Advances in Medical Education and Practice

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Psychometric Properties of Visual Indicators of Teaching and Learning Success "VITALS" Instrument for Evaluation of Clinical Teachers

Nada Al-Yousuf[1]
Salah Eldin Kassab[2,3]
Hasan Alsetri[4]
Hossam Hamdy[3]

[1]Department of Ophthalmology, King Abdullah Medical City, Manama, Kingdom of Bahrain; [2]Department of Basic Medical Sciences College of Medicine, Gulf Medical University, Ajman, United Arab Emirates; [3]Department of Surgery, Gulf Medical University, Ajman, United Arab Emirates; [4]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA

Purpose: We have previously developed an instrument for students' evaluation of clinical teachers that we called Visual Indicators of Clinical Teaching and Learning Success (VITALS). This study measures the reliability of VITALS as an instrument for student evaluation of clinical tutors. Additionally, the study explores the minimum number of student raters necessary for an acceptable reliability, and provides evidence of construct validity of the evaluation scores.

Materials and Methods: This retrospective study included 1825 evaluation forms completed by medical students evaluating clinical tutors using the VITALS instrument. Reliability was measured by applying generalizability theory (G-theory) analysis using a two-facet design (raters and items). A D-study was used to determine the minimum number of raters required to achieve a reliability ≥0.80. Face validity was tested by measuring tutors' degree of agreement with the items of the study instrument.

Results: The overall G-coefficient was 0.89. The subject of measurement (clinical tutors' scores) represented 15.8% of the variance across all tutors and items. The variance due to the interaction between raters (students) and tutors contributed to 43.5%, while the variance due to items was negligible. The remaining 40% of the variance was due to unexplained sources of error. The D-study demonstrated that a minimum of 12 raters (students) are required to achieve a reliability of 0.80. Finally, most of the clinical tutors agreed that all items in the instrument were appropriate.

Conclusion: We demonstrate that VITALS exhibits good psychometric properties. There should be at least 12 students rating each clinical tutor to have an acceptable level of reliability for the study instrument. Face validity of the study instrument was evidenced by its high level of approval among clinical tutors.

Keywords: clinical teaching, tutors' evaluation, reliability, validity, generalizability theory

Introduction

Clinical teaching is one of the important components of medical students' education. The quality of clinical teaching delivered by clinical faculty is an essential component of medical students' learning experience. The assessment of clinical teachers is often based on questionnaires completed by students. However, it is important that these questionnaires have good psychometric properties to reflect the quality of teaching by clinical faculty.[1]

Students describe the ideal clinical teacher as being a positive role model. The positive role model has been defined as being competent clinically and personally, while exhibiting quality teaching abilities. These role models were further described

Correspondence: Nada Al-Yousuf
King Abdullah Medical City, 61, King Abdulaziz Avenue, Manama, Kingdom of Bahrain
Tel +973 77310071
Fax +973 77310001
Email nyousuf10@gmail.com

Received: 5 May 2021
Accepted: 28 July 2021
Published: 16 August 2021

as being compassionate, supportive, just assessors, that provide constructive feedback and guidance, while providing opportunities for students to get involved in patients' care.[2,3] In addition, students valued active research, clear instruction, organization, and enthusiasm as core traits possessed by the ideal instructor.[4]

Clinical teachers also play an important role in teaching values and professionalism.[4] Indeed, one of the unique qualities of clinical teachers is scaffolding, which is an instructor's awareness of the level of their students, providing additional guidance according to their deficiencies.[5] Moreover, clinical teachers should promote articulation, which is when students are encouraged to be explicit about their knowledge and skills.[5]

Student evaluations are an important component of the evaluation of clinical teachers in medical education. It was shown that student evaluation of clinical teachers is reliable, and correlates well with student learning, peer evaluation and tutor self-rating.[6] Indeed, clinical tutors can get feedback directly from their students on their performances in multiple settings, which can be used to enhance the quality of teaching. Moreover, such feedback provides useful data for administrators to use. The feedback can be utilized to aid in tenure selection, promotions, faculty development programs and for allocating clinical teaching responsibilities.[7]

There are many published papers that study instruments developed for evaluating clinical tutors.[5,7–9,17,18,21–23] Examples include: Maastricht Clinical Teaching Questionnaire (MCTQ),[5] Cleveland Clinic Teaching Effectiveness Instrument (CCTEI),[7] and System for the Evaluation of Teaching Qualities (SETQ).[8] We have previously introduced the Visual Indicators for Teaching and Learning Success (VITALS) instrument which uniquely has a smaller number of items compared to the previously published methods for clinical teacher evaluation.[10] VITALS is based on the theoretical framework of cognitive apprenticeship for clinical teaching. In addition, VITALS can visually display the evaluations of clinical teachers graphically, which simplifies the interpretation of the data output.

Psychometric properties of evaluation methods include the assessment of both reliability and validity of score interpretations from these instruments. Validity is often divided into distinct subcategories such as face, content, and criterion validity. Emerging conceptualizations consider construct validity as a term that encompasses all the validity subcategories.[11] In this paradigm, construct validity represents the degree to which score interpretations from the psychometric instrument are used to provide evidence to support or refute the underlying construct. The sources of evidence include the content (do the instrument items represent the intended construct), relations to other variables (correlations with the scores from other instruments), internal structure (acceptable reliability and factor structure), response process (relations between the underlying construct and the thought process of subjects), and consequences (whether assessment scores make a difference or not).[11] Face validity represents subjective judgement of how well the items represent appropriate operationalization of the measured construct based on its face value.[12] On the other hand, reliability or consistency of the scores from one test to another is considered a requisite for, but not sufficient evidence of validity. Generalizability theory provides an integrated framework for measuring the various forms of reliability. Although we previously reported some evidence of validity for the VITALS instrument,[10] the other sources of evidence for validity including generalizability of the scores were not tested. After the widespread use of this instrument, especially in many other medical colleges in the Gulf region, we thought to provide further evidence of the psychometric properties of a study instrument. The findings from the current study could establish the evidence of its utility and stimulate further testing of the instrument in other contexts.

This study is conducted to answer the following research questions:

1. What is the reliability of student evaluation scores of clinical tutors using VITALS across raters and items?
2. What is the minimum number of student raters of clinical tutors that yields an acceptable reliability of the VITALS instrument?
3. What is the evidence of construct validity of the students' evaluation scores of clinical tutors using the VITALS instrument?

Materials and Methods

Study Context

This is a retrospective study measuring the validity and reliability of students' evaluation of clinical tutors using the VITALS instrument. The study was conducted at the College of Medicine and Medical Sciences, Arabian Gulf University (AGU) in Bahrain. The College of Medicine program of AGU

is six years long. It is composed of three phases: Phase I (year 1), Phase 2 or pre-clerkship (years 2, 3 and 4), and Phase 3 or clerkship (years 5 and 6). This study was conducted in 2015 during the clerkship phase of the program, which is composed of hospital-based clinical rotations of the following major disciplines: Medicine, Surgery, Pediatrics, Obstetrics and Gynecology. It was approved by the "Research and Ethics Committee" of Arabian Gulf University. A total of 25 students evaluated each clinical tutor. Inclusion criteria for the reliability study are: clinical tutors teaching in the clerkship phase with complete evaluations and no missing values. Out of 100 clinical tutors teaching at the clerkship phase, 73 were eligible for the reliability study. This produced a total of 1825 complete evaluation forms. For the face validity study, 72 tutors responded from the 100 tutors questioned.

Description of the VITALS Instrument

We have previously demonstrated the development of the VITALS instrument.[10] The evaluation form used by medical students to evaluate clinical tutors is composed of 10 items, with a 4-point Likert scale for each item. The maximum score is 4 and the minimum score is 1. All Likert scale item responses of "strongly disagree" or "disagree", were combined as "disagree", while all "strongly agree" or "agree" responses were combined as "agree". Clinical tutors are informed of their respective collective ratings in the form of a horizontal bar graph showing the teacher's strong points and weak points displayed at the end of the academic year.[10]

Face Validity of the VITALS Instrument

A questionnaire was distributed to tutors measuring their general acceptance of VITALS. The questionnaire was composed of 8 items. Each item was rated on a 1 to 4 response scale, where 1 = strongly disagree, 2 = disagree, 3 = agree and 4 = strongly agree. The items in the questionnaire were: 1) Tutors to be evaluated by medical students using VITALS. 2) Tutors should receive results of this evaluation. 3) Such evaluations are important to improve teaching. 4) Evaluations should be shown to academic administrators 5). Academic administrators should discuss this evaluation with clinical tutors. 6) This evaluation should inform decisions such as promotion. 7) It can be considered for tutors' contract renewal. 8) Medical students can judge tutors.

Reliability of the VITALS Instrument

Reliability was measured using statistical models based on generalizability theory (G-theory). We have previously described in detail the assessment of reliability using G-theory analysis.[11,12] The object of measurement in the current study was the tutor's evaluation scores by students. The design of the G-theory analysis included two facets (raters and items) with raters nested within tutors, and both crossed with items. We selected a design where items were fixed at 10, while the raters were considered random, as we were interested in generalizing the findings beyond the context used in this study. We used the generalizability coefficient (Φ), as we were interested in the relative inferences for inter-individual comparison of tutors' performance. A G-study was conducted to estimate sources of error in the tutors' ratings and to determine reliability of students with a different set of raters. A D-study was used to determine the reproducibility of these ratings, and to identify the minimum number of raters and items required to achieve reliability of ≥ 0.80. We have also reported the standard error of measurement (SEM) which represents the standard deviation of all the errors of measurement in the study.

The statistical software package GENOVA was used for the G-theory analysis. Data from the questionnaires from tutors and students analyzed using SPSS- version 16. Descriptive statistics were calculated based on the percentages of those who agreed and those who disagreed with each item in the form and in the questionnaire. Strongly agree and agree anchors were considered as "agree", and strongly disagree and disagree were considered as "disagree"

Results

Generalizability Study (G Study)

The overall G-coefficient of the VITALS instrument scores across the two facets of the study (25 student raters and 10 items) was 0.89. In addition, the percent variance attributed to the study subjects (object of study) was 15.8%. Because of the nested design of the study, we were not able to determine the percent variance due to the rater's facet. However, a large percent of variance (43.5%) was found due to the interaction between raters (students) and tutors. On the other hand, the percent variance due to items (0.2%) was negligible, and the variance due interaction between tutors and items was zero. This indicates that the relative ordering of tutors' scores did not differ when tested on different items. Finally, the interactions between tutors, raters and items represented 40.5% of total variance. This component represents both the variance attributable to the three-way interaction, and the variance ascribed to source of error (facets) that were unmeasured in the study. Table 1.

Table 1 Generalizability Study Showing the Variance Components and Their Percentages for Tutors' Evaluation by Students Using the VITALS

Facets and Interactions	df	Variance Component	% Variance	Standard Error
Tutors	72	0.171	15.80	0.031
Items	10	0.002	0.20	0.001
Raters: Tutor	1752	0.471	43.50	0.017
Tutor × Item	720	0.000	0.00	0.000
Rater× Item: Tutor	17,520	0.438	40.50	0.004
G coefficient = 0.89				
Absolute SEM = 0.58				

Notes: The proportion of observed variance explained by each facet is calculated by dividing the individual variance component by the total observed variance.
Abbreviations: df, degree of freedom; SEM, standard error of measurement; G Coefficient, Generalizability Coefficient.

Decision Study (D Study)

Figure 1 demonstrates the predicted changes in the G-coefficient by the combined effects of increasing the number of items and raters. As shown in the figure, changing the number of items used in the evaluation form, does not appear to influence the G-coefficient. However, G-coefficient increased as number of raters (students) increased. With 5 raters, the G-coefficient was 0.631, at 10 raters it was 0.774, at 15 raters it increased to 0.837 and became 0.872 at 20 raters. Generalizability coefficient was increased to 0.91, if the number of raters is increased to 30, and reaching to as high as 0.923 with 35 raters.

Evidence of Face Validity of VITALS

The response rate for the clinical tutors to their perception about VITALS was 72% (72 out of 100 tutors). Results indicated that 94% of tutors agreed with students' evaluation of tutors using the VITALS; 97% of the responding tutors agreed that clinical tutors should receive their evaluation by medical students; 93% agreed that such evaluation is important to improve teaching; 90% agreed that an evaluation should be discussed with tutors; 89% agreed that the evaluation should be shown to administrators; 78% agreed that such evaluation can be considered for tutors' promotion; 67% agreed that students are capable of judging their clinical tutors; and 65% agreed that the evaluation should be taken into consideration for tutors' contract renewal. Figure 2.

Discussion

This study is an extension to our previously published paper on developing and applying VITALS in reporting

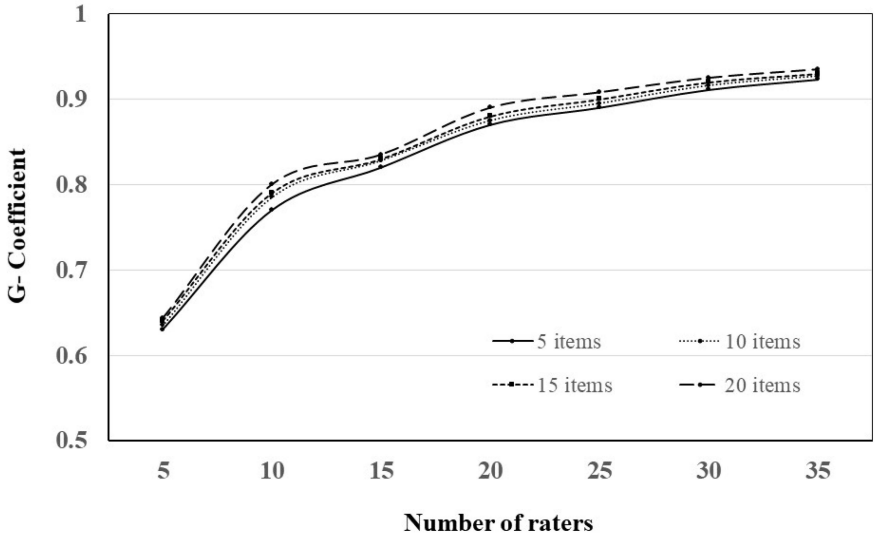

Figure 1 An example of the application of visual indicators for teaching and learning success (VITALS) of a clinical tutor evaluated by students (Reproduced with modifications from Hamdy et al, 2001).

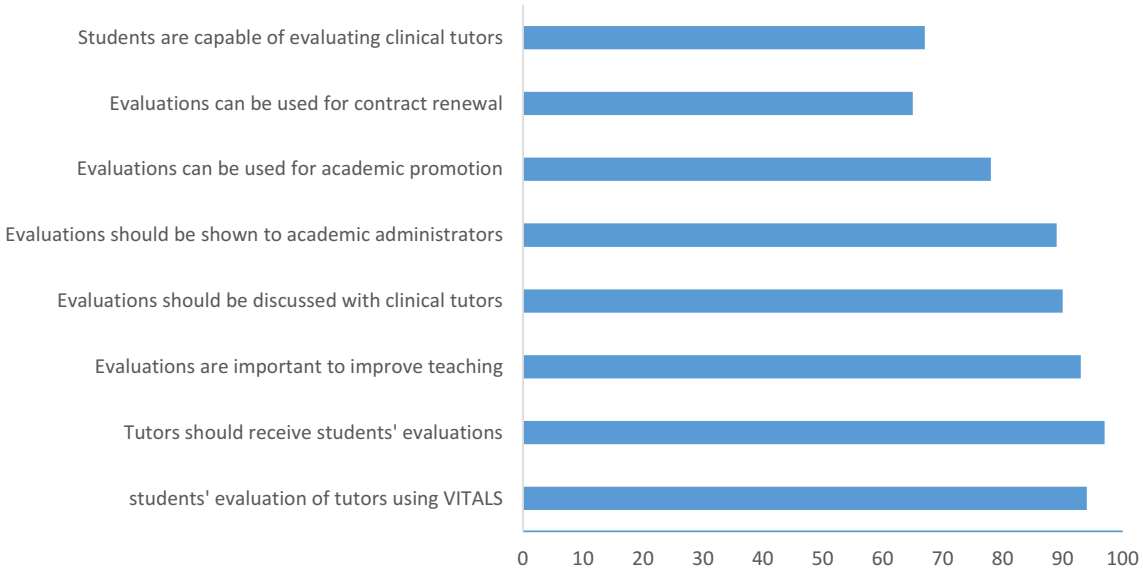

Figure 2 Decision study (D study) results for the evaluation scores of clinical tutors (n = 73) using VITALS instrument. The D-study estimated the projected G-coefficient using different numbers of student raters (from 5 to 35 raters) and number of items in the VITALS instrument (from 5 to 20 items).

evaluations of clinical teachers.[10] We have demonstrated in that study an evidence of content and face validity of the instrument by using a descriptive qualitative approach. In the current study, we evaluated the reliability and face validity of the instrument using a quantitative approach. We have demonstrated in this study that the VITALS instrument can be used for students' evaluation of clinical tutors with a good reliability. The decision study indicated that the main determinant for the changes in reliability was the number of student raters, while the items of the instrument did not contribute to the variance. In addition, we showed that at least 12 students are required to achieve a dependable estimate of the evaluation of clinical tutors. Furthermore, the study revealed that a large percent of variance is attributable to unmeasured sources of error, which opens the venue to explore other facets in future studies. The majority of clinical teachers were positive and agreed with the importance of VITALS as an instrument for the evaluation of clinical teaching, demonstrating evidence of face validity. This could be explained as a subjective judgment about the appropriate operationalization of the measured construct.[12]

The variance ascribed to the subject of measurement (clinical tutors) indicates that averaging over raters and items, clinical tutors differed somewhat systematically in their evaluation scores. These findings indicate an acceptable degree of variability in students' ratings of clinical tutors due to unsystematic sources of error. This relatively small percent of variance indicates that a larger percent of variance is attributed to different sources of error. Results of the present study demonstrated that students nested in tutors contributed 43.5% to variance components. Obviously, the nested model design of the study did not allow determining the percentage of variance imputed to differences between raters from the variance imputed to the interaction between raters and items independently.

The finding of the zero variance for the facet of item indicates that the score obtained on a particular item is representative of scores obtained on all similar items of the construct. In another way, it is an evidence of the internal consistency of the VITALS instrument. In addition, the interaction between items and raters was also zero, suggesting that the rank order of tutors did not change significantly across items, and there were small changes in rating behavior across items.

Gillmore et al, who used generalizability theory to study reliability of scores provided by students to tutors, found that increasing the number of students had a much greater impact on G-coefficient than increasing the number of items.[15] Mazor et al used decision study to measure the effect of changing the number of items on generalizability coefficient. They also demonstrated that increasing the number of items from five to ten did not have significant impact on the generalizability coefficient.[17] Interestingly, Mazor et al, who used the application of generalizability theory to students' ratings of tutors, also found that the

largest percentage of variance (45%) was associated with students (raters) nested within tutors. Furthermore, they have also found that relatively little variability (2%) was associated with the item facet, ie, students tended to rate a given tutor the same on all or most items.[17]

There are several published instruments for evaluation of clinical teachers with different types of evidence for their validity.[5,7–9,17–25] The G-theory analysis indicated good reliability of the scores from Student Evaluation of Clinical Teaching Questionnaire (SETQ),[22] the Maastricht Clinical Teaching Questionnaire (MCTQ),[5] and The Cleveland Clinic's Clinical Teaching Effectiveness Instrument.[7] However, the advantage of the VITALS instrument is the relatively smaller number of items compared with previously published instruments, and the visual appeal of the provided information enhances its impact, and increases the likelihood that the information will be transformed into action.[10]

The results of the D-study indicated that at least 12 student raters are required to achieve an acceptable reliability of the VITALS instrument, irrespective of the number of items. Increasing the number of student raters from 5 to 12 students resulted in significant improvement to achieve an acceptable dependability coefficient of 0.8. However, any increase above 12 students did not result in striking changes in the dependability coefficient. This message is important for academic administrators when taking this evaluation into consideration for promotion or contract renewal. Previous studies have reported inconsistent findings regarding the minimum number of raters for achieving an acceptable reliability.[5,7,15,21] The number of raters of these studies ranged from a minimum of two or three in some studies,[7,21] seven raters in a different study,[5] and fifteen raters in another study.[21] Differences between these studies and the current one could be related to the structure of the study instrument, and the context of the study.

This study has some limitations that are worth reporting. Firstly, applying the VITALS instrument has been restricted to one medical school in the clerkship rotations. Further studies will be required to test generalizing the results of the current study to other medical schools. Secondly, although we have provided an evidence of reliability and face validity of the instrument, other sources of validity evidence such as predictive validity and criterion-related validity need to be determined. Future studies need to examine the relationship between other measures of performance for clinical tutors, and their evaluations using the VITALS instrument. Finally, the large percent of variance due to the unmeasured error raises the flag for future studies using larger sample sizes to examine other relevant facets in the G-theory analysis. One important facet could be the occasion where temporality of the evaluation scores using the VITALS could be tested.

Conclusion

Results of this study provide an additional evidence to the construct validity of using VITALS for clinical tutors' evaluation by medical students in the clerkship phase. We demonstrated an evidence of reliability of the VITALS instrument taking into consideration the errors due to raters and items. In addition, there should be at least 12 students rating each clinical tutor in order to have an acceptable level of reliability for the study instrument. Finally, we have provided a quantitative evidence of face validity for the study instrument by the high level of agreement for using the instrument by clinical tutors. Results of this study may guide administrators when using evaluation results in supporting certain decisions such as promotion, retention, identification of tutors deserving recognition for teaching excellence, and general documentation of teaching quality. Additional studies are required before generalizing the use of this instrument in other medical schools.

Disclosure

The authors report no conflicts of interest and no financial interest in this work.

References

1. Fluit C, Bolhuis S, Grol R, et al. Assessing the quality of clinical teachers. *J Gen Intern Med*. 2010;25(12):1337–1345. doi:10.1007/s11606-010-1458-y
2. Bazrafkan L, Hayat A, Tabei S, Amirsalari L. Clinical teachers as positive and negative role models: an explanatory sequential mixed method design. *J Med Ethics Hist Med*. 2019;12:1–15.
3. Stenfors-Hayes T, Hult H, Dahlgren L. What does it mean to be a good teacher and clinical supervisor in medical education? *Adv Health Sci Educ Theory Pract*. 2011;6(2):197–210. doi:10.1007/s10459-010-9255-2
4. Doumouras A, Rush R, Campbell A, Taylor D. Peer-assisted bedside teaching rounds. *Clin Teach*. 2015;12(3):197–202. doi:10.1111/tct.12296
5. Stalmeijer R, Dolmans D, Wolfhagen I, et al. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. *Acad Med*. 2010;85(11):1732–1738. doi:10.1097/ACM.0b013e3181f554d6
6. Pritchard R, Watson M, Kelly K, Paquin A. *Helping Teachers Teach Well: A New System for Measuring and Improving Teaching Effectiveness in Higher Education*. San Francisco, Ca: New Lexington Press; 1998.
7. Copeland H, Hewson M. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic center. *Acad Med*. 2000;75:161–166. doi:10.1097/00001888-200002000-00015

8. Boerebach B. Evaluating clinicians' teaching performance. *Perspect Med Educ*. 2015;4(5):264–267. doi:10.1007/s40037-015-0215-7
9. Jochemsen-van der Leeuw R, Van Dijk N, Wieringa-de Waard M. Assessment of the clinical trainer as a role model: a Role Model Apperception Tool (RoMAT). *Acad Med*. 2014;89:671–677. doi:10.1097/ACM.0000000000000169
10. Hamdy H, Williams R, Tekian A, et al. Application of "VITALS": visual indicators of teaching and learning success in reporting student evaluations of clinical teachers. *Educ Health*. 2001;14(2):267–276. doi:10.1080/13576280110051064
11. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7–16. doi:10.1016/j.amjmed.2005.10.036
12. Drost EA. Validity and reliability in social science research. *Educ Res Perspect*. 2011;38:105–123.
13. Kassab S, Fida M, Radwan A, Bakri A, Abu-Hijleh M, O'Connor B. Generalizability theory analyses of concept mapping assessment scores in a problem-based medical curriculum. *Med Educ*. 2016;50:730–737. doi:10.1111/medu.13054
14. Kassab S, Du X, Toft E, et al. Measuring medical students' essential professional competencies in a problem-based curriculum: a reliability study. *BMC Med Educ*. 2019;19:155. doi:10.1186/s12909-019-1594-y
15. Gillmore G, Kane M, Naccarato R. The generalizability of student ratings of instruction: estimation of the teacher and course components. *J Educ Meas*. 1978;15:1–13. doi:10.1111/j.1745-3984.1978.tb00051.x
16. Kreiter C, Ferguson K, Lee W, et al. Generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Acad Med*. 1998;73:1294–1298. doi:10.1097/00001888-199812000-00021
17. Mazor K, Clauser B, Cohen A, Alper E, Pugnaire M. The dependability of students' ratings of preceptors. *Acad Med*. 1999;74:S19–S21. doi:10.1097/00001888-199910000-00028
18. Al-Ansari A, Strachan K, Hashim S, Otoom S. Analysis of psychometric properties of the modified SETQ tool in undergraduate medical education. *BMC Med Educ*. 2017;17:1–9. doi:10.1186/s12909-017-0893-4
19. Al-Ansari A, Arekat M, Salem A. Validating the modified system for evaluation of teaching qualities: a teaching quality assessment instrument. *Adv Med Educ Pract*. 2018;9:881–886. doi:10.2147/AMEP.S181094
20. Drieling K, Montano D, Poinsting L. Evaluation in undergraduate medical education: conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching. *Med Teach*. 2017;39(8):820–827. doi:10.1080/0142159X.2017.1324136
21. Wormley M, Romney W, Greer A. Development of the clinical teaching effectiveness questionnaire in the United States. *J Educ Eval Health Prof*. 2017;14:14. doi:10.3352/jeehp.2017.14.14
22. Boerebach B, Lombarts K, Arah O. Confirmatory factor analysis of the system for evaluation of teaching qualities (SETQ) in graduate medical training. *Eval Health Prof*. 2016;39(1):21–32. doi:10.1177/0163278714552520
23. Van der Leeuw R, Lombarts K, Heineman MJ, Arah O. Systematic evaluation of the teaching qualities of obstetrics and gynecology faculty: reliability and validity of the SETQ tools. *PLoS One*. 2011;6:1–7. doi:10.1371/journal.pone.0019142
24. Zuberi R, Bordage G, Norman R. Validation of the SETOC instrument – student evaluation of teaching in outpatient clinics. *Adv Health Sci Educ*. 2007;12:55–69. doi:10.1007/s10459-005-2328-y
25. Beckman T, Mandrekar J. The interpersonal cognitive and efficiency domains of clinical teaching: construct validity of a multi-dimensional scale. *Med Educ*. 2005;39(12):221–1229. doi:10.1111/j.1365-2929.2005.02336.x