Using Measurement Invariance to Explore the Source of Variation in Basic Medical Science Students' Evaluation of Teaching Effectiveness

Mahmoud Alquraan[1,2,*], Sulaf Alazzam[2] and Hakam Alkhateeb[3]

[1]*Al Ain University, Al Ain, UAE*

[2]*College of Education, Yarmouk University, Irbid, Jordan*

[3]*College of Medicine, Yarmouk University, Irbid, Jordan*

Abstract: *Introduction*: Many research studies have shown that students' evaluations of teaching(SET) are affected by different variables without testing the requirement of fair comparisons. These studies have not tested the measurement equivalency of SET surveys according to these variables. Measurement equivalency of SET refers to whether a SET survey is interpreted similarly across different groups of individuals (Variable Levels). Without evidence of measurement invariance across different variables under investigation, the SET ratings should not be compared across these variables and this is the goal of this study.

Methods: Measurement Invariance analysis of SET survey was investigated using 1649 responses to SET of four different medical core courses offered by the College of Science and College of Medicine and from different levels.

Results: The results showed the existence of teaching practices in the SET survey that are not equivalently loaded on its factor across the levels of targeted variables, and the college offered medical courses were a source of variation in basic medical science students' evaluation of teaching effectiveness. On the other hand, teaching practices in the SET survey are equivalently loaded on its factor across course levels.

Discussion: The study results showed that the SET of medical courses is comparable to the courses only taught by the College of Medicine. These results provide evidence that medical courses are different from other courses offered by other colleges. This means that comparing SET of the College of Medicine with other colleges and colleges of medicine needs to compare SET results at the college level only.

Keywords: Basic science medical students, student evaluation of teaching (SET), Measurement Invariance (MI), Health Education, Medicine Program.

INTRODUCTION

Universities should always use different indicators to monitor their institutional performance. In a large number of universities, student evaluation of teaching (SET) is the only measure and indicator of teaching effectiveness [1]. SET is a valuable source of feedback given by students to their teachers [2] and its relationship with students' learning requires universities to give serious attention to SET results [3]. SET is an important indicator of quality required by accreditation agencies which may be the reason behind using it in most universities across the world [4].

SET results have been used for important decisions related to faculty members and academic programs, such as accreditation, tenure, faculty members' promotion, and faculty members' contract renewal [5-7]. Students' feedback is an effective source of data about learning effectiveness [8-10]. A meta-analysis study done by Wright and Jenkins-Guarnieri provides evidence that SET can be used to improve the quality of teaching and to increase student achievement too, [11] assuming that students learn more from professors who receive higher ratings [12].

SET is an internal quality control process for all higher education institutions that are facing challenging academic environments and looking for an effective teaching methodology to enhance the learning process [3].

As it is important to take medical students' point of view in designing and implementing new innovations in medical education, [13] the results of SET are useful in adapting the teaching material and teaching methods in important decisions related to the faculty's development and promotions [14] Effective medical education relies on various factors, including the characteristics and competencies of the medical teacher, the integration of technology in the instructional process, and the teacher's ability to connect medical issues with social and historical events [15-16].

Previous research on variables affecting SET was categorized into three groups: the characteristics of the course itself and the learning environment; the attributes of the students including their perceptions and attitudes; and the qualities and attributes of the teacher [17]. According to Almakadma *et al.* factors considered by students when filling out SETs are language proficiency, easier exams, fewer lectures, decreased number of slides per lecture, giving clues to students about exams, being committed to class times

*Address correspondence to this author at the Al Ain University, Al Ain, UAE; E-mail: mahmoud.alquraan@aau.ac.ae

and schedule, being lenient about attendance, being lenient about class discipline, better teaching skills, better appearance (physical, clothing), being responsive and open to student feedback and suggestions, tutor relationship with the student (extracurricular, research, same ethnicity, personal relationship, etc.) [18]. In the same context, Haris *et al.* used a conceptual framework of the SET tool, which includes three factors that affect student's perception of faculty teaching evaluation: students related factors (level of seriousness, grade, honesty, and understanding), factors related to the teachers (personality, relationship, gender, age, qualification, knowledge, teaching methods, language), and factors related to the learning environment (class timing, class capacity) [19].

Constantinou & Wijnen-Meijer provide an overview of some of the variables that influence the scores of SETs which are: Low attention, lack of time and low response rates, Anonymity, Course difficulty, and grade expectation, Course type, course organization and motivation, Gender bias, and attractiveness, Issues of reliability [3]. While Lawrence mentioned that there are several factors unrelated to teaching effectiveness that can influence SET scores, such as the instructor's race, age, gender, and physical attractiveness, additional variables include students' grade expectations, enjoyment of the class, and even the weather conditions when the survey is conducted [20]. On the other hand Singh *et al*. determined a categories of the Characteristics of effective medical teachers: Classroom behavior/ Instructional delivery (good communication skills, good presentation skills, good sense of humor in teaching sessions, innovative in using technology in the classroom, well organized and possess excellent time management skills (good planner), inflexible regarding maintaining classroom discipline), Interaction with students/colleagues (aware of students' interests and needs, easily approachable/affable, not encouraging student's participation during theory lecture classes, work well with colleagues and administrators, Inspiring & motivational to students, very generous in assessing the performance of the students during exams, offer constructive criticism to the students, trust and respect the students, caring and shows empathy towards students), Personal qualities (leadership qualities, punctual, unbiased, have sound knowledge of subject, enthusiastic and has passion to teach, enjoys teaching, honest, moral & ethical), Professional development (up-to-date with the recent advancements in education technology, have publications and should be active in research, learning and open to change (Flexible)) [21].

Evaluating teachers poses a significant challenge, particularly in the context of medical schools, as it should be fair, and objective and ultimately contribute to the enhancement of the educational process. Unfortunately, the absence of a robust evaluation culture within our society hinders the recognition of teachers who consistently demonstrate high-quality teaching performance [22].

Medical education differs from other higher education curricula. The curriculum structure requires students to take predefined courses rather than making individual choices. Some teaching formats, like in-patient or bedside teaching, are unique to medical education. SET survey items are the most important teaching practices that are under investigation by the university administration. These teaching practices should be understood by the students in the same way despite all variables. The existence of variables that influence the understanding of these items differently means the explanations and uses of SET results may not be valid. Psychometrically, the procedure of examining the validity of SET is called testing measurement invariance (MI) [23].

Many research studies have shown that SETs are affected by different variables. Examples of these variables are academic achievement [24], student gender [25]; students academic college [26], instructor variables [27], course characteristics [28], and teaching methods [29]. These studies have not tested the measurement equivalency of SET surveys according to these variables. Measurement equivalency of SET refers to whether a SET survey is interpreted in the same way across different groups of individuals (Variable Levels). Without evidence of measurement equivalency across different variables under investigation, the SET survey may be susceptible to measurement errors. Psychometrically, this process is called testing Measurement invariance (MI).

Testing MI of SET survey items means investigating how these teaching practices are related to some variables. In the setting of this study, MI is supported if the quantitative relationships of the teaching practices included in the SET survey are equal across different medical courses, the academic college of the medical courses' teachers, and the level of the medical courses. If MI is supported, then the students who responded to the SET survey across all of these variables interpret the teaching practices (SET items) and the teaching effectiveness in the same way, and the relationships between these variables and SET can be investigated [30-31].

Asparouhov and Muthén recommended the use of multi-group confirmatory variable analysis (MGCFA) to test MI [32]. van de Schoot *et al.* and Putnick and Bornstein recommended several steps to assess MI

using MGCFA [33]. These steps start with a confirmatory factor analysis (CFA) model for each level of the under investigation variables separately, so the construct validity of the SET -in this study- can be evaluated using a number of model global fit indices. This stage is called configural invariance model testing. This step tests whether the SET has the same pattern of free and fixed loadings of SET items across the subgroups of targeted variables. Invariance at the configural level means that the structure of the SET is established in all the levels of targeted variables (i.e., teacher of the medical course from the college of medicine or college of science). Configural noninvariance on the other hand, indicates that the loadings pattern of the SET items on the SET theoretical construct differ across the levels of targeted variables (i.e., teacher of the medical course from the College of Medicine or College of Science) [33-31].

If configural invariance is established, the next step is to assess the metric invariance or equivalence of the item loadings. Metric invariance indicates that each teaching practice in the SET survey contributes to the SET theoretical structure to a similar degree across all the levels of targeted variables. Metric invariance is evaluated by constraining loadings (i.e., the loadings of the teaching practices on the SET) to be equivalent across all the levels of targeted variables. The metric model is then compared to the configural invariance to assess the fit. If the fit is significantly worse in the metric invariance model compared to the configural invariance model, it means that at least one teaching practice is not equivalent across the levels of targeted variables, and metric invariance is not established. If the fit is not statistically worse, it means that constraining the loadings across all the levels of targeted variables does not statistically affect the model fit, and metric invariance is established.

If metric invariance is established, the next step is to assess scalar invariance or equivalence of SET teaching practices intercepts. Scalar invariance is tested by constraining the teaching practices' intercepts to be equivalent across the levels of targeted variables. The scalar invariance model is compared to the metric invariance model to determine fit. If the model fit is statistically worse, it indicates that at least one teaching practice intercept differs across the levels of targeted variables, and scalar invariance is not established. If the fit is not statistically worse, it means that constraining the item intercepts across the levels of targeted variables does not statistically affect the model fit, and scalar invariance is established.

Model fit is evaluated using different global fit indices. In this study, the model fit indices which are suggested by Rutkowski and Svetina were used [34]. Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and standardized root mean squared residual (SRMR). CFI and TLI compare the fit of targeted models, and they should be > .90. SRMR should be < .08. For model comparisons, changes in the fit indices are insignificant if the CFI and TLI change more than 0.010, and the SRMR change more than 0.03 [35].

Few studies investigated MI in SET using different grouping variables. Bazán-Ramírez *et al.* evaluated MI of the teaching performance of the psychology course according to gender, age, and academic stage. The measurement invariance was not supported based on these variables [36]. Another study by Kalender and Berberoğlu evaluated MI in SET between high and low-achieving students, which showed that MI is not supported. Besides that, reviewing the related literature shows the absence of studies investigating MI on BMS SET. The study is trying to contribute to the existing literature by investigating this issue [37].

Most of the conducted studies that have aimed to test the differences in SET of BMS students according to course levels have not tested the MI of SET across the students' academic years or levels. To have a fair compression, measurement invariance must be tested and held [38]. Therefore, this study aims to test the measurement invariance (MI) of the SET questionnaire according to the college of the processors of the course.

Study Rationale and Aim

Most of the studies that tested the effect of different variables on SET have not taken MI into account. Violations of full MI make comparing groups' mean differences misleading [39]. Medical education courses differ from other higher education curricula. The structure of the medical study plan requires students to take predefined courses rather than making individual choices. This could raise a question about the comparability of SET between the College of Medicine and other colleges. Walsh raised a very important question "Medical education research: is participation fair" [40]. This question might be extended from depending heavily on medical students during most research on medical education to the fairness of the SET surveys in such research. This study aims to examine the MI in SET differences according to the Academic College of the teacher and Course level, and testing means differences in SET when it is fair to make these comparisons.

METHODOLOGY

Sample

The sample of this study represents all the students registered in four medical core courses, and each

Table 1: Sample Distribution According to Students' Academic College

Course Name	College of Teacher	Number of Students	Percent
Organic Chemistry	Science	369	22.4
General anatomy	Medicine	350	21.2
Circulatory system	Medicine	385	23.3
Nervous system 2	Medicine	545	33.1
Total		1649	100.0

course represents the year level of the BMS course (student level). Sample distribution to courses is presented in Table **1**.

SET Survey

The data analyzed in this study were provided to the authors by Yarmouk University. The University uses a survey that was developed and approved by the university to assess teaching effectiveness at the university level. It is a self-reported survey that is distributed at the end of each semester. This survey consists of 20 Likert-type items distributed to four factors: planning, instruction, management, and assessment teaching practices; each factor is measured by five items. The reliability and validity of this SET survey are presented in the results section of this study.

Statistical Analysis

To achieve the goals of the study, the following analyses were used:

1. Four Factor Confirmatory Factor Analysis (CFA) was conducted to assess the construct validity of the SET survey.
2. MI was used to assess the Course (Organic chemistry and three medical courses) as a source of variation in the SET survey.
3. MI was used to assess the Academic College of the teacher (Medical and Science Colleges) as a source of variation in the SET survey.
4. MI was used to assess the level of the medical courses offered by the College of Medicine (Years 1-3) as a source of variation in the SET survey.
5. When MI was supportive, mean differences were tested.

RESULTS

First, four-factor CFA was used for the full dataset (Model_ 0). Then, the three MI models were performed using MGCFA. The global fit indices of these models are shown in Table **2**. Using Chen's cut-off values, the content of Table **2** shows that Model_0 has a good model fit ($\mathrm{CFI_{All}} = 0.946$, $\mathrm{TLI_{All}} = 0.937$, $\mathrm{SRMR_{All}} = 0.029$, $\mathrm{RMSEA_{All}} = 0.107$) [35]. These results were evidence of SET construct validity.

The first source of variation tested in this study was the BMS medical courses (Four medical courses as shown in Table **1**), and the first tested model in MI was the Configural model. The fit indices of the Configural model, which is presented in Table **2**, reached the accepted level, and the same factorial structure holds across all BMS medical courses ($\mathrm{CFI_{Configural}} = 0.925$, $\mathrm{TLI_{Configural}} = 0.937$, $\mathrm{SRMR_{Configural}} = 0.033$, $\mathrm{RMSEA_{Configural}} = 0.131$). The configural invariance model was used as a reference model to compare the fit of metric invariance.

The results presented in Table **2** show that the fit indices of the Model 0- CFA and Model 1-configural have (Full dataset) reached the cut-off values suggested by Rutkowski and Svetina (TLI and CFI > 0.90. and SRMR < 0.08) [41]. This provides evidence of SET construct validity. The fit of Model 2- metric requires comparing Model 1- configural and Model 2-

Table 2: Fit Indices of the MGCFA for Evaluating MI Models According to Medical Courses Offered by the College of Medicine and College of Science (n=1649)

Model	χ^2	df	CFI*	TLI	RMSEA	SRMR	Δ CFI	Δ SRMR
Model 0- CFA	3270.44	164	0.946	0.937	0.107	0.029	--	--
Model 1- configural	5283.77	656	0.925	0.914	0.131	0.033	--	--
Model 2- metric	5910.15	704	0.916	0.909	0.134	0.092	0.009	0.059
Model 3- Scalar	6058.25	752	0.914	0.914	0.131	0.093	0.002	0.001

TLI: Tucker-Lewis Index, CFI: Comparative Fit Index, SRMR: Standardized root mean squared residual, RMSEA: Root mean square error of approximation.

Table 3: Fit Indices of the MGCFA for Evaluating MI Models According to The Teacher Academic College (Medicine and Science)

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	Δ CFI	Δ SRMR
Model 1- configural	3315.77	328	0.951	0.943	0.105	0.027	--	--
Model 2- metric	3921.60	344	0.941	0.935	0.112	0.087	0.010	0.060
Model 3- Scalar	4007.95	360	0.940	0.936	0.111	0.087	0.001	0.000

TLI: Tucker-Lewis Index, CFI: Comparative Fit Index, SRMR: Standardized root mean squared residual, RMSEA: Root mean square error of approximation.

metric fit indices. Δ CFI equals 0.009, and Δ SRMR equals 0.059. Based on Chen's criterion [35], metric invariance is not supported.

Since the model fit was significantly worse in the metric invariance model, this indicates that at least one teaching practice in the SET survey was not equivalently loaded on its factor across the levels of targeted variables, and the medical courses are a source of variation in BMS students' evaluation of teaching effectiveness. This indicates that the meaning and understanding of teaching practices were different from one course to another. The medical courses included in this study were offered by two different colleges (One offered by the College of Science: Organic Chemistry- Level 1, and three offered by the College of Medicine: General anatomy- level 1, Circulatory system- level 2, and Nervous system 2- level 3). This means that the teachers of these courses from two colleges and the MI according to this variable were tested, and the results are presented in Table **3**. Moreover, the three medical courses offered by the College of Medicine were from three different levels (First, second, and third year of BMS). Also, the MI according to this variable was tested, and the results are presented in Table **4** to find out the source of variation in SET of BMS students.

The second source of variation tested in this study was the college of the faculty members that offer the BMS medical courses (College of Medicine and College of Science), and the first tested model in MI was the Configural model. The fit indices of the Configural model, which is presented in Table **3**, reached the accepted level, and the same factorial structure holds across all colleges ($\mathrm{CFI}_{\mathrm{Configural}} = 0.951$, $\mathrm{TLI}_{\mathrm{Configural}} = 0.943$, $\mathrm{SRMR}_{\mathrm{Configural}} = 0.027$, $\mathrm{RMSEA}_{\mathrm{Configural}} = 0.105$). The configural invariance model was used as a reference model to compare the fit of metric invariance.

The results presented in Table **3** show that the fit indices of the Model 1- configural have reached the cut-off values suggested by Rutkowski and Svetina (TLI and CFI > 0.90. and SRMR < 0.08) [41]. This provided evidence of SET construct validity. The fit of Model 2- metric requires comparing Model 1- configural and Model 2- metric fit indices. Δ CFI equaled 0.01, and Δ SRMR equaled 0.06. Based on Chen's criterion, metric invariance was not supportive [35]

Since the model fit was significantly worse in the metric invariance model, this indicates that at least one teaching practice in the SET survey was not equivalently loaded on its factor between the two colleges, and the college that offered the medical courses was a source of variation in BMS students' evaluation of teaching effectiveness. This indicated that the meaning and understanding of teaching practices of the faculty members from the College of Medicine are different from those of the faculty members from the College of Science.

The third source of variation tested in this study was the medical course levels offered by the College of Medicine. First, four-factor CFA was used for the college of medicine courses only dataset (Model_ 0). Then, the three MI models were performed using MGCFA. The global fit indices of these models are shown in Table **4**. Using Chen's cut-off values [35], the content of Table **2** shows that Model_0 has a good model fit ($\mathrm{CFI}_{\mathrm{Medical}} = 0.958$, $\mathrm{TLI}_{\mathrm{Medical}} = 0.952$,

Table 4: Fit Indices of the MGCFA for Evaluating MI Models According to Course Levels Offered by the College of Medicine

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	Δ CFI	Δ SRMR
Model 0- CFA	2278.44	164	0.958	0.952	0.100	0.013	--	--
Model 1- configural	4246.45	492	0.928	0.917	0.134	0.020	--	--
Model 2- metric	4264.93	524	0.928	0.922	0.129	0.022	0.000	0.002
Model 3- Scalar	6058.25	752	0.928	0.926	0.126	0.023	0.000	0.001

TLI: Tucker-Lewis Index, CFI: Comparative Fit Index, SRMR: Standardized root mean squared residual, RMSEA: Root mean square error of approximation.

$\text{SRMR}_{\text{Medical}} = 0.013$, $\text{RMSEA}_{\text{Medical}} = 0.100$). These results were evidence of SET construct validity at the College of Medicine level.

The fit indices of the Configural model, which is presented in Table **4**, reached the accepted level, and the same factorial structure holds across all colleges ($\text{CFI}_{\text{Configural}} = 0.928$, $\text{TLI}_{\text{Configural}} = 0.917$, $\text{SRMR}_{\text{Configural}} = 0.020$, $\text{RMSEA}_{\text{Configural}} = 0.134$). The configural invariance model was used as a reference model to compare the fit of metric invariance.

The fit of Model 2- metric requires comparing Model 1- configural and Model 2- metric fit indices. Δ CFI equals 0.00, and Δ SRMR equals 0.002. Based on Chen's criterion [35], metric invariance is supportive.

As the MI was established, the estimated difference between SET means based on the level of the medical course (First, second, and third level of BMS medical courses) was fairly tested. SET means of the three levels are $\overline{\text{X}}_{\text{Level1}} = 78.93$, $\overline{\text{X}}_{\text{Level2}} = 62.16$, and $\overline{\text{X}}_{\text{Level3}} = 61.48$. ANOVA testing of the differences between the means showed that $\text{F}_{2,1277} = 57.48$, and it was significant ($p < 0.001$). This means that there is a true difference -since MI is supported- in BMS students' evaluation of teaching effectiveness according to the course level. The result suggests that teachers who teach first-year level medical courses were evaluated higher ($\overline{\text{X}}_{\text{Level1}} = 78.93$) when compared with teachers who teach second-year and third-year level courses ($\overline{\text{X}}_{\text{Level2}} = 62.16$ and $\overline{\text{X}}_{\text{Level3}} = 61.48$).

DISCUSSION AND CONCLUSIONS

MI in SET of medical courses was evaluated in this study according to the following variables: Course, the college of the offered courses, and the level of the courses offered by the College of Medicine. Testing MI according to these variables is a prerequisite for investigating the effect and the relationship between SET of medical courses and these variables.

As shown in Table **2**, Δ CFI equals 0.009, and Δ SRMR equals 0.059. Based on Chen's criterion [36], metric invariance was not supported, and the scalar model was not tested. Since the MI metric model of SET was not supported, it means that students from different courses had responded to the teaching practices in different ways so that the differences in the teaching practices cannot be compared across courses, as comparing the mean differences of theoretical constructs across groups requires scalar invariance [42], which was the case in this study. This indicated that the meaning and understanding of teaching practices are different from one course to another, and the source of variation in SET of the included courses in this study was different from one course to another.

The four medical courses included in this study were offered by two colleges (Medicine and Science), and the previous results showed a source of variation in the SET of these courses. MI was assessed according to college. The results presented in Table **3** showed that metric invariance was not supported, and college was a source of variation in the SET of BMS students. This indicated that at least one teaching practice in the SET survey was not equivalently loaded on its factor between the two colleges, and the college that offered the medical courses was a source of variation in BMS students' evaluation of teaching effectiveness. This indicated that the meaning and understanding of teaching practices of the faculty members from the College of Medicine are different from those of the faculty members from the College of Science.

The third source of variation that was tested in this study was the medical course levels offered by the College of Medicine (n=1280). The four-factor CFA results in Table **4** provided evidence of SET construct validity at the College of Medicine level. Also, the results supported full MI of SET of the three levels of the courses offered by the College of Medicine. This gave us the ability to compare SET across the level of medical courses. SET means of the three levels were $\overline{\text{X}}_{\text{Level1}} = 78.93$, $\overline{\text{X}}_{\text{Level2}} = 62.16$, and $\overline{\text{X}}_{\text{Level3}} = 61.48$. The result suggested that teachers who teach first-year level medical courses were evaluated higher ($\overline{\text{X}}_{\text{Level1}} = 78.93$) when compared with teachers who teach second-year and third-year level courses ($\overline{\text{X}}_{\text{Level2}} = 62.16$ and $\overline{\text{X}}_{\text{Level3}} = 61.48$).

The study results showed that the SET of medical courses was comparable according to the College that offered the courses, and the SET of medical courses is comparable according to the course levels. Students who studied the year one level course have higher SET compared with the students who studied second and third year level courses. This suggested the need for the College of Medicine to conduct an internal review to know the reasons behind giving first-level courses higher SET and benefit from these results to develop teaching practices at the college level.

REFERENCES

[1] Hande H, Kamath S, D'Souza J. Students' perception of effective teaching practices in a medical school. Education in Medicine Journal 2014; 6(3): 63-66. https://doi.org/10.5959/eimj.v6i3.247

[2] Boerebach B. Evaluating clinicians' teaching performance. Perspectives on Medical Education 2015; 4(5): 264-267. https://doi.org/10.1007/S40037-015-0215-7

[3] Constantinou C, Wijnen-Meijer M. Student evaluations of teaching and the development of a comprehensive measure of teaching effectiveness for medical Schools. BMC Medical Education 2022; 22(113): 1-12. https://doi.org/10.1186/s12909-022-03148-6

[4] Dodeen H. Validity, Reliability, and Potential Bias of Short Forms of Students' Evaluation of Teaching: The Case of UAE University. Educational Assessment 2013; 18(4): 235-250. https://doi.org/10.1080/10627197.2013.846670

[5] Abdallah A, Balla B. Students' Evaluation of Teaching Effectiveness: Level of Acceptance, Implementation, and Causes for Concern (A Case Study of Saudi Faculty Members at Jeddah University-Kholais Branch). International Journal of English Language Teaching 2022; 10(3): 24-36. https://doi.org/10.37745/ijelt.13/vol10no2pp.24-36

[6] Pan G, Shankararaman V, Koh K, Gan S. Students' evaluation of teaching in the project-based learning programme: An instrument and a development process. The International Journal of Management Education 2021; 19(2): 100501. https://doi.org/10.1016/j.ijme.2021.100501

[7] Kogan LR, Schoenfeld-Tacher R, Hellyer PW. Student evaluations of teaching: perceptions of faculty based on gender, position, and rank. Teaching in Higher Education 2010; 15(6): 623-636. https://doi.org/10.1080/13562517.2010.491911

[8] Park E, Dooris J. Predicting student evaluations of teaching using decision tree analysis. Assessment & Evaluation in Higher Education 2020; 45(5): 776-793. https://doi.org/10.1080/02602938.2019.1697798

[9] Urrutia-Aguilar M, Sánchez-Mendiola M, Guevara-Guzmán R, Martínez-González A. Comprehensive Assessment of Teaching Performance in Medical Education. Procedia - Social and Behavioral Sciences 2014; (141): 252-259. https://doi.org/10.1016/j.sbspro.2014.05.044

[10] Schönrock-Adema J, Boendermaker P, Remmelts P. Opportunities for the CTEI: disentangling frequency and quality in evaluating teaching behaviours. Perspectives on Medical Education 2012; (1): 172-179. https://doi.org/10.1007/S40037-012-0023-2

[11] Wright S, Jenkins-Guarnieri M. Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. Assessment & Evaluation in Higher Education 2012; 37(6): 683-699. https://doi.org/10.1080/02602938.2011.563279

[12] Uttl B, White C, Gonzalez D. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation 2016; (54): 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

[13] Oudkerk Pool A, Jaarsma A, Driessen E, Govaerts M. Student perspectives on competency-based portfolios: Does a portfolio reflect their competence development? Perspectives on Medical Education 2020; (9)1: 66-172. https://doi.org/10.1007/S40037-020-00571-7

[14] Müller T, Montano D, Poinstingl H, *et al*. Evaluation of large-group lectures in medicine – development of the SETMED-L (Student Evaluation of Teaching in MEDical Lectures) questionnaire. Müller *et al*. BMC Medical Education 2017; (17): 137. https://doi.org/10.1186/s12909-017-0970-8

[15] Ahmed M. Are good attributes of medical teachers more important than the learning style: a glimpse into the future of medical education and learning. Journal of Public Health and Emergency 2018; (2). https://doi.org/10.21037/jphe.2018.05.01

[16] Engbers R, de Caluwé L, Stuyt P, Fluit C, Bolhuis S. Towards organizational development for sustainable high-quality medical teaching. Perspectives on Medical Education 2013; (2): 28-40. https://doi.org/10.1007/S40037-013-0043-6

[17] Worthington A. The Impact of Student Perceptions and Characteristics on Teaching Evaluations: A case study in finance education. Assessment & Evaluation in Higher Education 2002; 27(1): 49-64. https://doi.org/10.1080/02602930120105054

[18] Almakadma A, Fawzy N, Baqal O, Kamada S. Perceptions and attitudes of medical students towards student evaluation of teaching: A cross-sectional study. Medical Education Online 2023; (28)1: 2220175. https://doi.org/10.1080/10872981.2023.2220175

[19] Haris S, Jamil B, Haris M, Deeba F, Khan MJ, and Khan IZ. Factors Affecting Students Perception towards Faculty Evaluation of Teaching at Nowshera Medical College. The Professional Medical Journal 2022; 29(2): 258-264. https://doi.org/10.29309/TPMJ/2022.29.02.6407

[20] Lawrence J. Student Evaluations of Teaching are Not Valid American Association of University Professors. Epub 2018.

[21] Singh S, Pai D, Sinha N, Kaur A, Soe H, Barua A. Qualities of an effective teacher: what do medical teachers think? BMC Medical Education 2013; 13(1). https://doi.org/10.1186/1472-6920-13-128

[22] Urrutia-Aguilar M, Sánchez-Mendiola M, Guevara-Guzmán R, Martínez-González A. Comprehensive Assessment of Teaching Performance in Medical Education. Procedia - Social and Behavioral Sciences 2014; (141): 252-259. https://doi.org/10.1016/j.sbspro.2014.05.044

[23] Dimitrov D. Testing for factorial invariance in the context of construct validation. Measurement and Evaluation in Counseling and Development 2010; 43(2): 121-149. https://doi.org/10.1177/0748175610373459

[24] Sánchez T, Gilar-Corbi R, Castejón J, Vidal J, León J. Students' evaluation of teaching and their academic achievement in a higher education institution of Ecuador. Frontiers in Psychology 2020; 11(233). https://doi.org/10.3389/fpsyg.2020.00233

[25] Boring A. Gender biases in student evaluations of teachers. Journal of Public Economics 2017; 145(13): 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

[26] Chen G, Watkins D. Stability and correlates of student evaluations of teaching at a Chinese university. Assessment & Evaluation in Higher Education 2010; 35(6): 675-685. https://doi.org/10.1080/02602930902977715

[27] Wolbring T, Riordan P. How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. Social Science Research 2016; 5(7): 253-272. https://doi.org/10.1016/j.ssresearch.2015.12.009

[28] Harnish R, Bridges K. Effect of syllabus tone: Students' perceptions of instructor and course. Social Psychology of Education 2011; 14(3): 319-330. https://doi.org/10.1007/s11218-011-9152-4

[29] Park B, Cho J. How does grade inflation affect student evaluation of teaching? Assessment & Evaluation in Higher Education 2022. https://doi.org/10.1080/02602938.2022.2126429

[30] Krammer G, Pflanzl B, Lenske G, Mayr J. Assessing quality of teaching from different perspectives: Measurement invariance across teachers and classes. Educational Assessment 2021; 26(2): 88-103. https://doi.org/10.1080/10627197.2020.1858785

[31] Van de Schoot R, Lugtig P, Hox J. A checklist for testing measurement invariance. European Journal of Developmental Psychology 2012; 9(4): 486-492. https://doi.org/10.1080/17405629.2012.686740

[32] Asparouhov T, Muthén B. Multiple-group factor analysis alignment. Structural Equation Modeling: A Multidisciplinary Journal 2014; 21(4): 495-508. https://doi.org/10.1080/10705511.2014.919210

[33] Putnick D, Bornstein M. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review 2016; (41): 71-90. https://doi.org/10.1016/j.dr.2016.06.004

[34] Rutkowski L, Svetina D. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educational and Psychological Measurement 2014; 74(1): 31-57. https://doi.org/10.1177/0013164413498257

[35] Chen F. Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal 2007; 14(3): 464-504. https://doi.org/10.1080/10705510701301834

[36] Bazán-Ramírez A, Pérez-Morán J, Bernal-Baldenebro B. Criteria for teaching performance in psychology: invariance according to age, sex, and academic stage of Peruvian students. Frontiers in Psychology 2021; (12): 4816. https://doi.org/10.3389/fpsyg.2021.764081

[37] Kalender I, Berberoğlu G. The measurement invariance of University students' ratings of instruction. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi 2019; 34(2): 402-417.

[38] Putnick D, Bornstein M. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review 2016; (41): 71-90. https://doi.org/10.1016/j.dr.2016.06.004

[39] Pokropek A, Davidov E, Schmidt P. A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal 2019; 26(5): 724-744. https://doi.org/10.1080/10705511.2018.1561293

[40] Walsh K. Medical education research: is participation fair? Perspectives on Medical Education 2014; (3): 379-382. https://doi.org/10.1007/S40037-014-0120-5

[41] Rutkowski L, Svetina D. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educational and Psychological Measurement 2014; 74(1): 31-57. https://doi.org/10.1177/0013164413498257

[42] Steinmetz H. Analyzing observed composite differences across groups. Methodology 2013; 9(1): 1-12. https://doi.org/10.1027/1614-2241/a000049

Received on 28-09-2023 Accepted on 21-10-2023 Published on 10-11-2023

https://doi.org/10.6000/1929-6029.2023.12.23