Original Article

Measuring assessment standards in undergraduate medical programs: Development and validation of AIM tool

Madiha Sajjad[1], Rehan Ahmed Khan[2], Rahila Yasmeen[3]

ABSTRACT

Objective: To develop a tool to evaluate faculty perceptions of assessment quality in an undergraduate medical program.

Methods: The Assessment Implementation Measure (AIM) tool was developed by a mixed method approach. A preliminary questionnaire developed through literature review was submitted to a panel of 10 medical education experts for a three-round 'Modified Delphi technique'. Panel agreement of > 75% was considered the criterion for inclusion of items in the questionnaire. Cognitive pre-testing of five faculty members was conducted. Pilot study was done with 30 randomly selected faculty members. Content validity index (CVI) was calculated for individual items (I-CVI) and composite scale (S-CVI). Cronbach's alpha was calculated to determine the internal consistency reliability of the tool.

Results: The final AIM tool had 30 items after the Delphi process. S-CVI was 0.98 with the S-CVI/Avg method and 0.86 by S-CVI/UA method, suggesting good content validity. Cut-off value of < 0.9 I-CVI was taken as criterion for item deletion. Cognitive pre-testing revealed good item interpretation. Cronbach's alpha calculated for the AIM was 0.9, whereas Cronbach's alpha for the four domains ranged from 0.67 to 0.80.

Conclusions: 'AIM' is a relevant and useful instrument with good content validity and reliability of results, and may be used to evaluate the teachers´ perceptions about assessment quality.

KEYWORDS: Faculty perceptions, Assessment, Quality, Standards, Tool, Development, Validation.

INTRODUCTION

Medical schools face the challenge of developing and implementing quality assessment programs

1. Dr. Madiha Sajjad, FCPS, MHPE.
    Department of Pathology,
2. Dr. Rehan Ahmed Khan, FCPS, FRCS, MHPE.
    Department of Surgery,
3. Dr. Rahila Yasmeen, MHPE.
    Department of Medical Education,
1-3: Riphah International University,
    Rawalpindi, Pakistan.

    Correspondence:

    Dr. Madiha Sajjad, FCPS.
    Professor, Pathology Department,
    Riphah International University,
    Peshawar Road,
    Rawalpindi, Pakistan.
    Email: madiha.sajjad@riphah.edu.pk

that are acceptable to accrediting bodies and society at large. Evaluating the assessment system at intervals assures that assessments remain effective and up-to-date.[1,2] Several instruments are available in literature to evaluate educational environment for both students[1] and teachers[3] at undergraduate institutes as well as in clinical environment.[4] However based on current literature, no validated survey instrument has been identified in an undergraduate medical context to reliably evaluate the quality of assessment. Perceptions about assessment process in an institute can provide important information about gaps between accepted standards and the implemented assessment practices.[5] Quality assurance in assessment requires involvement of the entire institutional team especially a faculty well acquainted and engaged in the culture of student assessment.[6] However, despite the important role of faculty in successful

implementation of assessments, little focus is given in the medical education literature about medical teachers' perceptions about assessment in institutes as an indicator of the quality of student assessment.

This study aims to develop and validate a tool in order to evaluate the quality of assessment practices and for institutional self-evaluation to inform, guide and improve assessment quality.

METHODS

A mixed methods study design was used with sequential qualitative and quantitative components in the following four-stage process for developing and validating the questionnaire:

1. Review of literature and preliminary questionnaire item development: Quality indicators of assessment were identified from literature search based on the quality standards provided by multiple sources such as the WFME document[7], LCME[8], CACMS[9] and students' perception questionnaire[1] etc. A preliminary draft questionnaire of 34 items was prepared for further amendments through the Delphi technique.

2. Modified Delphi technique for consensus development on questionnaire items and content validation of the AIM tool: A 3 round modified Delphi approach was used in which 18 medical education experts having a diploma/degree in medical education and working in undergraduate medical institutes, were invited to participate through email. In round one, the panellists were asked to grade 'relevance' of items, on a five point Likert scale. Percentage responses and median scores for each item were calculated. For round 2, items were added or amended based on results and the questionnaire was resent to the panelists. The panelists were instructed to either accept the suggested items, reject with reason and propose modifications where considered necessary. Panel agreement of > 75% on each statement was considered the criterion for inclusion of items in the subsequent round. In round 3, the panel was asked to indicate relevance of each item on a 4point Likert scale for final inclusion of the item into the questionnaire. The panel was also provided with a list of four domains, and the item statements allocated to it. They were requested to indicate agreement to the domain allotted for the statements and if not in agreement, to reallocate the statement to their preferred choice of domain. Panel agreement of > 75% was considered as inclusion criteria of item in the given domain.

Content validity index (CVI) for the individual items (I-CVI) and of the scale (S-CVI) was calculated using the ratings of item relevance by content experts in the last round.

3. Cognitive pretesting to check for faculty understanding: Five faculty members were selected for cognitive pretesting through convenience sampling method. Individual interviews were conducted through 'concurrent verbal probing method'.[10] The 4 cognitive validity criteria used were: 'correct item interpretation, comprehensible explanation, answer choice compatibility with interpretation, and overall item cognition' across the 5 participants.[11,12]

4. Pilot study on a sample of faculty to establish the reliability of the tool: The final 30 item questionnaire was given to 30 randomly selected participants, from both the basic and clinical sciences faculty, in order to determine the reliability of the final developed AIM tool. The summary of methodology is given in Fig.1.

RESULTS

Quality indicators of assessment identified after literature review: The 13 quality indicators identified from literature search were: Assessment principles regarding criteria for setting pass marks, grade boundaries and allowed retakes; conflict of interest policy; assessment methods; Assessment of the learning domains; number and type of assessment per educational objective; number and type of assessment per instructional method; students meeting each educational outcome; students right of appeal against assessment results; feedback received by students: frequency, timing and nature; formative and summative assessments; assessment utility: validity and reliability; integrated learning; use of external examiners. Based on these indicators, a preliminary 34 items questionnaire was developed.

Delphi Process results: In Round 1, 10 out of 18 (n=56%) panelists returned the completed preliminary questionnaire. Based on the consensus results, perceived double barreled statements were simplified and new items were added for the Round two questionnaire forming a total of 58 items. After Round 2, 24 statements were rejected and two were merged, making a total of 34 statements for the round 3 questionnaire. After Round three, two items were deleted and three were merged. The final tool consisted of 30 items under four suggested domains as given in Table-I.
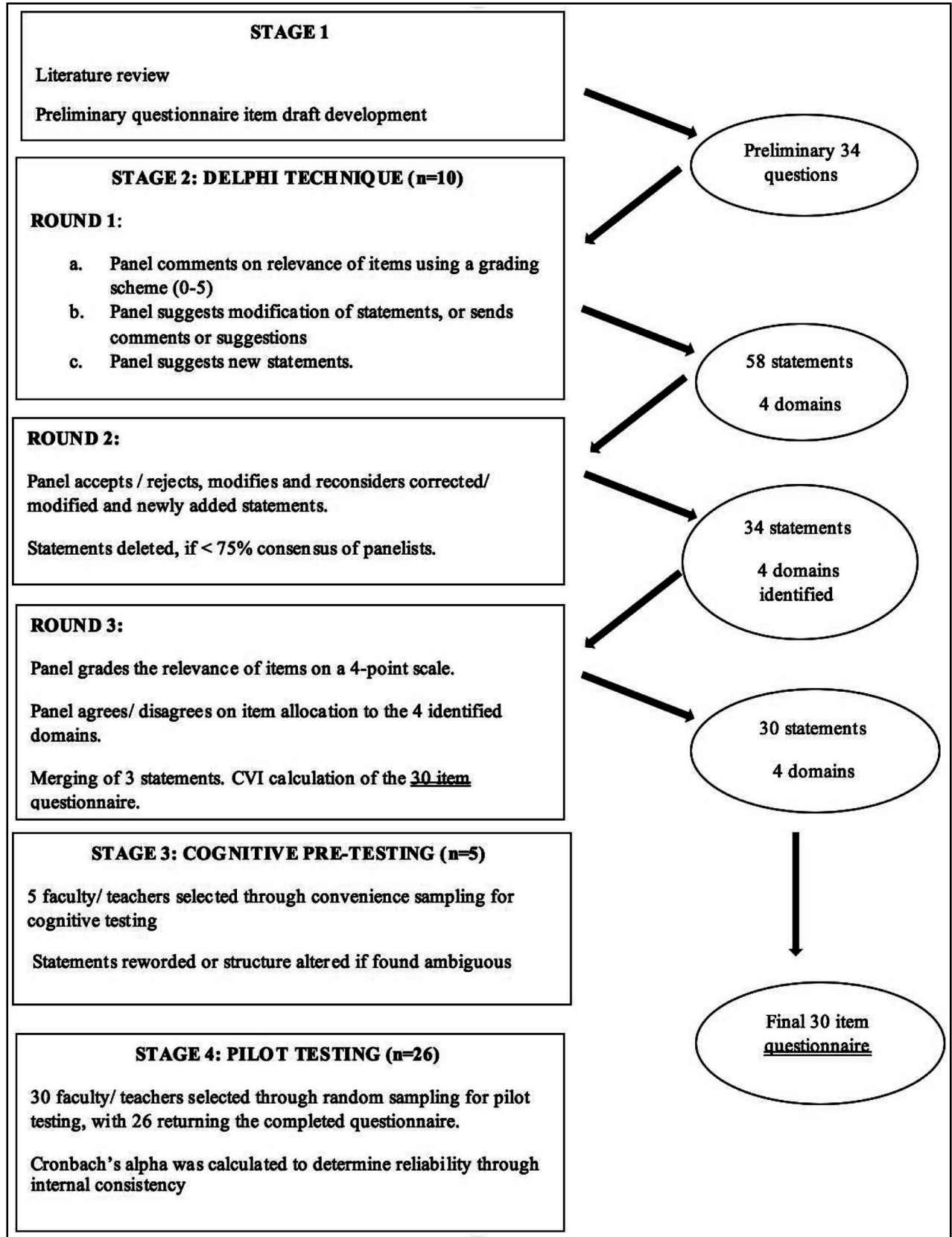
STAGE 1

Literature review

Preliminary questionnaire item draft development

Preliminary 34 questions

STAGE 2: DELPHI TECHNIQUE (n=10)

ROUND 1:

a. Panel comments on relevance of items using a grading scheme (0-5)
b. Panel suggests modification of statements, or sends comments or suggestions
c. Panel suggests new statements.

58 statements

4 domains

ROUND 2:

Panel accepts / rejects, modifies and reconsiders corrected/ modified and newly added statements.

Statements deleted, if < 75% consensus of panelists.

34 statements

4 domains identified

ROUND 3:

Panel grades the relevance of items on a 4-point scale.

Panel agrees/ disagrees on item allocation to the 4 identified domains.

Merging of 3 statements. CVI calculation of the 30 item questionnaire.

30 statements

4 domains

STAGE 3: COGNITIVE PRE-TESTING (n=5)

5 faculty/ teachers selected through convenience sampling for cognitive testing

Statements reworded or structure altered if found ambiguous

Final 30 item questionnaire

STAGE 4: PILOT TESTING (n=26)

30 faculty/ teachers selected through random sampling for pilot testing, with 26 returning the completed questionnaire.

Cronbach's alpha was calculated to determine reliability through internal consistency

Fig.1: Methodology flowchart.

Content validity index (CVI) calculation: Content Validity index of individual items (I-CVI) as well as the scale (S-CVI) was calculated. S-CVI was calculated with two methods, S-CVI/Avg and S-CVI/U, and is shown in Table-II.

Cognitive pre-testing: Cognitive pretesting of the questionnaire resulted in minor adjustments to the statements. A few technical terms like 'standard-setting' was considered ambiguous by some participants who considered 'pass/fail criteria' as the more relevant phrase for cognition.

Pilot testing for reliability of AIM tool: The response rate was 86% (n=26/30). Data was entered into SPSS 20, to calculate the reliability of the scale and its assigned domains. The calculated internal consistency for the composite score and all domains are given in Table-III.

DISCUSSION

The AIM tool was developed through an assorted methodology. A 'modified Delphi technique' was used rather than a focus group discussion forum, to assure respondent anonymity and to reduce unnecessary communication deterring from focussing on problem solving.[13]

For Delphi result analysis, the acceptable level of consensus needs to be determined beforehand. Different acceptable agreement levels between panelists are reported ranging from 51%-80%, or specified by stability of the response through the iterative process.[3] In our study, percentage responses along with the medians and ranges were calculated in Round 1. For subsequent rounds, we pre-specified a panel agreement of more than 75%

Table-I: AIM tool showing subdomains with allocated items.

Assessment policies
1   The medical school has a clearly defined assessment policy.
2   I have been oriented about the assessment policy in my college.
3   The procedures used for assessment of students are clearly laid down in assessment policy.
4   The criteria of student progression to next class are clearly documented.
5   The number of allowed exam retakes are clearly documented.
14   A system of appeal against assessment results is in place.
15   Assessments are open to scrutiny by external experts.
27   Standard setting is used to decide Pass/fail criteria before each individual summative assessment.

Assessment methods
6   The assessment methods used to assess knowledge component of course are appropriate for assessing the cognitive domain.
7   The assessment methods used to assess skill component of course are appropriate for assessing the psychomotor domain.
8   The assessment methods used to assess behavior component of course are appropriate for assessing attitude domain.
9   An appropriate weightage is given to knowledge, skills and attitude domains in assessments.
11   The assessment methods used are feasible.
16   Use of new assessment methods is encouraged, where appropriate.
23   Clear blueprints (table of specifications) are provided for each assessment.
25   Checklists or rubrics for performance assessments are clearly documented.

Purpose of assessment
17   The assessment system promotes student learning
18   Formative assessments are done at appropriate points during the curriculum to guide student learning.
19   There is an appropriate mix of formative and summative assessments.
20   The assessments encourage integrated learning by the students.
21   Feedback is given to students promptly after an assessment.

Assessment quality measures
10   Assessment system ensures that all assessments are conducted fairly
12   Adequate resources are provided for all assessments.
13   There is an adequate role of external examiners in summative examination.
22   Teachers are trained to provide feedback to students
24   Assessments adequately represent the exam blueprints (table of specifications).
26   There is an item bank which teachers contribute to and use for preparing exams.
28   Post examination item analysis is regularly conducted for summative assessments.
29   Post examination item analysis results are communicated to concerned departments.
30   Faculty development workshops are regularly conducted on various aspects of assessment.

Table-II: Content validity index calculation.

Item	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10	no. agreed	cvi-i
1	R	R	R	R	R	R	R	R	R	R	10	1.0
2	R	R	R	R	R	R	R	R	R	R	10	1.0
3	R	R	R	R	R	R	R	R	R	R	10	1.0
4	R	R	R	R	R	R	R	R	R	R	10	1.0
5	R	R	R	R	R	R	R	R	R	R	10	1.0
6	R	R	R	R	R	R	R	R	R	R	10	1.0
7	R	R	R	R	R	R	R	R	R	R	10	1.0
8	R	R	R	R	R	R	R	R	R	R	10	1.0
9	R	R	R	R	R	R	R	R	R	R	10	1.0
10	R	R	R	R	R	R	R	R	R	R	10	1.0
11	R	R	R	R	R	R	R	R	R	R	10	1.0
12	R	R	R	R	R	R	R	R	R	R	10	1.0
13	R	R	R	R	R	R	R	R	R	R	10	1.0
14	R	R	R	R	R	R	R	R	R	R	10	1.0
15	R	R	R	R	R	R	R	-	R	R	9	0.9
16	R	R	R	R	R	R	R	R	R	R	10	1.0
17	R	R	R	R	R	R	R	R	R	R	10	1.0
18	R	R	R	R	R	R	R	R	R	R	10	1.0
19	R	R	R	R	R	R	-	R	R	R	9	0.9
20	R	R	R	R	R	R	R	R	R	R	10	1.0
21	R	R	R	R	R	R	-	R	R	R	9	0.9
22	R	R	R	R	R	R	R	R	R	R	10	1.0
23	R	R	R	R	R	R	R	R	R	R	10	1.0
24	-	R	R	R	R	R	R	R	R	R	9	0.9
25	R	R	R	R	R	R	R	R	R	R	10	1.0
26	R	R	R	R	R	R	R	R	R	R	10	1.0
27	R	R	R	R	R	R	R	R	R	R	10	1.0
28	R	R	R	R	R	R	R	R	R	R	10	1.0
29	R	R	R	R	R	R	R	R	R	R	10	1.0
30	R	R	R	R	R	R	R	R	R	R	10	1.0
CVI values for 30 items questionnaire									Mean Expert proportion =0.98		S-CVI/Avg =0.98(30 items)	
	0.96	1.0	1.0	1.0	1.0	1.0	1.0	0.96	1.0	1.0		
											S-CVI/UA = 0.86	

R, Relevant items; -, Non relevant items; I-CVI , item-level content validity index; S-CVI/UA, scale-level content validity index/ universal agreement calculation method; S-CVI/Avg, scale-level content validity index/ average calculation method.

as a criterion for achievement of consensus. For round 3 final analysis, content validity index (CVI) of the individual items as well as that of the whole scale was also calculated using the ratings of item relevance by content experts. Good content validity of items is considered with I-CVIs of 1.00 with 3 to 5 experts and a minimum I-CVI of 0.78 with 6 to 10 experts. For scale level CVI, 0.90 or higher index is desired using the average calculating method and at least 0.80 is required using the universal agreement method, as it is more stringent in its approach.[14] In our study all the results were well above the desired range.

For 'Cognitive pre-testing' a respondent number of 10-30 or as few as 5-6 for a small scale research design, is considered sufficient.[12] We interviewed five faculty members using the 'Concurrent verbal probing method' as it eliminates the recall bias.[10]

Table-III: Cronbach's alpha for domains and full AIM tool.

Domains	No. of items	Cronbach's alpha
Assessment policies	8	0.78
Assessment methods	8	0.80
The purpose of assessment	5	0.67
The quality measures in assessment	9	0.73
Full questionnaire scale	30	0.915

To determine face validity, reliability and feasibility in certain large scale studies, pilot study is recommended.[1,3] For initial scale development, 30 representative participants from the population of interest is considered a reasonable minimum sample for pilot study, with a range from 25-40.[15] In our study, we selected 30 participants from the faculty. Cronbach's alpha was calculated for internal consistency of the tool and was calculated to be 0.9. The reported acceptable values of alpha, range from 0.70 to 0.95.[16]

Limitations: Construct validity could not be established because of small sample size.

CONCLUSION

The Assessment Implementation Measure (AIM) is a relevant and useful instrument to assess quality of assessment in undergraduate medical institutes. Further studies are needed for validation of AIM tool in variable contexts as well as its psychometric exploratory and confirmatory factor analyses.

REFERENCES

1. Hiong Sim J, Ting Tong W, Hong WH, Vadivelu J, Hassan H. Development of an instrument to measure medical students' perceptions of the assessment environment: initial validation. Med Educ Online. 2015;20(1):28612.
2. Tackett S, Grant J, Mmari K. Designing an evaluation framework for WFME basic standards for medical education. Med Teach. 2015;1–6. doi: 10.3109/0142159X.2015.1031737
3. Shehnaz SI, Premadasa G, Arifulla M, Sreedharan J, Gomathi KG. Development and validation of the AMEET inventory: An instrument measuring medical faculty members' perceptions of their educational environment. Med Teach. 2015;37(7):660–669. doi: 10.3109/0142159X.2014.947935
4. AlHaqwi AI, Kuntze J, van der Molen HT. Development of the clinical learning evaluation questionnaire for undergraduate clinical education: factor structure, validity, and reliability study. BMC Med Educ. 2014;14(1):44. doi: 10.1186/1472-6920-14-44
5. Rezaeian M, Jalili Z, Nakhaee N, Jahroomi Shirazi J, Jafari AR. Necessity of accreditation standards for quality assurance of medical basic sciences. Iran J Public Health. 2013;42:147-154.
6. Marrs H. Perceptions of College Faculty Regarding Outcomes Assessment. Int Electron J Leadership Learn. 2009;13(2).
7. World Federation for Medical Education. Basic Medical Education: WFME Global Standards for Quality Improvement. 2015. p. 1–64.
8. Sakai DH, Kasuya RT, Fong SF, Kaneshiro RA, Kramer K, Omori J, et al. Medical School Hotline: Liaison Committee on Medical Education Accreditation: Part I: The Accreditation Process. Hawai'i J Med Public Health. 2015;74(9):311.
9. Comittee on Accreditation of Canadian Medical Schools. CACMS standards and elements. 2015;1–25. Available from: https://www.afmc.ca/pdf/CACMS_Standards_and_Elements_June_2014_Effective_July12015.pdf
10. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Med Teach. 2014;36:463-474.
11. Gehlbach H, Brinkworth ME. Measure twice, cut down error: A process for enhancing the validity of survey scales. Rev Gen Psychol. 2011;15(4):380–387. doi: 10.1037/a0025704
12. Karabenick SA, Woolley ME, Friedel JM, Bridget V, Blazevski J, Bonney CR. Cognitive Processing of Self-Report Items in Educational Research : Do They Think What We Mean ? Cognitive Processing of Self-Report Items in Educational Research : Do They Think What We Mean ? Educ Psychol. 2007;1520:37–41. doi: 10.1080/00461520701416231
13. Kikukawa M, Stalmeijer RE, Emura S, Roff S, Scherpbier AJ. An instrument for evaluating clinical teaching in Japan: content validity and cultural sensitivity. BMC Med Educ. 2014;14(1):179. doi: 10.1186/1472-6920-14-179
14. Polit DF, Beck CT. The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. Res Nurs Health. 2006;29:489–497. doi: 10.1002/nur
15. Johanson GA, Brooks GP. Initial Scale Development: Sample Size for Pilot Studies. Educ Psychol Meas. 2010;70(3):394–400. doi: 10.1177/0013164409355692
16. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–55. doi: 10.5116/ijme.4dfb.8dfd

Authors' Contribution:

MS: Designed, did data collection, statistical analysis, manuscript drafting and editing.
RAK: Conceived, designed, statistical analysis, editing and critical review of manuscript.
RY: Did critical review and editing.