ORIGINAL ARTICLE
Volume 12 Issue 1 2020
DOI: 10.21315/eimj2020.12.1.3
ARTICLE INFO
Submitted: 28-12-2019
Accepted: 03-02-2019
Online: 10-04-2020

Using Item Response Theory (IRT) to Assess Psychometric Properties of Undergraduate Clinical Education Environment Measure (UCEEM) among Medical Students at the Faculty of Medicine, Suez Canal University

Sally Fouad, Shimaa El Araby, Rabab Abdel Ra'oof Abed, Mohamed Hefny, Moustafa Fouad

Faculty of Medicine, Suez Canal University, EGYPT

To cite this article: Fouad S, El Araby S, Abed RAR, Hefny M, Fouad M. Using item response theory (IRT) to assess psychometric properties of undergraduate clinical education environment measure (UCEEM) among medical students at the Faculty of Medicine, Suez Canal University. Education in Medicine Journal. 2020;12(1): 15–27. https://doi.org/10.21315/eimj2020.12.1.3

To link to this article: https://doi.org/10.21315/eimj2020.12.1.3

ABSTRACT

Undergraduate Clinical Education Environment Measure (UCEEM) has been used as a reliable and valid tool to evaluate clinical workplace, and it might be used for further purposes such as benchmarking and evaluating different clinical context. Thus, we aim to examine psychometric properties of UCEEM by using item response theory (IRT). This study is a cross-sectional field survey, with an explorative component of psychometrics conducted at the Faculty of Medicine, Suez Canal University. We used IRT which emphasises the fact that an individual's response to a questionnaire item is influenced by qualities of both the individual and the item. The results indicate that there are four factors obtained by exploratory factor analysis (EFA) form a reliable hypothetical model, and the goodness fit indices of the first order confirmatory factor analysis (CFA) indicated a good fit. The test characteristic curve (TCC) in IRT gives us information about the expected score in the questionnaire based on the level of agreement (ability = θ), e.g., the expected score is 42 if the level of agreement was 0. Based on study results, it was evident that the UCEEM questionnaire has a high reliability and acceptable evidence of construct validity to use it for further purposes.

Keywords: *Learning environment, Exploratory item analysis, Item response theory*

CORRESPONDING AUTHOR Rabab Abdel Ra'oof Abed, Medical Education Department, Faculty of Medicine, Suez Canal University, Round Road, Ismailia, Egypt | Email: dr.rababraoof@gmail.com

INTRODUCTION

Authentic Early Experience (AEE) means early contacts with patients and health system, but it differs in each educational phase depending on the learner's level of performance and the complexity of the clinical context, their participation may range from just observation to real contribution to practice. Participation leads to "real patient learning", a term that describes the processes and very immediate consequences of interaction between a learner and a patient, facilitated by a practitioner (1). Early experience might direct medical curricula towards the

social context of practice, help in students' transition to the clinical environment, motivate them, make them more confident to deal with patients, and enhance self-awareness and awareness of others. In addition, it might make their theoretical knowledge stronger, deeper and more contextualised, and improves their learning of behavioural and social sciences (2).

Undergraduate Clinical Education Environment Measure (UCEEM) represents a valid, reliable and feasible multidimensional instrument for the evaluation of the clinical workplace as a learning environment for the undergraduate medical student, it has the potential to become a valuable tool for benchmarking and evaluation of clinical learning climates in various contexts, in this regard further validation is needed with different population, psychometric methods and source of evidence (3).

Thus, authors used item response theory (IRT) which is a psychometric approach emphasising the fact that an individual's response to a particular questionnaire item is influenced by qualities of the individual and by qualities of the item for measuring this important construct. Thus, IRT provides procedures for obtaining information about individuals, items and tests (4).

As a matter of fact, one of the most useful aspects of IRT is its ability to construct scales that are short but still reliable and valid, whether researchers do that de novo for a new scale, or by eliminating redundant items from existing ones (5). In addition, there are other ways where IRT is helpful, such as adaptive testing. As can be seen, people are all accustomed to taking tests, not only in class but also to get admitted to a graduate or a professional school, where they began with items that were so easy that they were laughable, and then became so difficult that they feared for our future. The reason is that the test had to span the entire range of abilities of the test-takers, from those who have trouble filling in their name to the know-it-alls. Nevertheless, items that are far below or far above a person's ability level give us no useful information (6).

The same is true for questionnaires—items that tap low levels of the trait are useless for people who have a lot of it, and vice versa. The most useful items are those near the middle, and these will vary from one person to the next. In adaptive testing, the test administrator takes a guess at the person's ability level or selects an item whose level is near zero. If this item is passed, then it is unnecessary to give easier items; conversely, if it is failed, it would only frustrate the person to give more difficult ones. So, by choosing items judiciously, only a small proportion of all potential items need to be given. Because all of the items have a value along the same continuum, people can be compared with one another even though they may each have taken a different subset of questions (5).

Moreover, IRT produces a variety of data displays, encapsulating both student and item properties that enable test developers to monitor and improve the quality of test questions (6). Equally important the core assumption of IRT is that the probability of a student's answering a test question correctly depends on the examinee's underlying ability with regard to the trait being measured and on the statistical characteristics of the test item. Further, this relationship between the probability of answering the question correctly and the examinee's ability can be described by a mathematical function called an item characteristic curve (ICC) (7).

Before using IRT models in psychometric process, there are basic assumptions must be met. The first assumption is unidimensionality, only a single ability is measured by the items that make up the test. Covariance among the items can be explained by a single underlying dimension. This assumption is sometimes not met when cognitive; personality and test-taking factors might affect test performance. The unidimensionality of a scale can be evaluated by performing an item-level factor analysis, designed to evaluate the factor structure (8).

A second assumption is local independence. This means that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent. Responses for different items are not related. An item does not provide any clue to answer another item correctly. If local dependence does exist, a large correlation between two or more items can essentially affect the latent trait and it causes lack of validity. Such questions should be eliminated and not adequate to estimate an examinee's ability accurately (8).

IRT measurement can be a useful tool, but like all tools, it must be used properly or more harm than good may result. The assumptions for IRT measurement, especially those of unidimensionality and local independence, must be met for successful application of IRT models to real test data. Both assumptions are empirically testable using various correlation techniques, but to carry out these analyses successfully, sufficiently large representative samples of students must be available (9).

To carry out this IRT scaling analysis, an appropriate mathematical model must be selected: a model that can be empirically demonstrated to fit the data and meets all of the required assumptions. In this regard there are three IRT models commonly used for tests that are scored as "right" or "wrong" (i.e., as 0 or 1, or dichotomously). These models are named for the number of parameters they use to estimate examinee ability (10).

As an illustration, the One-Parameter IRT model is also known as the Rasch model, after its originator. The Rasch model uses only a single parameter, item difficulty; to estimate item and student characteristics. The one-parameter IRT model is widely used throughout the world in many different medical education settings as it requires the fewest number of examinees (11).

In the way two other IRT models, the two-parameter and the three-parameter models, are also widely used, especially for large-scale assessments. The two-parameter model adds an item discrimination parameter (in addition to item difficulty) and the three-parameter model adds a "guessing" parameter (pseudo-chance) to item difficulty and item discrimination. (The "guessing" parameter accounts for the probability of arriving at the correct answer, in a selected-response question, by chance alone.) The two- and three-parameter models typically fit large-scale one-dimensional achievement data well (10). In fact, all IRT models attempt to explain observed (actual) item performance as a function of an underlying ability (unobserved) or latent trait (11).

The objectives of this research are to examine the different types of validity evidence and reliability of the students' scores from the instrument used for measuring experiential learning, which is UCEEM questionnaire using IRT.

METHODS

This study is a cross-sectional field survey, with an explorative component of psychometrics. The sampling technique was a non-probability convenience sampling. The sampling frame was all the undergraduate students (from Year 1 to Year 6) at the Faculty of Medicine, Suez Canal University who were approached to participate in the study. The sample size was depended on the response rate of the students in the study. The sample of this study was 550 from all the undergraduate students (from Year 1 to Year 6) at the Faculty of Medicine, Suez Canal University. It is sufficient to evaluate the validity and reliability of the UCEEM questionnaire using graded response model (GRM) (one of the Polytomous IRT models). Each student received an informed consent as separate section at the beginning of the questionnaire. Students were informed that their participation in this study is optional and will not affect their score in anyway. In addition, they have the right to withdraw at any

time, and finally information will be kept confidential.

The proposal of this study was approved by research and ethics committee at Faculty of Medicine, Suez Canal University. The UCEEM, a 25-item instrument with two overarching dimensions (experiential learning and social participation) and four subscales (Figure 1), represents a valid, reliable and feasible multidimensional instrument for the evaluation of the clinical workplace as a learning environment for the undergraduate medical student. The instrument had not only proven useful in current quality improvement projects and studies of undergraduate clinical environments in the Swedish context, but also has the potential to become a valuable tool for benchmarking and evaluation of clinical learning environment in various contexts (3).

In this regard, the current study is testing the psychometric prosperities of UCEEM questionnaire through exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and reliability analysis followed by IRT analysis. GRM is the model used in this study specifies each step function using the 2PL whereby there is a common value for all *m* steps of the *ith* item, and a separate *bik* parameter for each step of the item. An appealing property of the GRM is that the item characteristics curve (ICC) can be obtained directly from the difference between adjacent step functions (12).

RESULTS

First: Exploratory Factor Analysis

Checking the suitability of data for factor analysis

Sample size: Sample size is 550 participants which is adequate for factor analysis.

Factorability of the correlation matrix: The correlation matrix reveals statistically significant, moderate correlations among the observed variables used in the analysis. None of the correlation coefficients are large; therefore, there is no need to eliminate any variables at this stage.

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity: This test revealed that the KMO measure of sampling adequacy was 0.93 (superb). This value indicates that there were sufficient items predicted by each factor. Furthermore, Bartlett's test of sphericity was statistically significant ($p < 0.001$) which indicates that the variables were significantly correlated shown in Table 1. Therefore, this output indicated the appropriateness of the data for factor analysis.

Table 1: KMO measure of sampling adequacy and Bartlett's test of UCEEM questionnaire

KMO measure of sampling adequacy		0.933
Bartlett's test of sphericity	**Approx. chi-square**	5094.314
	df	300
	Sig.	.000

Extraction of factors

Principal component analysis with varimax rotation was performed to identify and interpret the number of factors that could explain most of the common variance and to remove non-reflective or redundant items. The results revealed that the 25 items of the UCEEM questionnaire resulted in four factors with an eigenvalue >1.00. The four factors that emerged from the factor analysis accounted for 51.34% of the total variance as shown in Table 2. The number of factors was also confirmed with the visual inspection of the scree plot that indicated a sudden drop in the scree beginning with the fourth factor as shown in Figure 1. Finally, the questionnaire contained four factors and 25 items as shown in Table 2. Factor 1 included 8 items, Factor 2 included 11 items, Factor 3 included 2 items and Factor 4 included 4 items.

The four factors were labelled as follows:

Factor 1: Preparedness for student entry and engagement

The items map out perceptions of how the workplace prepares for and organises student participation and also map out how students perceive their own engagement and preparedness for learning and participation in workplace activities.

Factor 2: Opportunities to learn in and through work and quality of supervision

The items map out perceptions of work experiences and how these relate to expected learning outcomes of a clinical rotation and also map out perceptions of metacognitive, social and emotional dimensions of learning and supervision.

Factor 3: Equal treatment

The items map out perceptions of fair and equal treatment between the students and the staff.

Factor 4: Workplace interaction patterns and student inclusion

The items map out perceptions of aspects of social participation and interaction with and among people in the workplace.

Second: Reliability Analysis

The Cronbach's alpha coefficients of the four factors of UCEEM questionnaire were shown in Table 2. They were ranged between 0.58 and 0.85. The overall Cronbach's alpha for the total UCEEM items was 0.92. This result indicates high internal consistency (reliability). Alpha levels did not increase if any items were deleted.

Third: Confirmatory Factor Analysis

First order confirmatory factor analysis

The four factors obtained by EFA form a reliable hypothetical model. In order to confirm whether it would be a good fit model, CFA using analysis of a moment structures (AMOS) was performed. A structural equation model was built with four factors and 25 items were generated from the EFA (Figure 2).

Figure 2 shows that the four constructs' regression weights ranging from 0.35 to 0.86. Modifications of the model were aided by the use of modification indices, guided by the fitness indices. The results indicated that the fitness indices improved after the cross-loading between constructs by applying "modification indices". Also, the goodness fit indices of the first order CFA of UCEEM questionnaire was measured and it revealed significant chi-square (CMIN = 609.92, df = 264, $p < 0.01$) which indicated a good fit; nevertheless, it is well-established that chi- square statistics are sensitive to sample size. Authors then investigated other fit indices, which suggested that the hypothesised model had an excellent fit with the sample data (CMIN/DF = 2.31, CFI = 0.93, TLI = 0.92, RMSEA = 0.049) (see Table 3).

Table 2: Reliability statistics of UCEEM questionnaire using Cronbach's Alpha

Factors	Items	Cronbach's alpha if item deleted	Cronbach's alpha per factor	Cronbach's alpha of all items
(F1) Preparedness for student entry and engagement	(12) The supervisors are well prepared for supervising.	0.81	0.84	0.92
	(15) I have sufficient access to supervision.	0.81		
	(16) I have a supervisor (field tutor) to whom I know I can turn.	0.81		
	(14) It is clear that my supervisors (field tutor) are familiar with the learning objectives.	0.82		
	(13) My supervisors were expecting me when I arrived.	0.83		
	(17) I received useful induction in this placement.	0.82		
	(19) As a student I am received in a positive way by the staff here.	0.82		
	(18) I feel included in the team of people who work here.	0.82		
(F2) Opportunities to learn in and through work and quality of supervision	(2) My work tasks are suitably challenging for my level of knowledge and skills.	0.84	0.86	
	(1) My problem solving skills are developing well in this placement.	0.84		
	(3) I get the opportunity to provide a rationale for my actions during supervision sessions.	0.84		
	(4) I have the opportunity to put my theoretical knowledge into practice in this placement.	0.84		
	(5) I am encouraged to participate actively in the work here.	0.84		
	(6) I am sufficiently occupied with meaningful work tasks.	0.85		
	(9) I feel I have influence over my learning in this placement.	0.85		
	(10) I feel able to ask my supervisors any question I wish.	0.85		
	(11) I have the opportunity to learn together with other medical students in this placement.	0.85		
	(8) I receive useful feedback from my supervisors.	0.85		
	(7) My work tasks are relevant to the learning objectives.	0.85		

(*continue on next page*)

Table 2: (*continued*)

Factors	Items	Cronbach's alpha if item deleted	Cronbach's alpha per factor	Cronbach's alpha of all items
(F3) Equal treatment	(24) Everyone is treated equally here regardless of cultural background.	0.71	0.78	0.92
	(25) Everyone is treated equally here regardless of gender.	0.76		
(F4) Workplace interaction patterns and student inclusion	(23) I have adequate access to computers in this placement.	0.66	0.68	
	(20) I feel welcome in the staff room/lunch room here.	0.58		
	(22) There is sufficient physical space for the number of medical students on placement here.	0.62		
	(21) Communication between those working here is good.	0.59		

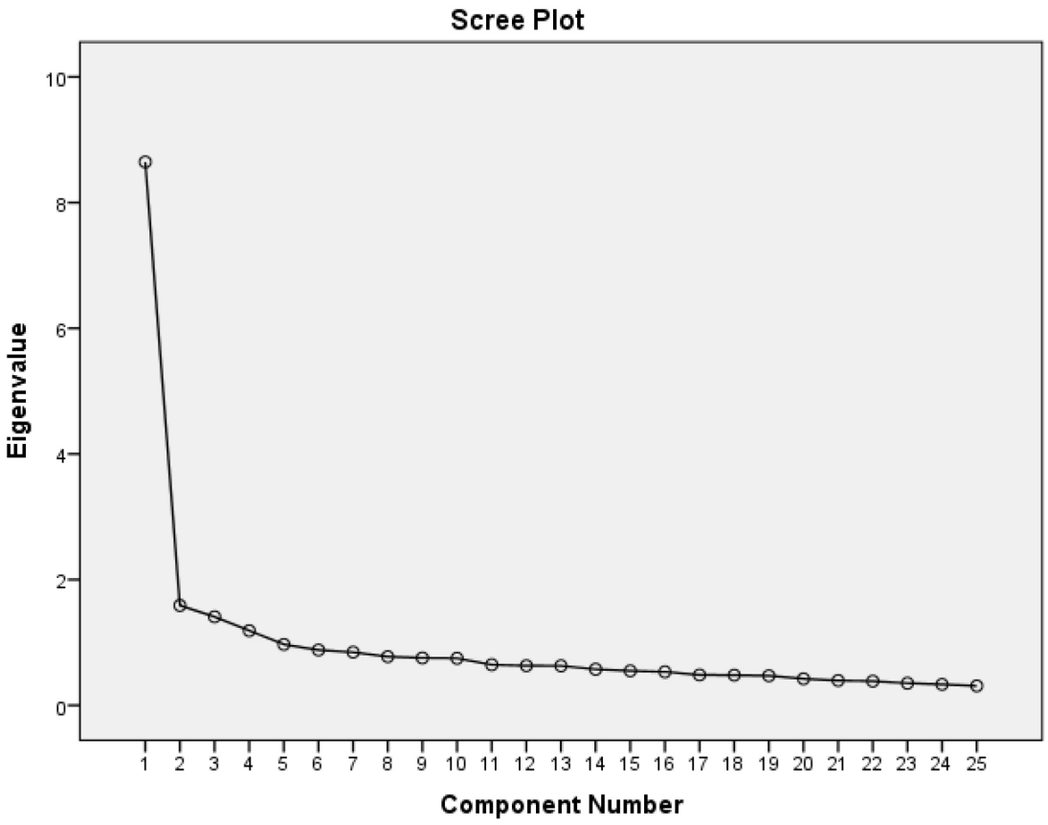

Figure 1: Scree plot of the eigenvalues of the factors of UCEEM questionnaire.

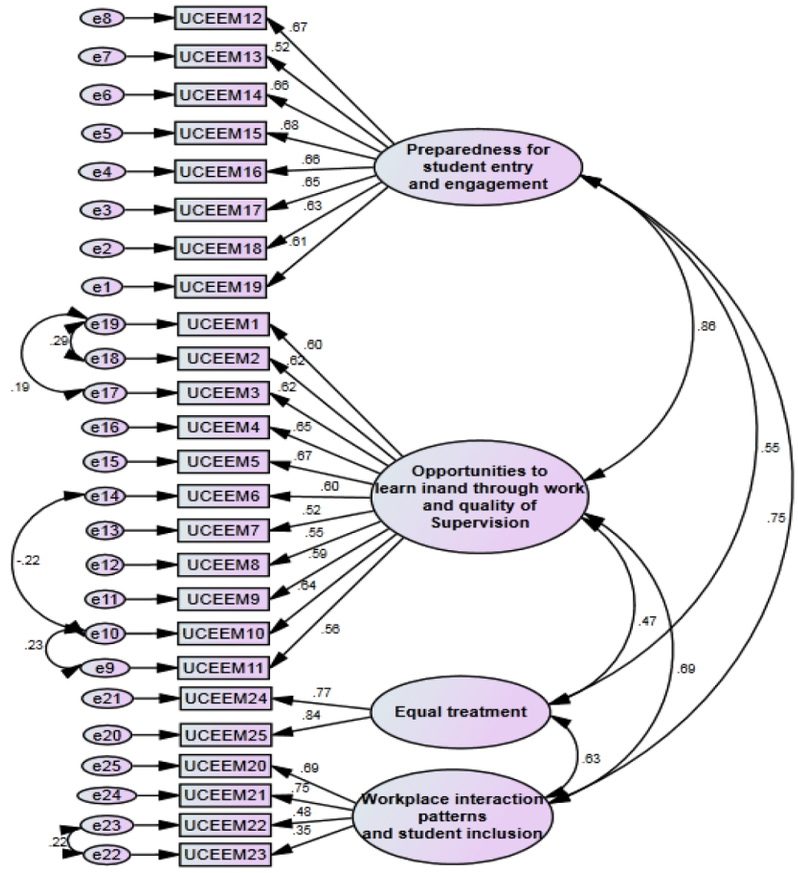

Figure 2: Linear model explaining the relationships among the different constructs belonging to the UCEEM questionnaire. Single headed arrows illustrate the standard regression coefficients between the construct and the corresponding items. Double headed arrows illustrate the covariance between different constructs.

Testing the validity evidence of the students' scores of the UCEEM using IRT

Table 4 illustrates the calibration of the 19-item scale with the reduced GRM (single slope for all items) resulted in a –2*Log Likelihood value of 25,571.27, whereas a fully specified GRM with unique slopes for all items yielded a value of 25,408.51. The difference in these two values (1,62.76) is distributed as chi-square with 19 degrees of freedom and is highly significant, indicating that the exclusion of the 19 unique item slope parameters in the more restricted GRM significantly detracts from the fit of the model. Consequently, the fully specified GRM was adopted as the more appropriate model for this item set.

The parameter estimates and their standard errors from the fully specified GRM calibration are listed in Table 5. The slope estimates ranged from 1.1 to 1.68, indicating minimal variation in item discrimination. The location parameters for the 19 items reflect a sizeable range for measuring the underlying construct (–2.74 to 3.13), but the majority of item response categories are only endorsed by respondents who have low and average levels of perception of learning experiences, implying that the item set as a whole is most useful in discriminating among students at the low and middle of the experiential learning continuum.

Figure 3 shows the test characteristic curve (TCC) that gives us information about the expected score in the questionnaire based on the level of agreement (ability = θ) (How will they agree to those items regarding their perception of learning experiences?), for example the expected score is about 42 if the level of agreement was 0 (middle of the continuum).

Table 3: Goodness fit indices of the first order CFA of UCEEM questionnaire

Model	CMIN	df	*p*	CMIN/DF	TLI	CFI	RMSEA
Default model	609.92	264	0.01	2.31	0.92	0.93	0.049

Note: Chi-square (CMIN), Degree of Freedom (df), Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA)

Table 4: The IRT model fit statistics

The model used	Reduced GRM	Fully specified GRM	The difference
2*Log Likelihood	25,571.27	25,408.51	1,62.76

Table 5: GRM item parameters estimates, standard error, and fit statistics for the 19 items of UCEEM scale (*n* = 550)

Items number	*a*	*b1*	*b2*	*b3*	*b4*	*S-X2*	*P*
UCEEM1	1.47	−2.07	−1.20	0.32	2.74	107.45	0.19
UCEEM2	1.44	−2.73	−1.10	0.34	2.70	135.61	0.00
UCEEM3	1.40	−2.63	−1.08	0.31	2.92	112.59	0.08
UCEEM4	1.47	−2.31	−0.98	0.28	2.45	152.09	0.00
UCEEM5	1.64	−2.00	−0.89	0.26	1.97	122.73	0.10
UCEEM6	1.30	−2.39	−0.90	0.69	3.08	121.67	0.10
UCEEM7	1.14	−2.97	−1.52	0.21	3.13	110.32	0.15
UCEEM8	1.29	−2.37	−0.96	0.41	2.12	155.13	0.00
UCEEM9	1.35	−2.45	−0.96	0.50	2.66	120.35	0.08
UCEEM10	1.65	−2.37	−1.55	-0.41	1.33	107.73	0.23
UCEEM11	1.38	−2.80	−1.70	-0.43	1.64	104.56	0.46
UCEEM12	1.59	−2.08	−1.05	0.22	1.80	127.48	0.04
UCEEM13	1.10	−2.74	−1.29	0.77	2.78	137.97	0.04
UCEEM14	1.64	−1.78	−0.88	0.46	2.11	144.92	0.00
UCEEM15	1.65	−2.50	−1.07	0.48	2.53	123.78	0.00
UCEEM16	1.56	−2.39	−1.05	0.45	2.32	130.34	0.01
UCEEM 17	1.60	−2.74	−1.18	0.47	2.31	110.75	0.10
UCEEM 18	1.54	−2.47	−1.19	0.24	2.03	125.77	0.04
UCEEM 19	1.43	−2.74	−1.30	0.41	2.43	158.87	0.00

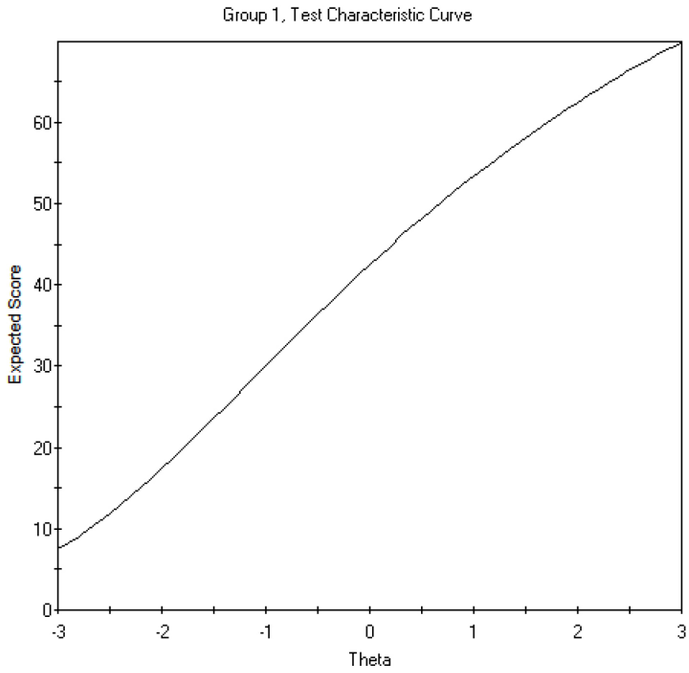

Figure 3: TCC showing the expected score in the questionnaire based on the level of agreement (ability = θ).

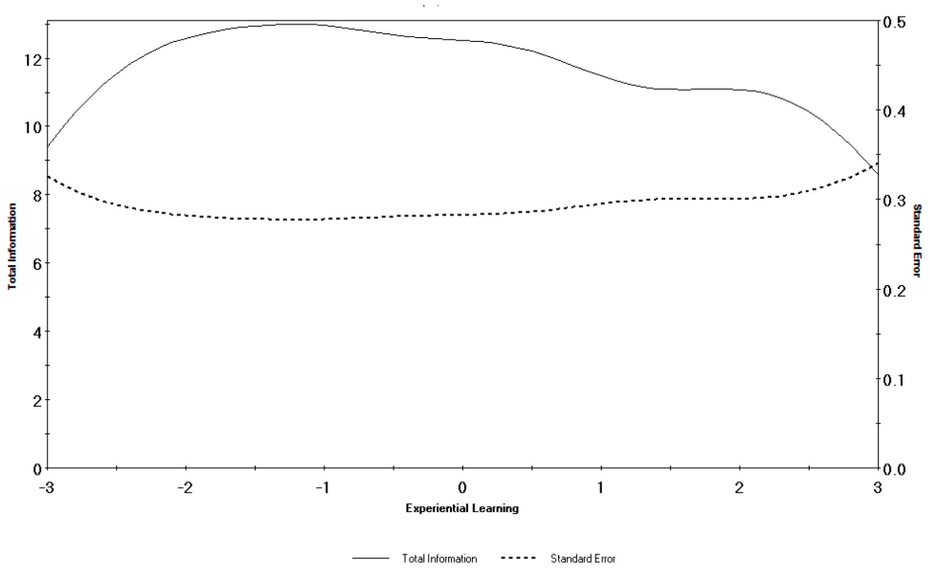

Figure 4: Total information curve (TIC).

Testing the reliability of students' scores of the UCEEM using IRT

The marginal reliability for response pattern scores is 0.92 indicating that this UCEEM questionnaire is a reliable tool for measuring experiential learning construct.

Figure 4 shows the total information curve (TIC) shows the all information provided by the 19 items where the items assess the lower and average levels of agreement more accurately than the upper level of agreement. The dotted line illustrates the standard error where it has smaller value in the lower and average levels of agreement that gives more information indicating more precise measurement at those levels and vice versa.

DISCUSSION

In this study, the psychometric properties of this instrument were tested for measuring the students' perception of different experiential learning activities in the two settings: the PHC centers affiliated to the FOM-SCU and the different clinical departments in Suez Canal University Hospital.

Thus, the collected and analysed data revealed that the UCEEM questionnaire can be categorised into four factors. These factors that reflected four dimensions of the clinical learning environment (CLE) as perceived by medical students were preparedness for student entry and engagement, opportunities to learn in and through work and quality of supervision, equal treatment, and workplace interaction patterns and student inclusion.

Obviously, the previous findings are comparable to the study of Strand and his collegues who originally developed the UCEEM questionnaire and before conducting the EFA, they emphasised that this questionnaire developed from five subscales which are preparedness for student entry, opportunities to learning in and through work, quality of supervision, workplace interaction patterns, and student inclusion learning climates in various contexts (3). All these factors were also emerged in our study, however two themes were integrated "Preparedness for student entry" and "quality of supervision"; in addition a new factor was added in the current study "organisation and supported participation" which is very important theme in measuring the quality of CLE.

The same previous authors used the UCEEM questionnaire at the University Hospitals of Lund and Malmo on Swedish medical students. They tested the questionnaire for construct validity using EFA. Their results showed that the questionnaire had four constructs which reflected four dimensions: opportunities to learn in and through work including quality of supervision, preparedness for student entry, workplace interaction patterns, and student inclusion and equal treatment (10). These findings are consistent with our study results although in factor two (opportunities to learn in and through work and quality of supervision) the number of items representing this factor in our study is 11 items while in their study was only nine items (3).

Moreover, the study of Strand and his collegues which was conducted on Swedish medical students at the University Hospitals of Lund and Malmo using the same instrument revealed that the first and second factors: preparedness for student entry and engagement, and opportunities to learn in and through work and quality of supervision cover experiential learning construct (3).

Therefore, the finding of our data demonstrated that the UCEEM questionnaire covers the most important dimensions of the learning environment which indicate an evidence for the internal structure validity of the UCEEM questionnaire. Additionally, a similar study was conducted in Islamic Azad University, Iran for testing the construct validity of the UCEEM questionnaire (Persian version)

using EFA. The EFA was close to the original five subscales of the UCEEM questionnaire (13).

A new addition to the existing literature was performed in the current study in the form of conducting item analysis using IRT of the UCEEM questionnaire using GRM. The item analysis of the data revealed that the hypothesised model had an acceptable fit with the proposed theory. It is worth-mentioning that up to our knowledge, this is the first study which examined the items for UCEEM questionnaire using one of the IRT models. Therefore, compared with previous studies, this study provided an additional source of evidence that support the construct validity of UCEEM questionnaire and fitting the measurement model with the theoretical model.

Furthermore, in our study the Cronbach's alpha coefficient value for the total scale was 0.92. This indicates high internal consistency (reliability) of UCEEM questionnaire. In addition, internal consistency reliability is, by itself, another evidence of construct validity of the questionnaire (12). This is congruent with the study of Strand and his collegues (3) who found that the Cronbach's alpha coefficients value was 0.93 and also with the study of Abbasi and her colleagues (13) who found that the Cronbach's alpha coefficients value was 0.93. Taken together, the findings in the current study indicate that the UCEEM questionnaire has a high reliability and acceptable evidence of construct validity.

CONCLUSION

The UCEEM questionnaire is a valid and reliable tool for measuring the experiential learning construct; this was evident by applying different psychometric tests to ensure its validity and reliability especially when used at the Faculty of Medicine, Suez Canal University.

ACKNOWLEDGEMENTS

The authors want to express their gratitude for all students at the Faculty of Medicine, Suez Canal University during the academic year (2015–2016) who participated in this study. We also wish to thank school administration who eases the work for this research and a gratefully thank members in medical education department for their continuous help.

REFERENCES

1. Yardley S, Teunissen PW, Dornan T. Experiential learning: AMEE Guide no. 63. Med Teach. 2012;34(2):e102–15. https://doi.org/10.3109/0142159X.2012.650741

2. Dornan T, Bundy C. What can experience add to early medical education? Consensus survey. BMJ. 2004;329(7470):834. https://doi.org/10.1136/bmj.329.7470.834

3. Strand P, Sjöborg K, Stalmeijer R, Wichmann-Hansen G, Jakobsson U, Edgren G. Development and psychometric evaluation of the undergraduate clinical education environment measure (UCEEM). Med Teach. 2013;35(12):1014–26. https://doi.org/10.3109/0142159X.2013.835389

4. Furr RM, Bacharach VR. Item response theory and Rasch models. Psychom An Introd. 2007;314–34.

5. Streiner DL. Measure for measure: new developments in measurement and item response theory. Can J Psychiatry. 2010;55(3):180–6. https://doi.org/10.1177/070674371005500310

6. Yang FM, Kao ST. Item response theory for measurement validity. Shanghai Arch Psychiatry. 2014;26(3):171–7.

7. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ. 2003;37(8):739–45. https://doi.org/10.1046/j.1365-2923.2003.01587.x

8. Baker FB. The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation; 2001, p. 25.

9. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44:109–17. https://doi.org/10.1111/j.1365-2923.2009.03425.x

10. Rasch G. Probabilistic models for some intelligence and achievement tests. Copenhagen Danish Inst Educ Res. 1960;4:382. https://doi.org/10.1016/S0019-9958(61)80061-2

11. Wright BD, Stone MH. Best test design: Rasch measurement. Chicago, IL: MESA PRESS; 1979. p. 1–17.

12. Ames AJ, Penfield RD. An NCME instructional module on item-fit statistics for item response theory models. Educational Measurement: Issues and Practice. 2015;34(3):39–48. https://doi.org/10.1111/emip.12067

13. Abbasi Zeinab, Ahmady S, Esmaeilpour S. Psychometric properties of undergraduate clinical education environment measure (UCEEM) in nursing and midwifery students in Iran. J Urmia Nurs Midwifery Fac. 2016;14(2):145–50.