysis. So, this	Cronbach's alpha coefficients were measured. Each instrument demonstrated internal consistency (alpha range= 0.731–0.825).	The alpha coefficients of 0.70 and above indicate that these scales are internally consistent (16, 19). In addition, test–retest reliability needs to be confirmed to assess the stability of the

	and norms of the social system that influence nurses' decisions to adopt evidence-based pain management practices.	Instrument, 0.94 for the Innovativeness Instrument, and 0.98 for the Norms of the Social System Instrument). The clarity, simplicity based on Content Validity Index (CVI) items was not reported. ***Factor analysis*** was examined through principal components factor analysis with varimax rotation and reported for each factor of instruments. Factors were established using the Kaiser rationale by retaining eigenvalues over 1.0. To establish salient factors, the items with correlations above 0.3 on more than one factor were deleted, as they were repetitious. The Kaiser–Meyer Olkin (KMO) measure of sample adequacy was then determined. The KMOs of Carlson's Prior Conditions Instruments (CPCIs) ranged from 0.655 to 0.841.	KMO measure is acceptable. For achieving more accurate instrument, another type of construct validity such as predictive validity is needed (6).	Inter-item correlations are between 0.2 and 0.7. After item analysis for internal consistency reliability, the Previous Practice Instrument was reduced to 13 items, the Felt Needs/ Problems Instrument to 14 items, the Innovativeness Instrument to nine items, and the Norms of the Social System Instrument to nine items. Alphas were 0.825, 0.76, 0.731 and 0.775 respectively.	measures over time (6).

Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Pisanti R et al (2008) (32)	Occupational Coping Self-Efficacy for Nurses Scale (OCSE-N)	In this article, ***content validity*** was not measured. ***Factor analysis***: exploratory factor analysis, and confirmatory factor analysis was done. Construct validity	***Content validity*** is an initial step in establishing validity (6), and it supports construct validity (3) that didn't measure in this study. ***Factor analysis:*** with	Internal reliability was estimated by calculating the ***Cronbach's alpha*** coefficient for the scale(s) derived from the analysis	The Internal consistency Cronbach's alpha reported. The alpha coefficients of 0.7 and above

	with exploratory factor analysis is (CFI) <0.75and RMSEA >0.15 for the first model, and CFI <0.92 and RMSEA <0.08 for the second model. ***Concurrent validity*** was assessed by estimating correlations between the Occupational Coping Self-Efficacy for Nurses Scale (OCSE-N) dimensions and two external criteria: Maslach Burnout Inventory (MBI) dimensions and coping dimensions. Pearson's correlation coefficients between the Occupational Coping Self-Efficacy for Nurses Scale (OCSE-N) dimensions and both the Maslach Burnout Inventory (MBI) variables and Coping Inventory for Stressful Situations – Short Version (CISS-SV) dimensions were all statistically significant. The OCSE-N dimensions were positively associated with task coping strategies (r = 0.07 to 0.08, P < 0.05) and negatively associated with both emotion-focused and avoidant strategies (r = _0.09 to _0.08, P < 0.01). The OCSE-N Scales also correlated with the burnout	exploratory factor analysis and confirmatory factor analysis was measured and reported. It is acceptable. ***Criterion validity*** was reported carefully.	and by checking whether every item increased Cronbach's alpha. Cronbach's alpha reliability were done for two subscales (For 'CSE to manage general nursing burden' alpha = 0.77; and for 'CSE to manage the relational difficulties in the workplace', alpha = 0.79).	indicate that these scales are internally consistent (16, 18). In addition, reliability such as test-retest needs to be confirmed to assess the stability of the measures over time (6).

		dimensions. They were negatively correlated with both emotional exhaustion (r= _0.31 to _0.21, P < 0.01) and depersonalization (r = _0.25 to _0.19, P < 0.01), and positively associated with personal accomplishment (r = 0.21 to 0.22, P < 0.01). These patterns of correlations support the construct validity of the Occupational Coping Self-Efficacy for Nurses Scale (OCSE-N).			
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Barnes C.R.& Adams on-Maced o E.N. (2007) (33)	Perceived Maternal Parenting Self-Efficacy (PMP S-E) instrument	In this study, ***Content validity*** was done by reviewing literature and theoretical definition and was supported through review by participants in a pilot study. ***Factor analysis*** was measured and cut-off value of 0.3 for factor loadings was applied as this is considered to indicate statistical significance. Factor 1 had an Eigen value of 8.235 and explained 41% of the variance, factor 2 had an Eigen value of 1.496 and explained 7.48% of the variance, factor 3 had an Eigen value of 1.314 and explained 6.57% of the variance, and factor 4 had an	***Content validity*** is an initial step in establishing validity, but the best method in this regard is Content Validity Index (14), that didn't measure in this study. ***Construct validity*** with exploratory factor analysis was measured and reported. It is necessary reporting of their results but the authors didn't report CFI and other results of factor analysis (23). In addition, cut-off point is low. ***Divergent validity*** was reported but the ranges of correlations were	***Cronbach's alpha coefficient*** was used to calculate internal consistency reliability estimates for the Perceived Maternal Parenting Self-Efficacy (PMP S-E) instrument; this reached an acceptable level (0.91). The internal consistency reliability estimates for each of the subscales were also acceptable	The alpha coefficients of 0.70 and above indicate that these scales are internally consistent (16, 18). The test–retest correlation coefficient was mentioned and it is acceptable.

		Eigen value of 0.255 explaining 6.27% of the variance. ***Divergent Validity*** by using the Maternal Self-Report Inventory was r_s = 0.4 (P < 0.05) and using the Maternal Postnatal Attachment Scale was r_s = 0.31, (P< 0.01).	moderate. The divergent validity value must be ranging from -1 to 0. If the convergent measure of instrument is negatively correlated with other measures, validity for each of the instrument is strengthened (4).	[subscale 1 (0.74), 2 (0.89), 3 (0.74) and 4 (0.72)]. In addition, item-whole correlation revealed that all items correlated statistically significantly with total scores (ranging from 0.30–0.77). The ***test–retest*** (The interval between the first and the second test was 10 days) correlation coefficients was 0.96.	
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
Van Laar, D et al. (2007) (34)	Work-Related Quality of Life (WRQoL) scale for healthcare workers	Survey of the literature and qualitative expert reviews were used to assess the ***content validity*** of the measure. For factor analysis, exploratory factor analysis and Confirmatory factor analysis were done. A cut-off value of 0.5 for factor loadings was applied. By using Split-half factor analysis for the full data, a first data set with 481 cases to be used in the exploratory step (hereafter referred to as data set EXPLORE), and a	***Content validity*** is an initial step in establishing validity, but the best method in this regard is Content Validity Index (14), that didn't measure in this study. Factor analysis with exploratory factor analysis and confirmatory factor analysis was measured and reported. The criterion for establishing model fit via goodness of fit indices statistics	Internal consistency by using ***Cronbach's alpha*** was 0.75–0.86 for subscales, and for the overall scale was 0.96.	The Internal consistency Cronbach's alpha reported. The alpha coefficients of 0.7 and above indicate that these scales are internally consistent (16, 18). In addition, other type of reliability such as test-retest needs to be confirmed to

second data set with 472 cases to be used in the confirmatory analysis (hereafter referred to as data set (CONFIRM). A preliminary principal component analysis (PCA) was carried out on the WRQoL EXPLORE data set. Twelve components with eigenvalues above 1.0 were generated. Using this procedure, 34 items were removed, leaving 24 items, which together represented six factors [Factor 1: Job and Career Satisfaction (JCS) contained six items, Factor 2: General Well-Being (GWB) also contained six questions, Factor 3: Home–Work Interface (HWI) reflected three items, Factor 4: Stress at Work (SAW) was represented by two items, Factor 5: Control at Work (CAW): Three items loaded on component five, Factor 6: Working Conditions (WCS) with three items].

Confirmatory factor analysis was conducted on the remaining 23 items and support was found for the model in the CONFIRM data set ($P < 0.01$, CFI = 0.93, GFI = 0.90, NFI = 0.89 and RMSEA = 0.06).

generally suggest that values around 0.90 are acceptable and values >0.90 or higher are considered good fit for the CFI, GFI and the NFI (35).

Values < 0.05 for the RMSEA indicate a close fit whereas values between 0.05 and 0.10 represent adequate to mediocre fit (36).

assess the stability of the measures over time (6).

Author /s	Instrument/s	Type of validity	Criticism of validity	Type of reliability	Criticism of reliability
Otieno O.G et al (2007) (37)	An instrument to measures nurses' use, quality and satisfaction with Electronic Medical Record (EMR) systems	***Content validity*** was addressed by basing the items on previous surveys and reviewing the instrument by a panel of nurses experienced in nursing informatics. ***Factor analysis,*** in this study was examined. A cut-off value of 0.4 for factor loadings was applied. Factor analysis revealed three subscales in use of Electronic Medical Record (EMR) scale. Also factor analysis revealed two subscales in 'quality of Electronic Medical Record (EMR)' and three-factor subscales in 'user satisfaction' are determined by factor analysis. ***Concurrent validity*** was assessed by calculating correlation coefficients between the scales of the instrument and the global measure. Criterion-related validity was not addressed explicitly in this study. However, the degree of correlation between the scores of the two subscales of EMR use (Nursing Care Management and Order Entry); two subscales of quality of EMR (Information	Measuring and reporting ***content validity*** in questionnaire developing is necessary and important (16, 19). In this study the reporting of content validity is acceptable. But CVI didn't report. ***Factor analysis:*** Exploratory factor analysis was measured and reported. It is acceptable. ***Concurrent validity*** was measured but the degree of correlation was not mentioned.	The reliability of each resultant factor was computed using Cronbach's alpha coefficient. Criteria were based on Cronbach's alpha coefficients ≥0.7 within a construct and item–total correlation ≥0.4 within the subscales. Items were deleted where necessary to achieve an alpha value of at least 0.7. In this study, overall Cronbach's alpha coefficient didn't mention, but it reported for each subscale. Three subscales with low Cronbach's alpha coefficient were removed from the final instrument.	The Internal consistency Cronbach's alpha reported. The alpha coefficients of 0.7 and above indicate that these scales are internally consistent (15, 16). In addition, reliability needs to be confirmed to assess the stability of the measures over time (6). In this study, validity and reliability of the instrument was reported together. It is recommended reporting of validity and reliability will be separated.

		Quality and Service Quality) and one subscale of user satisfaction (Impact of EMR systems on Clinical Care) revealed in all cases.			
Author /s	**Instrument/s**	**Type of validity**	**Criticism of validity**	**Type of reliability**	**Criticism of reliability**
FU M.R., McDaniel R.W. & Rhodes V.A. (2007) (38)	Adapted Symptom Distress scale: The Symptom Experience Index (SEI)	***Content validity*** of the Symptom Experience Index (SEI) was ensured by 15 general medical-surgical and oncology patients in the study who had tested the reliability and validity of the Adapted Symptom Distress Scale version 2 (ASDS-2). In addition, content validity of the SEI is supported by inclusion of symptoms that have been identified by patients in other studies as well as those perceived by patients with cancer in a series of the investigators' studies. ***Construct validity***: The authors used multiple comparisons (with Kruskal-Wallis test) to estimate construct validity by determining statistically significant differences between pairs of contrasting groups.	In this study, the reporting of ***content validity*** is acceptable. But CVI didn't report. The validity of this study isn't complete. ***Construct validity*** was measured through multiple comparisons. But, factor analysis can be used as an exploratory or confirmatory technique to estimate the underlying dimensions or to reduce redundant items in an instrument.	Cronbach's alpha was computed to measure internal consistency. Correlation analysis for the total experience revealed Cronbach's alpha 0.91; for total occurrence 0.85; for total distress 0.84. Reliability for the subscales was estimated using Cronbach alpha for each subscale: respiratory (0.8), cognitive (0.79), eating/gastrointestinal (0.73), pain/discomfort (0.76), neurological (0.78), fatigue/sleep/restlessness (0.81), eliminations (0.74) and appearance (0.77). To measure the	The reliability is reported correctly. The stability indicates a high degree of stability over a period of time and satisfactory degree of homogeneity (8). For test-retest procedures, the second administration generally is recommended about 2–14 days after the first (39). Because of the attributes of the phenomena being measured (symptom occurrence and distress), only healthy adult participants were asked to complete the

stability of the SEI, a test-retest method (during two different periods of 2–4 hours apart) was used with 63 healthy adult participants. Intra-class correlation coefficients were calculated to estimate test-retest reliability. Test-retest scores were strongly correlated for total symptom experience (r=0.93), occurrence (r=0.94) and distress (r=0.92).	Symptom Experience Index (SEI) during two different periods of 2–4 hours apart. This time lapse was sufficient to avert participants' recall of their previous response (i.e. absence of flu symptoms) and to preclude activities (i.e. onset of flu symptoms after 2 weeks) that may have affected the stability of the characteristic (symptom experience) being measured (40).

addition, it affects implications of research findings to the population under study (19).

In this study, review of 12 articles that published in the Journal of Advanced Nursing, 2007-2009, showed that psychometric properties did not present, since from 12 articles only 2 of the articles documented validity completely, and 5 of the articles reported incomplete content validity and 5 of them did not measured it. In regard to measuring construct validity, factor analysis is a useful method. From 12 articles that reviewed, 4 articles measured factor analysis completely, 4 of them measured or reported incomplete and 4 of the articles did not measure it. In regard to other type of validity, from 12 articles, only one article measured concurrent validity, one article measured discriminate validity, one article measured divergent validity and one article measured convergent validity.

As stated before, measuring 3 types of validity for new developed instruments is necessary. Therefore, measuring validity to determine the appropriateness of an instrument should be for a special group. The

findings showed that some of the articles did not measure psychometric properties properly, thus some of the developed scales need to measure other types of necessary validity.

In addition, reliability testing needs to be performed on each instrument used in a study before other statistical analysis are performed. From 12 articles, all of the articles measured and reported Cronbach's alpha and test-retest, but 4 of them did not measure test-retest.

Conclusion

It can be concluded that although researchers put a great emphasis on methodology and statistical analysis, they pay less attention to the psychometric properties of their new instruments. The authors of this article hope to draw the attention of researcher to the importance of measuring psychometric properties of new instruments.

Acknowledgements

The authors would like to thank Dr. Zagheri Tafreshi for commenting on a draft of this paper. Her feedback was much appreciated

References

1. Houser J. Nursing Research: Reading, Using, and Creating Evidence. 2nd Ed. Sudbury, Jones and Bartlett Publishers; 2011.
2. Watson R, Hugh McK, Seamus C, John K. Nursing Research: Designs and Methods. Elsevier Health Sciences; 2008.
3. Yaghmaie F. Subjective computer training: Development of a scale. Journal of Medical Education. 2004; 5(1):33-7.
4. Burns N, Grove SK. Practice of Nursing Research, Conduct, Critique and Utilization. 6th Ed. Philadelphia: Saunders Co; 2009.
5. Burns N, Grove SK. Understanding Nursing Research, Building an Evidence-Based Practice. 6th Ed. Philadelphia: Saunders Co; 2014.
6. Rattray J, Jones MC. Essential elements of questionnaire design and development. Journal of Clinical Nursing. 2007; 16(2):234–43.
7. Bush CT. Nursing research. Virginia: Reston Publishing Co; 1985.
8. Polit DF, Hungler BP. Nursing Research Principles and Methods. 7th Ed. Philadelphia: Lippincott Co; 2004.
9. Dempsy PA, Dempsy AD. Using Nursing Research, Process, Critical Evaluation and Utilization. 5th Ed. Philadelphia: Lippincott Co.; 2000.
10. Kerlinger FN. Foundations of behavioral research. 3rd Ed. New York: CBS Publishing; 1986.
11. Yaghmaie F. Content validity and its estimation. Journal of Medical Education. 2003; 3(1):25-7.
12. LoBiondo-Wood G, Haber J. Nursing Research; Methods, Critical Appraisal, and Utilization. 8th Ed. St. Louis: Mosby-Elsevier; 2014.
13. Polit DF, Beck CT. Essentials of Nursing Research, Methods, Appraisal and Utilization. 6th Ed. Philadelphia: Lippincott Williams & Wilkins; 2006.
14. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd Ed. New York: McGraw Hill Inc; 1994.
15. Yaghmaie F. Development of a scale for measuring user computer experience. Journal of Research Nursing. 2007; 12(2):185-90.
16. Yaghmaei F. Measuring Behavior in Research by Valid and Reliable Instruments. 2nd Ed. Tehran: Shahid Beheshti Medical University Publishing; 2009. (Persian)
17. Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychological Assessment. 1995; 7(9):309–19.
18. Yaghmaie F. Reliability and its measurement in quantitative studies. Journal of Faculty of Nursing & Midwifery, Shahid Beheshti University of Medical Sciences. 2003; 13(42):22-7. (Persian)
19. Zagheri-Tafreshi M, Yaghmaie F. Factor analysis of construct validity: A review of nursing articles. Journal of Medical Education. 2007; 10(1):19-26.
20. Ushiro R. Nurse–Physician Collaboration Scale: development and psychometric testing. Journal of Advanced Nursing. 2009; 65(7):1497–508.
21.Chang HJ, et al. Chinese version of the positive and negative suicide ideation: Instrument development. Journal of Advanced Nursing. 2009; 65(7):1485–96.
22. Eizenberg MM, et al. Moral distress questionnaire for clinical nurses: instrument development. Journal of Advanced Nursing. 2009; 65(4):885–92.

23. Munro BH. Statistical Methods for Health Care Research. 5th edition. Philadelphia: Lippincott Williams and Wilkins; 2005.
24. Liu M, Yin L, Ma E, Lo S, Zeng L. Competency Inventory for Registered Nurses in Macao: instrument validation. Journal of Advanced Nursing. 2009; 65(4):893–900.
25. Zisberg A, Young HM, Schepp K. Development and psychometric testing of the Scale of Older Adults' Routine. Journal of Advanced Nursing. 2009; 65(3):672–83.
26.Woods-Dauphinee S, Berg K, Daley K. Monitoring status and evaluating outcomes: An overview of rating scales for the use with patients who have sustained a stroke. Topics in Geriatric Rehabilitation. 1994; 10(2):22–41.
27. Cyr L, Francis K. Measures of clinical for nominal and categorical data: The kappa coefficient. Computers in Biology and Medicine. 1992; 22(4):239-46.
28. Pelander T, Leino-Kilpi H, Katajisto J. The quality of pediatric nursing care: developing the Child Care Quality at Hospital instrument for children. Journal of Advanced Nursing. 2009; 65(2):443–53.
29. Stotts NA, Aldrich KM. How to try this defines your terms, Evaluating instruments for use in nursing practice. Advanced Journal of Nursing. 2007; 107(10):71-2.
30. Ferketich S. Focus on psychometrics: Aspects of item analysis. Research in Nursing & Health. 1991; 14(2); 165–8.
31. Carlson C. Development and testing of four instruments to assess prior conditions that influence nurses' adoption of evidence-based pain management practices. Journal of Advanced Nursing. 2008; 64(6):632–43.
32. Pisanti R, Lombardo C, Lucidi F, Lazzari D, Bertini M. Development and validation of a brief Occupational Coping Self-Efficacy Questionnaire for Nurses. Journal of Advanced Nursing. 2008; 62(2):238–47.
33. Barnes CR, Adamson-Macedo EN. Perceived Maternal Parenting Self-Efficacy (PMP S-E) instrument: development and validation with mothers of hospitalized preterm neonates. Journal of Advanced Nursing. 2007; 60(5):550–60.
34.Van Laar D, Edwards J A, Easton S. The Work-related quality of life scale for healthcare workers. Journal of Advanced Nursing. 2007; 60(3):325–33.
35. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin. 1980; 88(3):588–606.
36. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit in Testing Structural Equation Models. Sage, Newbury Park, CA, USA; 1993.
37. Otieno OG, Toyama H, Asonuma M. Nurses' views on the use, quality and user satisfaction with electronic medical records: questionnaire development. Journal of Advanced Nursing. 2007; 60(2):209–19.
38. Fu M, McDaniel RW, Rhodes VA. Measuring symptom occurrence and symptom distress: development of the symptom experience index. Journal of Advanced Nursing. 2007; 59(6):623–34.
39. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th edition. Oxford University Press Inc.: New York; 2015.
40. Fu MR, Rhodes VA, Xu B. The Chinese translation: The index of nausea, vomiting, and retching (INVR). Cancer Nursing. 2002; 25(2):134–40.