ORIGINAL RESEARCH

Do medical students' scores using different assessment instruments predict their scores in clinical reasoning using a computer-based simulation?

Mariam Fida[1]
Salah Eldin Kassab[2]

[1]Department of Molecular Medicine, College of Medicine and Medical Sciences, Arabian Gulf University, Manama, Bahrain; [2]Department of Medical Education, Faculty of Medicine, Suez Canal University, Ismailia, Egypt

Purpose: The development of clinical problem-solving skills evolves over time and requires structured training and background knowledge. Computer-based case simulations (CCS) have been used for teaching and assessment of clinical reasoning skills. However, previous studies examining the psychometric properties of CCS as an assessment tool have been controversial. Furthermore, studies reporting the integration of CCS into problem-based medical curricula have been limited.

Methods: This study examined the psychometric properties of using CCS software (DxR Clinician) for assessment of medical students (n=130) studying in a problem-based, integrated multisystem module (Unit IX) during the academic year 2011–2012. Internal consistency reliability of CCS scores was calculated using Cronbach's alpha statistics. The relationships between students' scores in CCS components (clinical reasoning, diagnostic performance, and patient management) and their scores in other examination tools at the end of the unit including multiple-choice questions, short-answer questions, objective structured clinical examination (OSCE), and real patient encounters were analyzed using stepwise hierarchical linear regression.

Results: Internal consistency reliability of CCS scores was high (α=0.862). Inter-item correlations between students' scores in different CCS components and their scores in CCS and other test items were statistically significant. Regression analysis indicated that OSCE scores predicted 32.7% and 35.1% of the variance in clinical reasoning and patient management scores, respectively ($P<0.01$). Multiple-choice question scores, however, predicted only 15.4% of the variance in diagnostic performance scores ($P<0.01$), while students' scores in real patient encounters did not predict any of the CCS scores.

Conclusion: Students' scores in OSCE are the most important predictors of their scores in clinical reasoning and patient management using CCS. However, real patient encounter assessment does not appear to test a construct similar to what is tested in CCS.

Keywords: medical education, computer-based simulations, virtual patients, student assessment, PBL, Bahrain

Introduction

The goal of assessment in medical education remains the development of reliable instruments of student performance which, as well as having predictive value for subsequent clinical competence, also have a formative educational role.[1] Physicians arrive at proper diagnosis of patients' conditions through a process of problem solving. This process requires the physician to listen to the patients' complaints, ask relevant questions, conduct proper physical examinations, and order appropriate investigations. To develop

Correspondence: Salah Eldin Kassab
Department of Medical Education, Faculty of Medicine, Suez Canal University, Ring Road, P.O. Box 41111, Ismailia, Egypt
Email kassabse@gmail.com

expertise in clinical diagnosis, physicians require possessing clinical reasoning and physical examination skills, in addition to having acquired domain knowledge. The processes underlying the cognitive behavior of medical professionals when handling patient problems and how they develop clinical reasoning expertise has not been well understood. The process of clinical reasoning includes the use of two different approaches: 1) a slow hypothetico-deductive approach relying mainly on biomedical knowledge and mostly used by novices; and 2) a fast pattern recognition approach through retrieval of illness scripts and mostly used by experts.[2] However, recent studies demonstrated that the clinical reasoning process includes the use of both approaches in tandem.[3] Aside from the physician's level of expertise, other variables could also affect the reasoning process including the case specificity and the context of the patient's case.[4]

Clinical reasoning is a process that evolves over time and requires integrated basic and clinical knowledge and repeated practice or training. Therefore, developing instructional methods which help medical students to become efficient in solving patients' problems is a central issue in medical education. Early clinical exposure can help students to develop scripts very early and to incorporate biomedical and clinical knowledge that they would acquire subsequently within their scripts, if appropriate care is taken about integration of this knowledge.[5]

Virtual patients or computer-based case simulations (CCS) have been defined as a "specific type of computer program that simulates real-life clinical scenarios; learners emulate the roles of health care providers to obtain a history, conduct a physical exam, and make diagnostic and therapeutic decisions".[6] Learners ask the computer program questions related to the clinical presentation of the virtual patients and collect information about medical history and physical examination and laboratory investigations. By practicing on this program, learners use a hypothetico-deductive approach to clinical reasoning until they arrive at a final diagnosis and a treatment plan. Although CCS have been advocated as useful tools for teaching clinical skills and evaluating clinical competence of students, the validity and reliability of assessment scores using this method are still unclear.

A previous study using web-based virtual patients (DxR Clinician) demonstrated no correlation between students' scores in clinical reasoning and their scores in a Diagnostic Thinking Inventory, a tool used for diagnostic reasoning.[7] Oliven et al found a significant correlation between the students' scores in OSCE stations using computer-based virtual patients and stations using trained actors (standardized

patients).[8] Previous studies examined the relationship between scores in CCS and written assessment instruments. In the United States Medical Licensing Examination (USMLE) step 3 examinations, the CCS correlated moderately with the multiple-choice questions (MCQs) component of the examination.[9] In addition, a correlation between CCS and the National Board of Medical Examiners (NBME) examination MCQ scores in a pediatrics clerkship was found to be low.[10] Therefore, it remains unclear whether assessment using CCS brings added value to the traditional written examinations. This study provides an additional insight into the predictive value of the CCS scores in a multisystem integrated problem-based module. The relationship between students' scores in the CCS and their scores in other assessment tools such as real patient-based clinical encounter (PCE), MCQs, integrated short-answer questions (SAQs), and objective structured clinical examination (OSCE) are explored.

Methods

The medical curriculum at the College of Medicine and Medical Sciences (CMMS), Arabian Gulf University (AGU), Manama, Bahrain consists of a 6-year program divided into three phases: phase 1 (year 1), pre-clerkship phase (years 2 to 4), and the clerkship phase (years 5 and 6). The college adopts problem-based learning (PBL) as the main instructional method in the pre-clerkship phase. During this phase, students are exposed to nine different PBL units; three units in each of the three years. In PBL tutorials, students learn to integrate different aspects of knowledge related to each problem, including basic medical sciences, clinical sciences, and community health. Along with the PBL units, there is vertical representation of the professional clinical skills training and community health activities.

Study context

This study was conducted during the academic year 2011–2012. The study sample included all medical students in year 4 (n=130) studying the last unit (Unit IX: Multi-system Integration) of the pre-clerkship phase of the medical program. The duration of the unit spans over 6 weeks and students study one main PBL case per week in addition to applying the knowledge acquired from each case to another set of mini-problems. The main goals of this unit are: 1) to emphasize vertical integration through the use of multisystem problems; and 2) to prepare students for the clinical training during the clerkship phase. Students learn through multisystem paper-based cases in PBL tutorials. This runs parallel with

the clinical skills training on real patients in Primary Health Care (PHC) centers. In addition, CCS has been recently introduced for enhancing clinical reasoning skills through cases selected with similar pathologies to what they studied in PBL tutorials. The paper-based PBL case scenarios are designed with cues to help the students in generating "learning needs" covering integrated basic sciences, clinical sciences, and psychosocial and community aspects related to the problem. On the other hand, computer-based cases are designed with the objective to help students in applying their knowledge in clinical reasoning.

Assessment instruments

Written assessment included a set of 75 context-rich MCQs of the A-type (single best response), usually based on a clinical scenario, and six integrated SAQs based on a clinical context with a number of questions linked to the scenario. The questions included in each SAQ were also related to different disciplines and aimed mainly to test the student's ability to integrate medical knowledge in relation to 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 written assessment was applied using modified Angoff's method[11] for determining the borderline pass of students, moderated by eight expert judges.

CCS

The DxR Clinician Program is a web-based patient-simulation package that trains students on clinical reasoning using the hypothetico-deductive approach (DxR Development Group, Inc., Carbondale, IL, USA). Students accessed the required cases via a local webserver and cases were provided to the students every week in parallel with the problems discussed in PBL tutorials. Students were encouraged to complete each case independently and feedback was given to the whole class, and individually through emails, by the program coordinator (MF). The students' encounters with the case started with an online patient presenting with a chief complaint and, as students progressed through the case, they collected patient history, conducted virtual physical examinations, and ordered investigations and tests. While the students went through the case, they compiled a list of working hypotheses and narrowed their list of hypotheses into a final diagnosis. Based on their diagnosis, they developed a patient management plan. Students received performance feedback immediately on completing the task.

Clinical reasoning scores were calculated based on the student's ability to list the diagnostic hypotheses related to the case, to arrive at the correct diagnosis, and to select the investigative items from the patient information needed to justify the correct diagnosis, and therefore rule out competing hypotheses short-listed earlier. As for the level of diagnostic performance, a descriptive measure of what students include in their investigative inquiry classified a student's performance by one of ten descriptions. Each of these descriptions was assigned a value between 0 and 100. Management, on the other hand, was scored based on the four subcategories of Required, Recommended, Related History and Physical Examination, and Related Lab, where each subcategory is assigned a numerical value based on the relative importance of each category.

The scoring system in the DxR Clinician is calculated through the "Record Utility" software, which tracks the students' interactions and provides a separate score for three categories of student performance, namely clinical reasoning, diagnostic performance, and patient management. The scores of the students in these three categories are weighted 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 calculates the student scores.

OSCE

OSCE was comprised of ten stations, 5 minutes each, and the students were divided into three groups. In seven OSCE stations, standardized patients were used, and these stations included the following required skills: vital signs, history taking, examination of the heart, superficial and deep palpation of the abdomen, musculoskeletal system examination, examination of fundus using an ophthalmoscope, and testing for the fifth and seventh cranial nerves. In the remaining three stations, models were used to test the following procedures: palpation of a breast mass, using an auroscope, and demonstrating how to collect a pap smear. Each station was scored by one faculty examiner, using a structured checklist. The content of the OSCE stations was reviewed by the clinical experts concerned and approved by the program examination board.

Real patient encounter examination

In the PHCC, students were assessed in terms of 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 PHCC, and end-unit direct observation of patient encounter. Continuous assessment represents longitudinal evaluation of the students' competence in undertaking clinical examination under supervision, based on a structured checklist of skills. At the end of the unit, a student's proficiency in clinical skills was evaluated using a 5-point rating scale (excellent to poor) addressing seven skills: vital signs, history taking, head and neck examination, chest examination, musculoskeletal system examination, and neurological examination. Each student was evaluated by a single examiner in each of these competencies. Clinical faculty involved in the students' training in PHCC attended an orientation session conducted by the program coordinator about teaching and clinical assessment of students in PHCC.

Statistical analysis

The Statistical Package for the Social Sciences (SPSS) software version 19 was used for data entry and analysis in this study. Data are presented as mean ± standard deviation of each variable. Descriptive statistics using means, frequencies, and percentages were tabulated. Internal consistency reliability of the examination scores was calculated using Cronbach's alpha statistics. The relationships between students' scores in the CCS as the dependent variables and the students' scores in other test item formats (such as MCQs, SAQs, OSCE, and PCE), as independent variables, were analyzed using hierarchical stepwise linear regression. A P-value of <0.05 was considered to be statistically significant.

Results
Internal consistency reliability of the examination scores

Internal consistency reliability of the examination scores in different test item formats (CCS, SAQs, OSCE, and the PCE) of the unit summative examination showed good to acceptable levels. The reliability coefficients from highest

to lowest were CCS (α=0.862), SAQs (α=0.817), OSCE (α=0.767), and PCE (α=0.644).

Inter-item correlations between different examination scores

Correlation analysis was conducted to identify the relationships between the students' performance in the CCS and its three components (diagnostic performance, clinical reasoning, and patient management) with their performance in the other test item formats: MCQs, SAQs, OSCE, and the PCE. As shown in Table 1, examination scores in different test item formats were positively and significantly ($P\leq0.01$) correlated with each other and with the total DxR score. The highest degrees of correlations were between the different components of the CCS examination scores.

Predictors of CCS scores

The stepwise hierarchical multiple regression model shown in Table 2 indicates that SAQ scores predicted 16% of the variance in the total DxR score (adjusted R^2=0.158, β=0.398, $P<0.01$) and increased by 4% when OSCE scores were introduced. However, OSCE scores predicted approximately 33% of the variance in clinical reasoning scores (adjusted R^2=0.327, β=0.572, $P<0.01$) and increased by 2.2% when the SAQ scores were included. In addition, OSCE scores predicted 35% (adjusted R^2=0.351, β=0.592, $P<0.01$) of the variance in clinical management scores in the CCS and increased by 2% when PCE was introduced in the model ($P<0.05$). On the other hand, MCQ scores predicted approximately 15% of the diagnostic performance scores (adjusted R^2=0.154, β=0.392, $P<0.01$), while the OSCE scores predicted only 3.1% ($P<0.05$). However, PCE scores did not significantly predict any of the CCS scores.

Discussion

This study examined the psychometric properties of CCS assessment. The validity-related evidence included testing the

Table 1 Pearson's correlations between the students' scores in the three components of the CCS (diagnostic performance, clinical reasoning, and patient management) and their scores using other assessment tools (MCQs, SAQs, OSCE, and PCE)

Variables	Diagnostic performance	Clinical reasoning	Management	MCQs	SAQs	OSCE	PCE
Total CCS score	0.98**	0.93**	0.57**	0.39**	0.40**	0.40**	0.23**
Diagnostic performance score		0.86**	0.48**	0.39**	0.38**	0.36**	0.20*
Clinical reasoning score			0.69**	0.42**	0.46**	0.57**	0.39**
Management score				0.27**	0.38**	0.59**	0.53**

Notes: *$P<0.05$; **$P<0.01$.
Abbreviations: CCS, computer-based case simulations; MCQs, multiple-choice questions; SAQs, short-answer questions; OSCE, objective structured clinical examination; PCE, real patient-based clinical encounter.

Table 2 Stepwise hierarchical linear regression analysis of the students' scores in CCS components (diagnostic performance, clinical reasoning, patient management scores) as dependent variables and their scores in other assessment tools (MCQs, SAQs, OSCE, and PCE) as predictors of CCS scores

Dependent variables	Predictors	R^2	f^2	ΔR^2	b	Standard error	β	P-value
Total CCS score	Model 1							
	SAQs	0.158	0.05		0.740	0.151	0.398	0.000
	Model 2							
	SAQs	0.158			0.465	0.184	0.250	0.013
	OSCE	0.198		0.040	0.687	0.274	0.248	0.013
Diagnostic performance score	Model 1							
	MCQs	0.154	0.04		0.711	0.147	0.392	0.000
	Model 2							
	MCQs	0.154			0.511	0.171	0.282	0.003
	OSCE	0.185		0.031	0.566	0.256	0.208	0.029
Clinical reasoning score	Model 1							
	OSCE	0.327	0.03		0.994	0.126	0.572	0.000
	Model 2							
	OSCE	0.327			0.805	0.155	0.463	0.000
	SAQs	0.349		0.022	0.213	0.104	0.182	0.043
Patient management score	Model 1							
	OSCE	0.351	0.03		0.915	0.110	0.592	0.000
	Model 2							
	OSCE	0.351			0.789	0.125	0.510	0.000
	PCE	0.371		0.020	0.158	0.078	0.164	0.045

Abbreviations: CCS, computer-based case simulation; MCQs, multiple-choice questions; SAQs, short-answer questions; OSCE, objective structured clinical examination; PCE, real patient-based clinical encounter.

internal consistency reliability and relationships between the students' scores in the CCS exam and their scores using other tools of assessment such as MCQs, SAQs, OSCE, and PCE. The main findings in this study are the presence of high levels of internal consistency reliability of the CCS exam scores and significant positive correlations between the different components of the CCS exam scores. In addition, we have demonstrated that the students' scores in clinical reasoning and management are strongly predicted by the scores of the OSCE while the diagnostic performance scores are predicted better by the MCQs.

In this study, the OSCE scores predicted 33% and 35% of the variance in the scores of the clinical reasoning and patient management components of the CCS exam, respectively. These data indicate that the CCS examination tool is testing a construct which is related to what is tested in OSCE. These findings are in line with a study done by Oliven et al, where a total of 262 students were evaluated with both exam modalities (CCS and OSCE). The correlation between the two exam scores was highly significant, with a high internal consistency reliability of the CCS exam scores. They concluded that the CCS is a reliable assessment tool, with the advantage of also providing a training modality.[8]

The OSCE in this study consisted of ten stations testing different aspects of clinical skills including history taking, physical examination, communication skills, and procedural skills on models as well as on standardized patients. Even though OSCE is implemented at many institutions, there have been some concerns over the validity and reliability of this assessment tool.[12,13] Auewarakul et al concluded that there is sufficient validity evidence to support the utilization of OSCEs in internal medicine rotation.[14] An explanation 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.[15] In our study, the importance of OSCE scores as predictors of CCS scores indicates that the practice of the students on the DxR cases includes different sets of constructs of knowledge recall and its application rather than simply testing clinical problem-solving skills.

In the current study, the PCE scores did not significantly predict the students' scores in the patient management or clinical reasoning components of the CCS exam. These data indicate that the CCS as an examination tool is not testing the same construct which is tested by clinical skills exams. A previous study examining the same web-based simulated-patient case software (DxR Clinician) found no correlation between students' scores in clinical reasoning and diagnostic performance and their scores in a Diagnostic Thinking Inventory, a validated tool for diagnostic reasoning.[7]

Alternatively, Triola et al have shown that simulated patient encounters using computer-based virtual patients had an equivalent impact on learners when compared with those exposed to live cases. They also reported that objective measures of performance, knowledge, and diagnostic abilities were equivalent between live and virtual standardized patients, and that the virtual patients may be superior in certain specific applications.[16] Furthermore, several studies compared novice learners' to expert learners' performances in simulation exercises. These studies indicated that simulation exercises are valid measures of clinical proficiency that could differentiate novice from expert.[17–19] The relatively low internal consistency reliability of the PCE scores in the current study is another factor which could affect the validity-related evidence of the scores emerging from this tool. Other factors such as the differences in the examination environment in patient-based versus computer-based settings and the subjectivity in applying the evaluation criteria by the clinical examiners in PCE could also be involved.

We have demonstrated that MCQs predicted 15% of the variance in diagnostic performance scores. This component of the CCS examines the cognitive processes during the patient encounter, which require the student to possess core knowledge related to the case and apply this knowledge in making decisions about diagnosis, physical examination, ordering laboratory investigations, and arriving at a final diagnosis. Therefore, the significant relationship between the students' scores in CCS and MCQs could be explained by the shared components of cognitive domains common in both test formats. A previous study also indicated that the CCS correlated modestly with the multiple-choice component of the USMLE.[9] Feldman et al have shown that, CCS assessment tool provides objective information that can complement a student's NBME score and course grade and may assist in evaluating clinical problem-solving ability.[10]

A limitation of this study is that students were evaluated on three cases only out of six multisystem integrated cases in this PBL unit. Taking into consideration that clinical reasoning is case specific[20] and that the study was conducted in one PBL medical school, the generalizability of the study findings to other settings is limited. Future studies should address the latent constructs emerging from the assessment instruments, including CCS, in PBL programs and how each construct relates to the common domain of competence assessment.

Conclusion

We conclude from this study that the scores from the CCS (DxR Clinician) exhibit good internal consistency reliability

and that they measure a student's clinical competence in a construct related to those measured by OSCEs. However, it appears that the CCS scores measure a construct which is different from that measured through real patient encounter examinations.

Author contributions

MF: coordinated the computer-based clinical reasoning examination for students. Collected the data and conducted the statistical analysis. Drafted the first version of the manuscript. SEK: initiated the study idea and designed the study protocol. Revised the manuscript critically for important intellectual content. Both authors approved the final form of the manuscript.

Disclosure

References

1. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357:945–949.
2. Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ*. 2007;41(12):1133–1139.
3. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Med Educ*. 2010;44(1):94–100.
4. Groves M. Understanding clinical reasoning: the next step in working out how it really works. *Med Educ*. 2012;46:444–446.
5. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med*. 2000;75(2):182–190.
6. Association of American Medical Colleges. *Effective Use of Educational Technology in Medical Education: Summary Report of the 2006 AAMC Colloquium on Educational Technology*. Washington, DC: AAMC; 2007:7. Available from: https://members.aamc.org/eweb/DynamicPage.aspx?webcode=PubsByTopic. Accessed December 22, 2014.
7. Jerant AF, Azari R. Validity of scores generated by a web-based multimedia simulated patient case software: a pilot study. *Acad Med*. 2004;79(8):805–811.
8. 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. *Stud Health Technol Inform*. 2011;169:233–237.
9. Dillon GF, Clyman SG, Clauser BE, Margolis MJ. The introduction of computer-based case simulations into the United States medical licensing examination. *Acad Med*. 2002;77(10 Suppl):S94–S96.
10. Feldman MJ, Barnett GO, Link DA, Coleman MA, Lowe JA, O'Rourke EJ. Evaluation of the Clinical Assessment project: a computer-based multimedia tool to assess problem-solving ability in medical students. *Pediatrics*. 2006;118(4):1380–1387.
11. Zieky MJ. So much has changed: how the setting of cutscores has evolved since the 1980s. In: Cizek GJ, editor. *Setting Performance Standards: Concepts, Methods, and Perspectives*. Majwah, NJ: Lawrence Erlbaum Associates; 2001:19–51.
12. [No authors listed]. Editorial – inverting the pyramid. *Adv Health Sci Educ Theory Pract*. 2005;10(2):85–88.
13. Barman A. Critiques on the Objective Structured Clinical Examination. *Ann Acad Med Singapore*. 2005;34(8):478–482.

14. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. *Med Educ*. 2005; 39(3):276–283.

15. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med*. 2008;40(8):574–578.

16. Triola M, Feldman H, Kalet AL, et al. A randomized trial of teaching clinical skills using virtual and live standardized patients. *J Gen Intern Med*. 2006;21(5):424–429.

17. Adler MD, Trainor JL, Siddall VJ, McGaghie WC. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. *Ambul Pediatr*. 2007;7(2):182–186.

18. Girzadas DV Jr, Clay L, Caris J, Rzechula K, Harwood R. High fidelity simulation can discriminate between novice and experienced residents when assessing competency in patient care. *Med Teach*. 2007;29(5):472–476.

19. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW. Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: a pilot study. *Acad Med*. 2003; 78(10 Suppl):S45–S47.

20. Kreiter CD, Bergus G. The validity of performance-based measures of clinical reasoning and alternative approaches. *Med Educ*. 2009; 43(4):320–325.