Education in Medicine Journal ISSN 2180-1932

ORIGINAL ARTICLE

Volume 6 Issue 4 2014
DOI:10.5959/eimj.v6i4.271
www.eduimed.com

Construct Validity of Five Assessment Instruments at the End of Pre-Clerkship Phase in a PBL Curriculum.

Salah Eldin Kassab[1], Mariam Fida[2], Ahmed Al Ansari[3]

[1]Department of Medical Education, Faculty of Medicine, Suez Canal University, Ismailia, Egypt. [2]College of Medicine and Medical Sciences, Arabian Gulf University, Manama, Bahrain. [3]Department of General Surgery, Bahrain Defense Force Hospital, Bahrain.

ARTICLE INFO

Received	: 13/05/2014
Accepted	: 10/09/2014
Published	: 01/12/2014

KEYWORD

Medical education
Problem-based learning
Student assessment
Construct validity

ABSTRACT

Introduction: Developing a framework of educational outcomes which is aligned with student assessment helps in driving the student learning to achieve these outcomes. The validity of this framework is essential for ensuring fairness to learners and to society. **Objective**: The goal of this study is to investigate the validity of the underlying domains of the assessment framework emerging from the students' scores using five instruments used at the end of a pre-clerkship, problem-based medical curriculum. **Method**: Medical students (n=245, Year 4) were enrolled in the study during two consecutive academic years (2011 and 2012). We examined the construct validity of students' scores from the following assessment instruments: multiple choice questions, integrated short answer questions, objective structured clinical examinations, real patient clinical examinations and computer-based clinical simulations. An exploratory factor analysis was carried out followed by confirmatory analysis using structural equation modelling to evaluate the constructs emerging from the students' scores of the five instruments. **Result**: The analysis yielded four inter-related constructs, called medical knowledge, clinical skills, procedural skills, and reasoning skills. In addition, the four constructs loaded on a higher order construct (called competence) with high regression weights. Although the overall fitness of the separate assessment instruments was poor, the fitness indices of the three factor model after cross-loadings between variables of different constructs were improved [(χ^2 [94] = 149.2, p < 0.01 (Bollen-Stine p value = 0.104), Comparative Fit Index (CFI) = 0.96, Tucker Lewis Index (TLI) = 0.94 and Root Mean Square Error of Approximation (RMSEA) = 0.04 (90% CI = 0.03-0.07)]. **Conclusion**: The study indicates better validity evidence of the structure from the combined scores of the five instruments in explaining 'clinical competence" than the individual assessment instruments. The study also demonstrates that computer-based case simulations measure a separate construct which is different from measuring clinical skills in real or standardized patients.

CORRESPONDING AUTHOR: Salah Eldin Husseiny Kassab, Faculty of Medicine, Suez Canal University, Ring Road, Ismailia, PO Box 41111 Egypt. Email: kassabse@gmail.com

Introduction

The design of medical curricula has recently shifted from focussing on what teachers expect their students to learn in medical schools to what is expected from the physicians in clinical practice. Accordingly, many frameworks describing aspects of competence have been proposed based on the outcomes which the students are expected to acquire before graduation. With the advent of this outcome-based approach in education, frameworks for assessment of learners' competence have been developed to be congruent with this approach [1]. Consequently, the *n: n* approach, where one instrument can assess various competence domains and a competence domain is assessed using information from various instruments, has been proposed to replace the old trend of 1:1 (one instrument is testing one domain of competence) [2]. Therefore, determining the quality of assessment has been proposed to move from evaluation at the level of an individual assessment method to evaluation across methods [3].

Essential to constructing quality assessment instruments is gathering the required reliability and validity evidence to support the instrument's scores, purpose, use, and interpretation. A validation procedure needs to consist of a series of critical studies to determine whether the test actually assesses the construct it purports to measure [2]. To establish the construct validity evidence inference, the degree to which test scores indicate the amount of an unobservable trait the test purports to measure should be measured [4].

Two previous studies examined the construct validity of the combined scores from three assessment instruments used for clerkship students [5, 6]. Although these studies helped in contributing to the understanding of clinical competence, these three instruments did not capture many aspects of clinical competence. Taking into consideration the developmental nature of clinical competence, analysing the underlying structures of clinical competence in pre-clerkship students has been lacking. Therefore, this study provides an insight into understanding the validity of test score from five assessment instruments during the pre-clerkship phase of a PBL curriculum. The students' scores from five assessment instruments, namely MCQs, integrated SAQs, OSCEs, computer-based case simulations (CCS), and real patient-based clinical examination (PCE) were analyzed. In that context, the study was designed to answer the following research questions:

1. What are the underlying latent variables (constructs) which can be explored from the student assessment by the five instruments?
2. How much is the degree of relationship between these constructs and also between the primary constructs and possible emerging higher order constructs of competence?
3. How much is the degree of model fitness between the constructs emerging from the five assessment instruments and the observed structure of the measured variables (students' scores)?

Method

This study included a total of 245 medical students enrolled in a problem-based medical curriculum at the College of Medicine and Medical Sciences (CMMS), Arabian Gulf University (AGU) in the Kingdom of Bahrain, during the academic years of 2011-2012 and 2012-2013. The CMMS curriculum consists of a six-year program divided into three phases: phase 1 (Year 1), pre-clerkship phase (Years 2 to 4) and the clerkship phase (Years 5 & 6). The college adopts PBL as the main instructional method in the pre-clerkship phase. During this phase, students are exposed to nine different PBL units; three units are studied in each of the three years. Along with the PBL units, there is vertical representation of the professional clinical skills training and community-based activities. The main objectives of the professional clinical skills program are early clinical exposure with opportunities of practicing clinical skills in a safe environment using models and simulated patients.

The theme of the last unit of year 4 is "Multisystem Integration", in which students learn through prototype PBL cases which address multiple body systems. The main objectives of this unit are to emphasize vertical integration through the use of multi-system problems, and to prepare students for the clinical training during the clerkship phase. Students learn through multi-system paper-based cases in PBL tutorials. This runs in parallel to training using computer-based case simulations (CCS) and clinical skills training on real patients in Primary Health Care Centers. CCS has been recently introduced to teach clinical reasoning skills by selecting cases with similar pathologies to what is taught in PBL tutorials. The paper-based PBL case scenarios are designed with cues to help the students in generating "learning needs", through the integration of basic and clinical sciences, psychosocial and community aspects of the problem. On the other hand, computer-based cases are designed with the objective to help students in applying their acquired knowledge in clinical reasoning. At the end of the unit, students were assessed using summative examination using different instruments such as MCQs and SAQs, OSCE, CCS and PCE.

Assessment Instruments

Written assessment

Written assessment included a set of seventy five context-rich MCQs of the A-type (single best response), usually based on a clinical scenario. In addition, six integrated SAQs based on a clinical vignette with multiple stems were included. The questions included in each SAQ are generated from different disciplines, with the aim of testing the student's ability to integrate medical knowledge in different clinical and community contexts. An examination blueprint was constructed as a template for student assessment in this unit, which guided the selection of examination topics. Standard setting of the assessment scores was applied using modified Angoff's method for determining the borderline pass of students, through eight expert judges. The marking process for SAQs is structured so that a score is allocated for each question related to the scenario. Each SAQ is marked by the first author of the test item, reaching to a total of seven faculty raters.

Computer-based clinical simulations (CCS)

The CCS program is a web-based patient simulation package that trains students the clinical reasoning approach, using the hypothetico-deductive model (*DXR Development Group, Inc., Carbondale, IL*). The details of the different functional aspects of the software have been described previously [7]. Students access the assigned cases that mainly correlate with the cases discussed in PBL tutorials every week, via a local web server. Students were encouraged to complete each case independently, and feedback was given to the whole class and individually through emails sent by the program coordinator (M.F). The initial student encounter with the case begins with an online virtual patient presenting with a chief complaint. Students then progress through the case by collecting patient history, conducting virtual physical examination and ordering laboratory tests. While the students go through the case, they compile a list of working hypotheses and narrow down the differential diagnosis to a final diagnosis. Based on the selected diagnosis, they are required to design a patient management plan. Performance feedback is generated immediately after completing the case study.

At the end of the unit, 130 students from the academic year 2011-2012 were examined on three cases: 1) Rosetti - Chest pain # 4 (Unstable angina, atherosclerotic coronary vascular disease, hypertension and hyperlipidemia), 2) Bilroth - Syncope #1 (Obesity, hypertension, and hypercholesterolemia), and 3) Swenson - Back pain # 4 (Vertebral compression fracture secondary to osteoporosis). Students from the academic year 2012-2013 included 115 students and were examined on three cases: 1) Ray - Well-child visit # 2 (Membranous nephropathy and systemic lupus erythematosis), 2) Cohen - Cough # 1 (Pneumocystis carinii pneumonia, AIDS), and 3) Lancaster - Shortness of breath # 7 (Pulmonary embolism, deep vein thrombosis, and protein c deficiency).

The scoring system in the D_XR Clinician is calculated through the "Record Utility" software which tracks the students' encounter, and provides a separate score for each of the following three categories of student performance: clinical reasoning score, level of diagnostic performance and patient management. In the clinical reasoning score, students are assessed based on their ability to list their diagnostic hypotheses, arrive at the correct diagnosis, and select the investigations needed to justify the selected diagnosis by eliminating the rest. Evaluation of the level of diagnostic performance is a descriptive measure of what students include in their investigative inquiry by using one of ten descriptions. Each of these descriptions is assigned a value between zero and 100. Patient management is scored based on the four subcategories of Required, Recommended, Related H&P (History & Physical Examination), and Related Lab where each subcategory is assigned a numerical value based on the relative importance of each category [8]. The scores of the students in each of the three categories are analyzed and then combined to give the overall performance score. The relative weight of each category to the overall performance score can be adjusted by the examination coordinator, and then the program uniformly applies these parameters to calculate the scores for all the students. In this study, the overall CCS score per case was calculated as follows: 50% for clinical reasoning score, 40% for the level of diagnostic performance & 10% for patient management. The program gives the option for content knowledge assessment, but this option was not used in this study.

OSCE examination

OSCE examination was composed of ten stations in the year 2011-2012 group and 12 stations for the 2012-2013 group of students. This study included the 10 common stations in both cohorts of examinations. Students were allowed five minutes to finish each station, and they were divided into three groups.

In seven OSCE stations, standardized patients were used (table 1), and these stations included the following: 1) testing for vital signs (measuring pulse, blood pressure and reading thermometer), 2) history taking, 3) examination of the heart apex and demonstrating two maneuvers to elicit the apex beat if it is not palpable in supine position, 4) superficial and deep palpation of the abdomen, 5) musculoskeletal system, which included a scenario of a patient who cut his middle finger with a knife and cannot flex his finger and asking the student to evaluate the patient's problem and explain the steps to the examiner, 6) examination of fundus using an ophthalmoscope including labeling of the extra-ocular muscles on a diagram, and 7) students were provided with a written scenario of a patient with right intermittent facial pain and were asked to discuss trigeminal neuralgia and to name the cranial nerve affected, and to demonstrate the examination of the cranial nerve affected on a standardized patient. In the remaining three stations, models were used to test the following procedures: 1) palpation of breast quadrants and identifying a breast mass, 2) identifying an auroscope, and performing an examination on the model provided in addition to identifying different pathologies, 3) demonstrating how to collect a pap smear on a model.

Standardized patients were selected from the existing trained pool that is used in clinical teaching in the Professional Skills Program. Each station was scored by one faculty examiner, using a structured checklist and the cumulative scoring of the items was calculated out of 10. The content of the OSCE stations was reviewed by the concerned clinical experts and approved by the program examination board.

Real patient clinical examination (PCE)

In Primary Health Care (PHC) Centers, students were evaluated based on their competence in different aspects of clinical skills on real patients. There were two types of clinical assessment in this training period: continuous assessment, which is based on the weekly training in outpatient clinics of PHC, and represents longitudinal evaluation of the students' competence in undertaking clinical examination under supervision based on a structured checklist of clinical skills. At the end of the PBL unit, student's proficiency in clinical skills was evaluated using 5-point rating scale (excellent to poor), comprising of six skills components (vital signs, history taking, head and neck examination, chest examination, musculoskeletal system and neurological examination). Each student was evaluated by a single examiner in each of these competencies.

Statistical analysis

Data analysis in this study was conducted using Statistical Package for the Social Sciences (SPSS) software version 19 and SPSS AMOS version 20. Data are presented as mean ± SD of each variable. A *p*-value of <0.05 was considered to be statistically significant. Exploratory factor analysis (EFA), using principal axis factoring with promax rotation, was carried out to identify the different factors underlying the students' scores in the items related to the five assessment instruments. The number of factors that was extracted and used was based on Kaiser Rule (i.e., eigenvalues > 1.0), and on results from previous research.

On the basis of EFA, confirmatory factor analysis (CFA) using maximum likelihood estimation was carried out using the AMOS software for the different factor models, both the original instruments and the extracted factors. In addition, the structural model was added (structural equation modeling) to test the contribution of each construct to other higher order constructs. Missing data (4 samples, <2%) were treated by listwise deletion of missing variables. In case losses of 5% or less, removal of the overall missing data by listwise deletion is a defensible strategy for handling the incomplete data problem [9]. Multivariate normality was assessed using Amos by examining Mardia's normalized estimate of multivariate kurtosis [10]. We found statistically significant levels of multivariate skewness and kurtosis in the data suggestive of violating the multivariate normality assumption. To take account of this violation of assumptions, we conducted bootstrapping in AMOS using Bollen-Stine bootstrap approach to estimate the chi-square p value to be as an alternative p-value in consideration [11]. The number of bootstrap samples for this study was set at 250 samples as having bootstrap samples beyond this size does not give added quality in bootstrapped standard error estimates [12].

Different indices were used to evaluate the goodness-of-fitness of the different models compared with the data model: Comparative Fit Index (CFI) assesses overall improvement of a proposed model over an independence model where the observed variables are not correlated [13]. A good model fit is indicated by a CFI value of 0.90 or greater [14]. The *Chi-square* test indicates the amount of difference between expected and observed covariance matrices. The Root Mean Square Error of Approximation (RMSEA) is related to the residuals in the model, and a good model fit is typically indicated by RMSEA value of 0.06 or a value of 0.08 or less is often considered acceptable [15]. Tucker Lewis Index (TLI) or Non-normed Fit Index (NNFI) which is another indicator that is commonly used to measure model fitness [16].

Result

Descriptive statistics

Table 1 shows the descriptive statistics of the data generated from assessment of students' scores using the five instruments in the current study. The table illustrates the number of scoring items in each assessment component and the students' mean scores (SD) using different assessment instruments.

Table 1**:** Descriptive statistics of the data generated from evaluating the students (n=241) using the different assessment instruments used in the current study.

Assessment Instrument	**Type of assessment**	**# of scoring items**	**Students' scores**
			Mean (SD)
Written Assessment			
MCQs	Paper-based	75	65.28 (14.55)
Integrated SAQs	Paper-based cases	6	72.52 (14.44)
CCS	Computer-based cases		71.50 (15.19)
- Clinical Reasoning		3	76.50 (15.33)
- Diagnostic Performance			60.10 (17.77)
- Patient management			77.89 (12.46)
PCE	Real patient encounter		77.01 (10.90)
Continuous Assessment		6	79.21 (9.52)
End-Unit Assessment		1	74.80 (12.27)
OSCE			69.61 (11.38)
History taking	Observed (SP)	1	7.07 (1.21)
Vital signs	Observed (SP)	3	7.24 (1.69)
Chest examination	Observed (SP)	2	6.58 (1.24)
Abdominal examination	Observed (SP)	2	7.34 (1.42)
Musculoskeletal examination	Observed (SP)	5	6.58 (1.09)
Breast examination	Observed (Model)	1	6.89 (1.53)
Ear examination	Observed (Model)	3	6.09 (1.65)
CNS examination	Linked (SP + written case scenario)	3	7.11 (1.02)
Eye examination	Linked (SP + Diagram)	3	7.87 (1.30)
Gynecological examination	Observed (Model)	4	6.85 (1.60)

MCQs = multiple choice questions, SAQ = short answer questions, OSCE = objective structured clinical examination, CCS = computer-based case simulation, PCE = real patient-based clinical examination, and SP = standardized patient. Students' scores in different OSCE stations are marked out of 10, while the rest of scores are marked out of 100.

What are the underlying latent variables (constructs) which can be explored from the student assessment by the five instruments?

Exploratory factor analysis (EFA) yielded four extracted factors as shown in Table (2). Scores of PCE and some OSCE stations such as history taking, and examinations of chest, abdomen and musculoskeletal system loaded heavily on factor 1(clinical skills). The scores which loaded highly on factor 2 (procedural skills) included the stations of measuring vital signs, gynecological examination (collecting a Pap smear), ear examination (conducting hearing tests) and examination of the breast. Factor 3 (medical knowledge) had heavy loadings from the scores of written assessment (MCQ and SAQ), CNS examination and opthalmoscopic examination (assessment of extra-ocular muscles). Finally, factor 4 (reasoning skills) had heavy loadings mainly from the three CCS scores (i.e. clinical reasoning, diagnostic performance and patient management).

In the current study, the multivariate kurtosis was 57.83 and the CR was 17.00 indicating that the data violated the multivariate normality assumption. Therefore, the Bollen-Stine bootstrap approach was used to estimate the chi-square p value to be as an alternative for the normal-theory chi-square statistic.

Table 2**:** Exploratory factor analysis of the students' scores using five assessment instruments in pre-clerkship phase with four factor loadings extracted. Extraction is conducted using principal axis factoring with promax rotation. Only the pattern matrix is shown.

Assessment method	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
Written Assessment				
MCQs	-.097	-.061	**.789**	.263
Integrated SAQs	.148	-.107	**.758**	.179
CCS				
- Clinical Reasoning	.023	.027	.080	**.880**
- Diagnostic Performance	-.048	-.053	.076	**.946**
- Patient management	.370	.134	-.096	**.563**
PCE				
Continuous Assessment	**.571**	.254	.194	-.059
End-Unit Assessment	**.491**	.331	.069	-.116
OSCE				
- History taking	**.528**	.174	.116	-.038
- Vital Signs	-.172	**.891**	-.033	-.044
- Chest examination	**.736**	-.194	.085	.050
- Abdominal examination	**.875**	-.175	-.159	.003
- Musculoskeletal examination	**.701**	.034	-.042	.155
- Breast examination	.217	**.728**	-.026	-.014
- Hearing test	.010	**.587**	.384	-.162
- CNS examination	.205	-.217	**.741**	-.242
- Extraocular muscle examination	-.282	.299	**.680**	.065
- Collecting Pap smear	-.040	**.798**	-.244	.210

MCQs = multiple choice questions, SAQ = short answer questions, OSCE = objective structured clinical examination, CCS = computer-based case simulation, and PCE = real patient-based clinical examination. Factor 1 = clinical skills, Factor 2 = Procedural skills, Factor 3 = Medical knowledge & Factor 4 = reasoning skills.

How much is the relationship between these constructs and also between the primary constructs and possible emerging higher order constructs?

Subjecting the four extracted factors with the related variables to CFA using structural equation modeling (SEM) indicated high regression coefficients between each latent variable (construct) and the underlying variables, as shown in Figure 1. The four constructs (after cross-loadings) moderately or highly correlated with each other with coefficients ranging from 0.32 to 0.66. In addition, second order CFA indicated that the four constructs tapped on a common construct (called competence) with regression weights ranging from 0.49 to 0.87 (Figure 2).

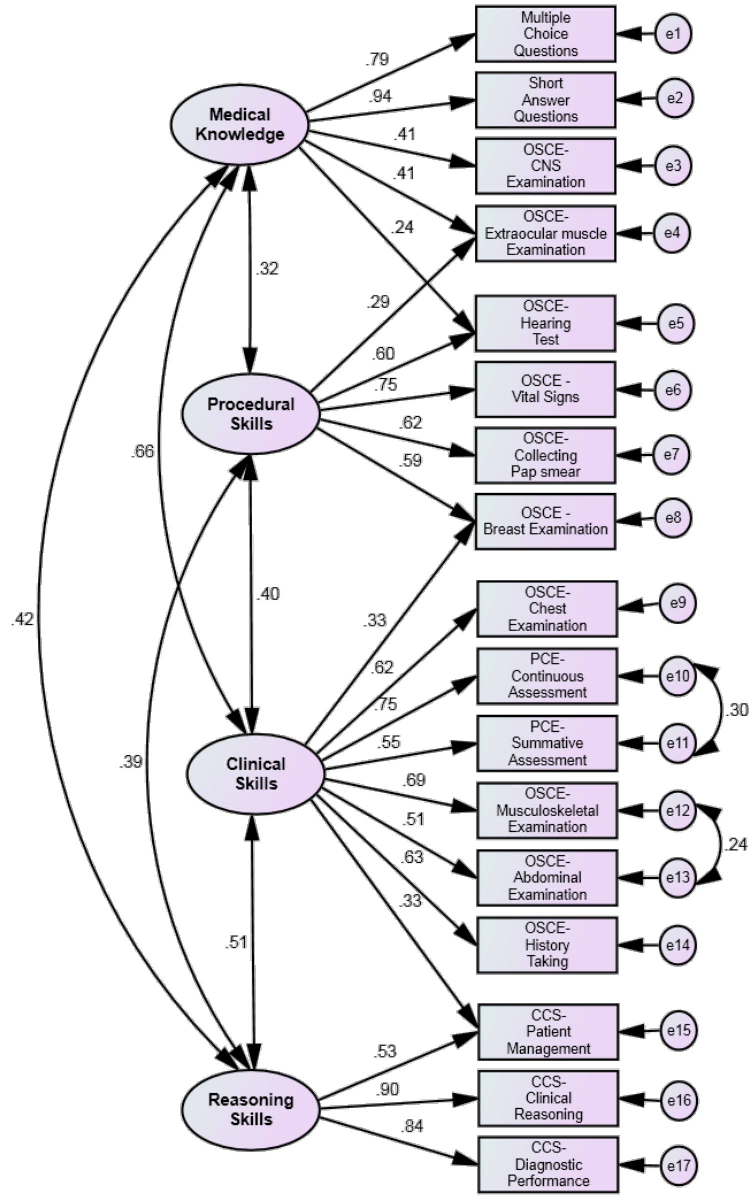

Figure 1: Linear model explaining the relationships among the different constructs belonging to the five assessment instruments. Standard regression coefficients show that the students' scores from the five assessment instruments tap on four latent constructs (medical knowledge, procedural skills, clinical skills and reasoning skills). Double headed arrows illustrate the correlation coefficients between different constructs.

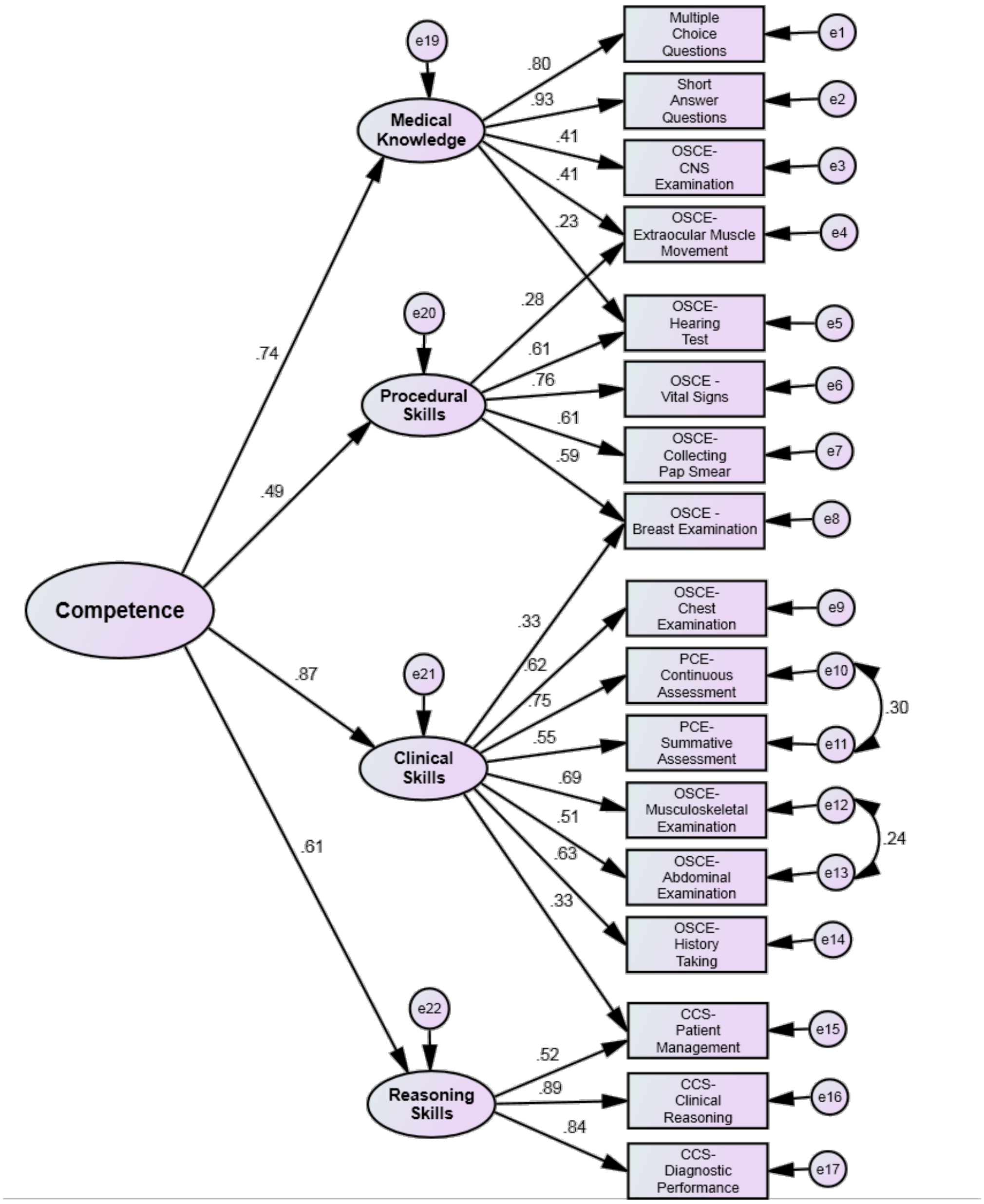

Figure 2: Higher order confirmatory factor analysis using structural equation modeling in the students' scores from the five assessment instruments. Linear model explaining the relationships among the different constructs belonging to the five assessment instruments. Standard regression coefficients show that knowledge, psychomotor skills and reasoning skills significantly affect the overall assessment of student competence.

How much is the degree of model fitness between the constructs emerging from the five assessment instruments and the observed structure of the measured variables (students' scores)?

Table 3 show the fitness indices in four different models: five original instruments, four factor model extracted from EFA, four extracted factors model with cross loadings using CFA, and second-order CFA with three extracted factors model with cross-loading. *Post-hoc* modifications of the model were aided by the use of modification indices, guided by the fitness indices and restricted to theoretically justifiable parameters. The results indicated that the fitness indices progressively improved as we shifted from the first model to the fourth model. Fitness indices improved after the cross-loading between constructs by applying “modification indices”. The cross-loadings were between the following: a) “medical knowledge” construct and ear examination in OSCE which is clustered with “procedural skills” construct; b) “procedural skills” construct and another variable, from “medical knowledge” construct (testing extra-ocular muscles- OSCE); and c) “clinical skills” construct and two other variables, one from “procedural skills” construct (breast examination- OSCE) and the other from “reasoning skills” construct (patient management). The overall fit of the three factor model after conducting the second order CFA was acceptable, resulting in χ^2 [94] = 149.2, p = 0.004 (Bollen-Stine p value = 0.104 which is not statistically significant). Other fitness indices include a comparative fit index (CFI) of 0.96, TLI of 0.94 and root mean square error of approximation (RMSEA) of 0.04 (0.03-0.07).

Table 3: Results for the confirmatory factor analysis (CFA) of the five instruments showing the fitness indices with four different models: five original instruments, four extracted factors, four extracted factors with cross loadings, and three extracted factors with second order CFA.

Model & no. of factors	CMIN	DF	P-value	CFI	TLI	RMSEA (90% CI)	Bollen-Stine bootstrap
First order CFA							
5 original instruments	303.9	115	0.000	0,81	0.77	0.11 (0.10 – 0.13)	0.004
4 (Extracted factors)	238.6	113	0.000	0.87	0.84	0.09 (0.08 – 0.11)	0.016
4 (Extracted factors with cross loadings)	160.4	108	0.001	0.91	0.93	0.06 (0.04 – 0.08)	0.072
Second order CFA							
3 (Extracted factors with cross loadings)	149.2	94	0.004	0.96	0.94	0.04 (0.03 – 0.07)	0.104

DF = degrees of freedom; CMIN = Chi Square; CFI = comparative fit index; TLI = Tucker Lewis Index or Non-normed Fit Index (NNFI); RMSEA = root mean square error of approximation; 90% CI = 90% confidence interval.

Discussion

This study examined the underlying latent structure among five assessment instruments in order to understand how they individually and jointly contribute to measuring medical competence of medical students in a pre-clerkship, PBL curriculum. The study provided evidence that the scores of students (measurement model) fits better with the structural model of the latent constructs (competence domains) emerging from five instruments than the structure of individual instruments. The results indicated that the students’ scores from the five instruments yielded four correlated latent constructs: called medical knowledge, clinical skills, procedural skills and reasoning skills. Furthermore, the four separate constructs tapped into a common higher order construct of “competence”. Results of the CFA indicated that the fitness indices progressively improved and reached a level of acceptable fitness in the third model of tapping the four constructs into a common domain “competence”. The improved fitness of this structural model with the measurement model of the students’ scores indicates that this model better explains the understanding of the complex domain of competence in undergraduate students in a PBL setting.

An interesting finding in the current study is that the CCS scores represented a separate construct with high regression coefficients with the three variables: clinical reasoning, diagnostic performance and patient management. We have also reported that the "reasoning skills" construct correlates moderately with "clinical skills", which represents the scores from examining students on real patients and from OSCE stations using SPs. These findings corroborate a recent study which reported a significant correlation between the students' scores in OSCE stations using computer-based virtual patients, and stations using standardized patients [17]. Another study of residents in a tertiary care setting found a moderate positive correlation between scores of using CCS and standardized patients [18]. On the other hand, a study on third year students on primary care clerkship using the same computer software found no significant correlations between the students' scores in CCS (clinical reasoning and diagnostic performance) and their scores on any of the Diagnostic Thinking Inventory (DTI), and they concluded that CCS using this software lacks criterion validity [7]. In the current study, the regression weight of 0.61between "reasoning skills" and "competence" indicates the significant impact of this construct to the common domain of "competence" and could justify the use of the CCS as a separate indicator of competence in the assessment profile of medical students. Although the current findings indicated that CCS represented a separate construct, the method effect may also contribute to the finding [19]. Further studies, using larger sample of students in different contexts, will be required to test the validity of CCS scores in relation to other validated tools of diagnostic reasoning, such as script concordance test.

Although the written examination scores clustered on one separate construct "medical knowledge", the OSCE scores were distributed among the three constructs, namely clinical skills, procedural skills and medical knowledge. A previous explanation was put forward to justify the low validity of OSCE scores is that OSCEs measure multiple constructs of knowledge and skills, and therefore, are not expected to correlate well with standard testing formats [20]. The finding in the current study that OSCE scores in stations testing nervous system and extra-ocular muscles tapped on the "medical knowledge" could be explained by the fact that students' performance on the two OSCE stations relied more on their integrative knowledge of linking basic science concepts to clinical conditions. On the other hand, the high loading of the OSCE stations using standardized patients, along with scores of real patient clinical examination (PCE), into the "clinical skills" construct indicate that these two instruments are measuring similar aspects of student performance.

Validity is currently seen as building a train of arguments of how best observations of behavior can be translated into scores and how these can be used at the end to make inferences about the construct of interest [3]. In the current study, although the standard psychometric methods (CFA and SEM) were used for validity inferences of multiple assessment instruments in the PBL program, there are other relevant validity arguments which should be considered. These arguments include examination blueprinting, criterion-based assessment, test item review, examiner training, scoring rules, and judgment processes [21]. Future studies should address a more comprehensive view on the role of these arguments in the construct validity of the student assessment in integrated PBL programs.

Although we believe that the methodology used in the current study did not lack rigor, there are some limitations. First, we have defined competence in this study based on the "analytical framework" of student assessment, which is characteristic for most outcome-based curricula [1]. The analytical assessment framework in the current study included only the outcomes of medical knowledge and skills (clinical, procedural and reasoning). However, there are other important components of competence (e.g. professional values, interpersonal skills, communication skills, etc.) which were assessed in this PBL unit, but were not included in the data analysis of this study. Another limitation is that students were evaluated on relatively small number of cases used for assessment in OSCE, CCS and PCE. Taking into consideration the impact of case specificity on the different domains measured in this study, and that the study is conducted in one PBL medical school, the generalizability of the study findings to other settings is limited. Finally, this study focused on examining construct validity based on CFA alone and no other auxiliary measures were used to confirm the findings.

Conclusion

We conclude that the combined scores of the five assessment instruments used in pre-clerkship phase tapped into four latent constructs (medical knowledge, clinical skills, procedural skills and reasoning skills). This model provided better validity evidence of its internal structure compared with the model using individual scores of each assessment instrument. Furthermore, computer-based case simulations using software for assessment of clinical reasoning are measuring a unique construct, which is not measured by other assessment instruments.

Reference

1. Pangaro L, Cate OT. Frameworks for learner assessment in medicine: AMEE Guide No. Medical Teacher 2013; 35: e1197–e1210.
2. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Medical Teacher 2011; 33: 478–85.
3. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical Education 2005; 39: 309–317.
4. Anastasi A, Urbina S. Psychological testing (7th Ed.). Upper Saddle River, NJ: Prentice-Hall. 1997.
5. Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, Kleinhenz ME. Validity of three clinical performance assessments of internal medicine clerks. Academic Medicine 1995; 70:517–22.
6. Lee M, Wimmers PF. Clinical competence understood through the construct validity of three clerkship assessments. Medical Education 2011; 45: 849–857.
7. Jerant AF, Azari R. Validity of scores generated by a web-based multimedia simulated patient case software: a pilot study. Academic Medicine 2004; 79(8): 805-11.
8. DxR Clinician Instructor Manual, D$_X$R Development Group, Inc., Carbondale, IL, (2011). [Retrieved 25 Dec 2011]. Available from: (http://www.dxrgroup.com/dxronline/downloads/v3/DxRC_InstrManv3_2011.pdf).
9. Roth P. Missing data: A conceptual review for applied psychologists. Personnel Psychology 1994; 47, 537-560.
10. Arifin WN, Yusoff MSB, Naing NN. Confirmatory factor analysis (CFA) of USM Emotional Quotient Inventory (USMEQ-i) among medical degree program applicants in Universiti Sains Malaysia (USM). Education in Medicine Journal, 2012; 4(2).
11. Bollen KA, Stine RA. Bootstrapping goodness-of-fit measures in structural equation models. Sociological Methods and Research. 1992; 21:205–229.
12. Nevitt J, Hancock GR. Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. Structural Equation Modeling 2001; 8, 353-377.
13. Byrne BM. Structural equation modeling with AMOS: Basic concepts, applications, and programming. 2nd ed. New York: Taylor & Francis Group; 2010.
14. Violato C, Hecker K. How to use structural equation modeling in medical education research: A brief guide. Teaching and Learning in Medicine 2007; 4: 362-71.
15. Browne MW, Cudeck R. Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), testing structural equation models. Newbury Park, CA: Sage Publications. 1993; 136–62.
16. Bentler PM, Bonett DG. Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin 1980; 88: 588-600.
17. Oliven A, Nave R, Gilad D, Barch A. Implementation of a web-based interactive virtual patient case simulation as a training and assessment tool for medical students. Studies in Health Technology and Informatics 2011; 169: 233-37.
18. Hawkins R, MacKrell Gaglione M, LaDuca T, Leung C, Sample L, Gliva-McConvey G, Liston W, De Champlain A, Ciccone A. Assessment of patient management skills

and clinical skills of practising doctors using computer-based case simulations and standardised patients. Medical Education 2004; 38(9): 958-68.

19. Brown TA. Confirmatory Factor Analysis for Applied Research. New York, NY: Guilford 2006; 2.
20. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Family medicine 2008; 40(8):574-8.
21. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity. Medical Education 2012; 46: 38–48.