Journal of Medical Education Development, 2018, Vol 10, No 28, 106 - 119
http://zums.ac.ir/edujournal/

Zanjan University of Medical Sciences

Content Validity and Reliability of the Measurement Tools in Educational, Behavioral, and Health Sciences Research

Vakili MM[1*], Jahangiri N[2]

[1]Department of Health Education & Health Promotion, Zanjan University of Medial Sciences, Zanjan, Iran.
[2]Educational Development Center, Zanjan University of Medial Sciences, Zanjan, Iran.

Article Info

Article Type:
Review Article

Article history:
Received 13 Oct 2017
Accepted 22 May 2018
Published 11 Mar 2018

Keywords:
Psychometrics
Content Validity
Face Validity
Reliability
Questionnaire

Abstract

Development and assessment of measurement tools are important stages in the research processes regarding social, educational, and medical sciences, which mainly focus on the measurement of characteristics, qualitative variables, and abstract variables. Validity and reliability are two important components of researcher-made tools. The quality of assessment and confirming of validity and reliability are major concerns in research. Before publishing their findings, researchers are required to provide a report on the quality assessment of the validity and reliability of measurement tools. Precision in explaining these features could lay the ground for commenting on the trustworthiness and validation of the obtained findings, as well as their comparison with previous studies. If the validity and reliability of research instruments are not confirmed, the endeavors of researchers might be degraded. The present study aimed to elaborate on the significance of validity and reliability indices in medical research and the qualitative and quantitative assessment of content validity on confirming the reliability of measurement tools.

[*]***Corresponding author:*** Vakili MM, ***Email:*** Vakili@zums.ac.ir

This article is referenced as follows: Vakili M M, Jahangiri N. Content Validity and Reliability of the Measurement Tools in Educational, Behavioral, and Health Sciences Research . J Med Educ Dev. 2018; 10 (28) :105-117

Introduction

Investigation of abstract concepts and human behavior are major elements of research processes regarding social, behavioral, educational, and health care. In such studies, evaluation and measurement of features and characteristics are of paramount importance. Common approaches to measure such variables include the use of researcher-made scales, tests, and questionnaires. Therefore, researchers and experts are constantly attempting to develop their applied measurement tools in accordance with their research objectives and study variables (e.g., cognitive, behavioral, emotional, and psychoanalytical variables) in the target populations (1). Tools or instruments could also be used for the examination of abstract concepts, such as knowledge, emotional values, attitudes, psychomotor skills, and clinical simulations or they could be employed as demographic surveys.

Medical educators have endeavored to codify and develop valid and reliable tests and questionnaires in order to promote the reliability of the evaluation of educational programs (2). Considering that most of the studies in educational, psychological, and behavioral fields involve such approaches, the acceptable validity and reliability of assessment and measurement tools have gained greater importance (3). Results of every study must be reliable as far as possible, and the procedures in the study must be assessed in terms of their objectives and subject matters (4). As a result, validity and reliability are essential to the credibility of every measurement tool (2).

Although researchers constantly emphasize on the importance of the accurate assessment of the validity and reliability of research instruments, a high proportion of the published studies in the fields of behavioral sciences and health education do present the reports on the validity (40-93%) and reliability (35-80%) of their instruments, which is considered to be grievous (1). It should be noted that such limitations do not only apply to the studies in the fields of health research and education or social and behavioral sciences. Considering the wide application of measurement and assessment tools in health sciences, such shortcomings may also be witnessed in many studies in other fields.

Without the evaluation of such significant indices of measurement tools, researchers may be led toward inaccurate and unreliable research conclusions and recommendations. In other words, lack of confidence in whether

an applied tool in a study has been able to provide stable, accurate scores makes researchers unsure of the accuracy of the reported findings. In order for researchers and experts to increase the benefits and applicability of their findings, it is essential to pay attention to the features of the applied psychometric instrument through confirming its validity and reliability. Otherwise, all their endeavors and financial investment might be dissipated, leading to the provision of invalid recommendations and findings for policymakers and planners (1). Another consequence of such failure is the loss of opportunity for other researchers to apply and retest of previous findings (5, 6).

The present study aimed to elaborate on the stages of the methodology used for the evaluation of the validity (with an emphasis on content validity) and reliability of research instruments. Meanwhile, we have discussed the approachesusedtocalculate the quantitative indices and interpret research findings.

What is Validity?

Validity refers to the relevant interpretations regarding the obtained scores in a test used for a specific purpose, as well as their compatibility with scientific evidence and theories. In other words, validity determines the methods to accurately interpret the results of a test for a specific objective. Interpretation of the significance of the obtained results from the instruments used to assess physical quantities (e.g., height, blood pressure) could be achieved directly and simply. However, recognition of complex, abstract concepts (e.g., awareness, attitude) are not as easy to obtain. Consequently, researchers attempt to collectively examine a set of abstract concepts in a relative manner on a structural basis. The results obtained from a psychometric analysis are merely significant in the context where the construct is to be assessed (7). Validity is basically associated with the accuracy and reliability of the obtained scores in a measurement tool, which is conventionally known as the ‘holy trinity’, including content validity, construct validity, and validity of criteria (1).

Content validity is one of the most common assessment methods for the reliability of researcher-made instruments, which is often determined in the initial stage of developing the instrument. Content validity consists of qualitative and quantitative approaches, with an emphasis on two main study groups (target group and elites).

Face validity involves a qualitative approach and is considered to be an objective judgment of the structure of the research instrument in

the initial stages of a study. This concept refers to the rationality of a test in the view of the respondents, through which the researcher aims to ensure the relevance of the instrument with the objective of the study in appearance, similar conception of the target group of the items based on the opinion of the researcher, agreement of the target group with the statements used in the items of the instrument, and acceptability of the components and generality of the instrument from the perspective of the target group (3).

Evaluation of content validity helps the researcher to provide reliable evidence to ensure the inclusion of all the important aspects and key concepts in the evaluation of the subject matter, as well as the acceptability of all the components of the tool in the view of the expert panel (8). In this process, it is of paramount importance to assess the reliability, face validity, and content validity of the tool both qualitatively and quantitatively, with the centralized role of the target group in developing the research tool.

In the assessment of the methodology of the current literature in Iran with the aim of evaluating abstract concepts, such as knowledge and attitude, or examining the effects of educational interventions on the knowledge and attitudes of target groups, it is evident that researchers have faced major problems in the accurate reporting of the process of the validity of research instruments, thereby disregarding or presenting inadequate information in this respect. In several studies, the exact number of the experts on the panel is not determined or less than five experts are employed to confirm validity, while in some cases, the comments of the experts are not thoroughly mentioned in regards to the quantitative indices or qualitative assessment of the validity of the tool (confirmation of validity by some experts). In addition, consulting with a panel of experts may be totally neglected in many studies (5, 9, 10).

The minimum number of the experts required to evaluate the validity of a research instrument and calculating the content validity is five (11). In many cases, the researcher only mentions the reports in the previous studies, in which the quality and assessment of the process of validity may still be unclear.

Content Validity

Evaluation of the Validity of an Instrument Focusing on the Target Group

This stage of evaluation is primarily focused on the target group of the study and is performed in both a qualitative and quantitative manner. To verify the validity of

a tool quantitatively, it is essential to grade the items of the instrument within a chart based on a five-point Likert scale, including absolutely essential (score 5), essential (score 4), moderately important (score 3), slightly important (score 2), and not important at all (score 1). Afterward, the questionnaire must be completed by some individuals, so that they could provide their comments regarding the importance per each item.

To determine the impact score of an item, the frequency of the respondents with the scores 4 or 5 should initially be determined (Table 1), followed by the separate calculation of the total score attributed to each item and the mean score of each item. In the next stage, using the formula of impact item=frequency (%) × importance, the impact of each item is estimated. If the mentioned index is estimated at >1.5 for the items, the item is regarded as important by the target group and will be preserved for the following phases of psychometric analysis. Otherwise, the item should be eliminated from the instrument.

In the qualitative evaluation of the validity of an instrument, the researchers are concerned with issues such as the problematic understanding of the statements, proportionality and proper relevance of the items with each other, possibility of ambiguity and misinterpretations regarding the statements or word meanings. To do so, a small sample of the target group is asked to determine the items that appear unclear, difficult or improper (12). In this regard, selecting a minimum of 10 samples with similar demographic characteristics to the main target group seems effective (10, 13).

Table1: Five Point Likert Scale for Target Group Opinions About Items of Questionnaire

No.	Items	not necessary at all 1	Slightly Important 2	Moderately Important 3	Important 4	Strongly Important 5
1						
2						
3						
…						

Evaluation of Content Validity by a Panel of Experts

To verify the content validity by a panel of experts, each of the items must be assessed qualitatively and quantitatively.

Content Validity Ratio (CVR)

Content validity ratio (CVR) is a useful statistical item for the rejection or acceptance of the items in a questionnaire, which is internationally acknowledged as an assessment technique to confirm content validity (14). This index was first proposed by Lawshe in 1975 and has been used to confirm the selection of the optimal and most accurate content (essentiality of items) based on the poll taken from a panel consisting of five experts, and the number of these experts could also increase to 40. The experts must be specialized in the subject matter of the research. Ideally, these experts are selected from a wide spectrum of relevant specialties. Normally, a panel consisting of 5-10 experts would be appropriate, while more than 10 experts may be unnecessary. To this end, the items of a tool should be designed in the form of a table and separately provided to each expert, so that they would present their views toward the essentiality of the items in the research instrument (14). Each item is discussed in three scales of essential, useful but not essential, and not essential, and the panel of experts decides whether each item could be recorded in the tool.

After obtaining the comments of the experts, the value of content validity is calculated based on the formula (14), ne is represents the number of panel members who have chosen the item of " essential", and N is the number of panel members. The value of each item is compared with the values in the Lawshe's table, and if the calculated content validity value is equal to or higher than the determined value in Lawshe's table, the item is preserved; otherwise, it should be eliminated from the list of the items.

$$CVR = \frac{n_e - (N/2)}{N/2},$$

Table2: Content Evaluation Form by Experts Panel

No.	Items	Essential	Useful but not Essential	Not Essential
1				
2				
3				
…				

Table3: Minimum Values of Content Validity Ratio

Minimum Value	No. of Experts Panel	Minimum Value	No. of Experts Panel
0.51	14	0.99	5 - 7
0.49	15	0.75	8
0.42	20	0.78	9
0.37	25	0.62	10
0.33	30	0.59	11
0.31	35	0.56	12
0.29	40	0.54	13

Content Validity Index (CVI)

Content validity index (CVI) provides the data on the validity of the items separately. CVI could be used to determine the content validity of the entire tool. CVI represents the mean values in proportion to the content validity of an instrument in all the items with the minimum CVR of 0.79 that have been maintained in the tool (14). Tilden et al. have verified the proper CVI of more than 0.70 to confirm the acceptability of items in a questionnaire (15), while Davis mentions the values of more than 0.80 in this regard (16).

CVI has been proposed to ensure that the items in a questionnaire are optimally designed to measure the contents. To this end, all the items in a questionnaire should be classified in a table and separately provided to the panel of experts, so that they would present their views about the three parameters of relevance, simplicity, and clarity of each item based on four-point Likert scale (8, 17).

After obtaining the comments of the experts, data extraction should be performed, followed by estimating CVI using the specific formula for each of the three mentioned parameters independently. Finally, the mean values are calculated for each item, and total CVI is determined for the items as well.

$$CVI = \frac{(\text{the number of experts giving a rating of either 3 or 4})}{\text{the number of experts}}$$

After estimating the CVI for all the items, the acceptability of each item is assessed based on the following criteria: acceptable items (scores of>0.79), items requiring modification (scores 0.70-0.79), and unacceptable items (scores <0.70). The acceptable items are preserved in the questionnaire, unacceptable items are removed, and modifiable items are revised and corrected by the panel of experts.

Table 4: Criteria for Measuring Content Validity Index According To Experts Panel Opinion

No.	Content validity Index (CVI)																Average total score of the index
	Items	Relevance				simplicity					clarity						
		Very Relevant	Relevant but Need Minor Revision	But Need Some Revision	Not Relevant	Point	Very Simple	Simple but Need Minor Revision	But Need Some Revision	Not Simple	Point	Very Clear	Clear but Need Minor Revision	Ned Some Revision	Not Clear	Point	
1																	
2																	
3																	
…																	

Qualitative Assessment of Content Validity

Following the evaluation of the validity of a research tool through content validity, it is important to qualitatively assess the items of the questionnaire by the panel of experts. To do so, the experts are asked to provide their written feedback and recommendations to modify each item in terms of their content, Persian grammar, number of the words and length of the phrases, sequence, addition of new items, proportionality to the sociocultural characteristics of the target group, and overall structure of the instrument.

After obtaining the feedback of the experts, the research team must collect and apply the comments and recommendations in order to make the necessary modifications in the items of the questionnaire. In the case of drastic changes in the number or contents of the items, it might be necessary to repeat the process of evaluating the validity of the instrument by the panel of experts and target group. The items in the research instrument must be simple and clear with a proper sequence and exquisite font and design, so that the target group would be able to complete a legible questionnaire without any ambiguities.

What is Reliability?

In psychometric and educational studies that focus on a specific behavior or abstract feature, the reliability of the measurement tool is a common concern among researchers. A measurement tool is considered trustworthy

when it has proper reliability (2). Reliability determines whether measurements could be repeated, and any random effect leading to the variations of the measured variables is regarded as the source of measurement errors (18). According to Joppe, reliability or repeatability is defined as the stability of the findingsthroughtime,and aresearch instrument is reliable when the study results could be repeatedusingthesame cognitive methods (19). Miller and Kirk propose three types of reliability in quantitative studies, including the level of obtaining similar results with repeating the evaluation, stability of the measured variables through time, and similarity of the measurements within a specific period (20, 21).

The reliability of a research instrument could be improved by considering such principles as the clarity of the items, applying the instructions on the completion of the test, so that it could be easily comprehended by the respondents, and effective training of the examiners on accurate grading (3). Despite various techniques for assessing the reliability of research tools, the two most common methods in this regard are the evaluation of internal consistency and test-retest.

Internal Consistency of the Items

Undoubtedly, the Cronbach's alpha coefficient is one of the most critical and common statistical indices in the research on the development and application of tests and is used in many studies in the fields of psychology, education, social sciences, sociology, economic sciences, law, and particularly medical and nursing sciences (2, 18). This is because in comparison with the other methods of assessing reliability, the Cronbach's alpha is easier to apply and is measured only once (2).

Despite the wide application of the Cronbach's alpha coefficient in scientific research, its use and interpretation seem insufficient (2, 22). This index was first introduced by Lee Cronbach in 1951 for the assessment of the internal consistency of the items in a test or scale, which covers a score range of 0-1 (2). This index indicates whether all the items of a designed scale have proper consistency with the studied subject matter. Inaccurate use of Cronbach's alpha may lead to the incorrect disregard of the test or inaccurate interpretations of the research findings. To avoid this, it is essential to properly recognize the associations between conceots such as internal consistency, homogeneity,orsingularity, thereby enhancing the use of Cronbach's alpha (2). Furthermore, it is important to assess the internal

consistencyoftheitemsofaquestionnaire before its use in a study. If the research tool consists of multiple dimensions or constructs, the internal consistency should beassuredbetween alltheitemsofeach dimension.

The number of the items in an instrument affects their consistency with each other, while the dimensions of the instrument affect the Cronbach's alpha value. There are several reports on the acceptable values of Cronbach's alpha; accordingly, coefficient values of ≥0.9, 0.8-0.9, 0.7-0.8, 0.5-0.6, and <0.5 are considered to be optimum, good, acceptable, questionable, and unacceptable, respectively. Low Cronbach's alpha values may be due to the low number of the items in the tool, poor internal consistency among the items or non-homogenous constructs (2).

Test-Retest

Charles believes that the concept of reliability or stability is achieved when the provided responses to the items of the questionnaire or the scores of the respondents remain stable in the case of test-retest within two different periods. In other words, if a characteristic is evaluated with the same tool, the obtained results must be identical. High stability indicates high reliability, which assures the repeatability of the results. However, test-retest might reduce the reliability of an instrument since responding to the items again may increase the sensitivity of the subjects and influence their responses in the retest phase (21). Therefore, we cannot guarantee no changes in the perceptions of individuals under the influence of external factors. This issue is known to affect the responses of the study subjects.

Repetition of a set of the test items at consecutive intervals is likely to intensify the effects of some of inherent characteristics of the subjects on the different choice of the answer compared to the previous tests, thereby changing the reliability of the research instrument and reduce the accuracy and stability of the test and test scores. Therefore, researchers are responsible for gaining the trust of others regarding the stability and accuracy of the tests and their scores in a study. It is also noteworthy that the interval between the initial test and retest is a matter of debate in this regard (23), which has been reported to be from a few hours to six months, while the interval of two weeks to one month has been agreed upon by the majority of researchers. The interval between the two tests should be long enough for the subjects to not remember their initial responses, while it should not be so short that the contents on the knowledge and attitude

undergo drastic changes. Principally, the reliability of a research instrument is expected to decrease with the length of the interval between the two tests (23).

Figure 1 helps the better recognition of the concepts of validity and reliability, as well as the significant correlation of these indices. As is seen in this figure, reliability is synonymous with the repeatability of similar scores, and validity is to be placed exactly at the center of a specific objective. In the upper and lower portions of Figure 1, moving from the left to the right is associated with the falling trend of reliability, which consequently affects validity.

As illustrated in the circle on the right and upper portion of Figure 1, reduction of reliability is associated with an increase in the inclination of validity toward a random event. Meanwhile, the circle on the left and lower portion of Figure 1 shows very high reliability although high reliability never guarantees high validity (24). As such, if the research tool lacks proper validity, high reliability is of no value. In Figure 1, the optimal state in terms of validity and reliability is shown in the circle on the left and upper portion, which demonstrates that the obtained scores are identical and at the center of the objective in repeated tests.

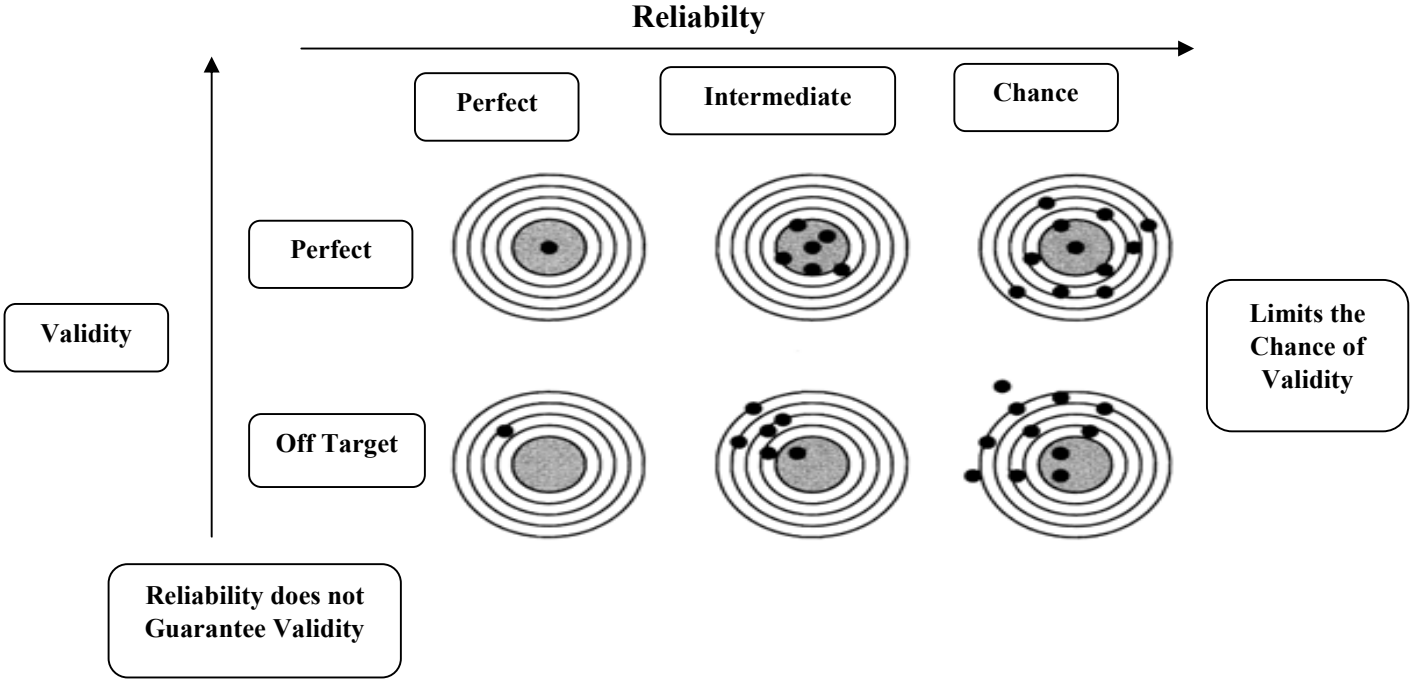

Figure 1: The Relationship Between Reliability and Validity

In the studies conducted in Iran, there are few problems in the comparison of the validity of a research tool and reporting its reliability. However, researchers have only sufficed to evaluating the internal consistency of questionnaire items and estimating the Cronbach's alpha value in confirming the reliability of an instrument, while the stability of the obtained results are not properly discussed. This could be attributed to the difficulty in the repeated access to the samples or undermining the repetition of investigations within a short period on behalf of the researcher or subjects.

Conclusion

Review of the methodologies in the investigations focusing in abstract concepts (e.g., knowledge, attitude, health beliefs) in Iran clearly shows that researchers are faced with numerous challenges in the proper reporting of the process of assessing the validity and reliability of research instruments. This process is either not thoroughly performed or the presented data are not adequate. On the other hand, in many studies, providing the feedback of the target group is not carried out although it is considered to be one of the most critical stages in the assessment of the validity of a research tool. With respect to applying the comments of the panel of experts, this stage remains ambiguous without mentioning the number and specialties of the experts, as well as presenting the data on the qualitative and quantitative indices in the psychometric analysis; the majority of these studies only generally state that the validity of the research tool has been confirmed by some experts (fewer than five experts).

Compared to validity, there are fewer issues associated with the reports on the reliability of measurement tools. Studies mostly report the obtained value for the Cronbach's alpha coefficient and reliability status of the instrument, while the assessment of the stability of the instrument using the test-retest method is disregarded. Acknowledging the paramount importance of the reliability and validity of every measurement tool, it should be noted that validity is prioritized over the reliabilityof thetool. If the research instrument lacks proper validity (not able to accurately measure the study variables), it cannot be evaluatedintermsofthe quality of its reliability. Undoubtedly, there are limitations in all studies; nevertheless, the credibility of a researchprimarilydepends on its methodology. Despite the importance of all the components of research methodology, it could be claimed

that the methods and instruments used in a study to evaluate the variable (qualitative and quantitative) are of utmost importance. No matter how accurately a study is conducted, invalid instruments challenge the judgment regarding the findings, meeting the objectives, and acceptability of the research hypotheses. Consequently, interpretations and discussions about the findings or their comparison with other studies are unreliable.

In the present study, we attempted to emphasize on the critical importance of validity and reliability as the two fundamental features of research measurement tools. Furthermore, we elaborated on each stage of the assessment of content validity and presented two of the most common methods used to evaluate the reliability of research instruments. It is expected that the study of such basic concepts in the methodology of research in medicine and other fields attract the attention of researchers and authors since today, the techniques used for the quality, collection, and presentation of the validity and reliability of measurement tools is among the foremostcriteriaforaccepting research findings by credible domestic and foreign journals.

References

1- Barry AE, Chaney B, Piazza-Gardner AK, Chavarria EA. Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*. 2014;41(1):12-8.

2- Tavakol M, Dennick R. Making sense of Cronbach's alpha. *International journal of medical education*. 2011;2:53-5.

3- Drost EA. Validity and reliability in social science research. *Education Research and perspectives.* 2011;38(1):105-23.

4- Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse education today*. 2004; 24 (2):105-12.

5- Haghdoost A, Pourkhandani A, Motaghipisheh S, Farhoudi B, Fahimifar N, Sadeghirad B. Knowledge and Attitude concerning HIV/AIDS among Iranian Population: a Systematic Review and Meta- Analysis. *Iranian Journal of Epidemiology*. 2011;6(4):8-20.

6- Nejat S, Feyzzadeh A, Asghari S, Keshtkar A, HESHMAT R, Majdzadeh S. HIV risk factors in Iran; systematic review, meta-analysis and generalized impact fraction approaches. *PAYESH*. 2007;6(1):45-54.

7- Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. The *American journal of medicine*. 2006;119(2):166.e7-16.

8- Polit DF, Beck CT. The content validity index: are you sure you know what's being reported?

Critique and recommendations. *Research in nursing & health*. 2006;29(5):489-97.

9- Mohammadbeigi A, Mohammadsalehi N, Aligol M. Validity and Reliability of the Instruments and Types of MeasurmentS in Health Applied Researches. *Journal of Rafsanjan University of Medical Sciences*. 2015;13(12):1153-70.

10- Vakili MM, Hidarnia AR, Niknami S. Development and Psychometrics of an Interpersonal Communication Skills Scale (A.S.M.A) among Zanjan Health Volunteers. *Hayat*. 2012; 18 (1):5-19

11- Lawshe CH. A quantitative approach to content validity. Personnel psychology. 1975;28(4):563-75.

12- Lacasse Y, Godbout C, Series F. Health-related quality of life in obstructive sleep apnoea. *European Respiratory Journal*. 2002;19(3):499-503.

13- Vakili MM, Hidarnia AR, Niknami S, Mousavinasab N. Development and psychometrics of Health Belief Model instrument about HIV/AIDS. *Zahedan Journal of Research in Medical Sciences zahedan*. 2012;14(9):64-71.

14- Gilbert GE, Prion S. Making Sense of Methods and Measurement: Lawshe's Content Validity Index. *Clinical Simulation in Nursing*. 2016;12(12):530-1.

15- Tilden VP, Nelson CA, May BA. Use of qualitative methods to enhance content validity. *Nursing Research*. 1990;39(3):172-5.

16- Davis LL. Instrument review: Getting the most from a panel of experts. *Appl Nurs Res*. 1992;5(4):194-7.

17- Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in nursing & health*. 2007;30(4):459-67.

18- Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98-104.

19- Joppe M. The Research Process. Retrieved February 25, 1998. 2000.

20- Kirk J, Miller ML. Reliability and validity in qualitative research: Sage; 1986.

21- Golafshani N. Understanding reliability and validity in qualitative research. *The qualitative report*. 2003;8(4):597-606.

22- Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assessment*. 1996;8(4):350-3.

23- DeVon HA, Block ME, Moyle□Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*. 2007;39(2):155-64.

24- Krippendorff K. Content Analysis An Introduction to Its Methodology. London: Sage Publications; 2004. P. 211-214.