Measuring Clinical Reasoning Competency Using a Virtual Patient Model

DANE M. CHAPMAN, MD, PhD[1,2,3]
JUDITH G. CALHOUN, PhD, MBA[1]
ADRIAN P. VAN MONDFRANS, PhD[2]
WAYNE K. DAVIS, PhD[1]

Institutional Affiliations:

1. *University of Michigan Medical School, Department of Medical Education, Ann Arbor, MI*
2. *Brigham Young University, McKay School of Education, Department of Instructional Psychology and Technology, Provo, UT*
3. *University of Missouri-Columbia, School of Medicine, Center for Clinical Reasoning and Procedural Competency, Department of Emergency Medicine, Columbia, MO*

Abstract

Background:
Physicians must be thorough yet efficient in data gathering and must use decision-making strategies that limit diagnostic studies and costs, but still promote maximal diagnostic proficiency. These clinical reasoning skills are neither adequately taught nor measured in medical schools and residencies.

Objective:
To define clinical reasoning constructs *a priori* and develop clinical reasoning indices to be used with a virtual patient simulation model for teaching and assessing clinical reasoning competency.

Methods:
We used an experimental, pretest-posttest design to assess expected gains in clinical reasoning competency after three hours of virtual patient simulation practice. Computer transcripts (N=486) were generated by 81 medical students with complete data who solved one pretest, three practice, and two posttest simulations.

Results:
Four clinical reasoning constructs were identified *a priori*: proficiency, efficiency, thoroughness, and strategy, and nineteen clinical reasoning indices were defined. Multivariate ANOVA and correlational analyses revealed significant pretest-posttest differences for posttest 1 (13/19 indices) and posttest 2 (14/19 indices), supporting the instructional effectiveness of virtual patient simulation practice and the construct validity of four clinical reasoning constructs and their corresponding nineteen clinical reasoning performance indicies. Reliability (stability) and concurrent validity of indices varied with case content.

Conclusions:
Instructional effectiveness, validity and stability of four constructs and nineteen corresponding clinical reasoning indices were established for a computer-based, free-inquiry virtual patient simulation model.

Keywords:
virtual patient, simulation, clinical reasoning, clinical decision-making, competency, assessment

Introduction

High-fidelity, virtual-reality training simulations are increasingly being used for procedural training until proficiency is reached, and before allowing trainees to perform certain high-risk procedures on patients.[1-4] The Federal Drug Administration (FDA) endorsement of procedure-based simulation training is expected to cause a ripple effect throughout all of medicine.[1] Traditional methods of procedural training, including practicing upon patients, will no longer be acceptable as currently performed. While it is doubtful that the use of patients for training will ever be completely substituted with simulations, physicians

will be held to higher standards of training and remediation to reduce medical errors, just as pilots have been mandated with flight simulators.[2-6]

Despite the popularity and rapid advance of procedure-based simulations in medicine, the application of cognitive-based virtual patient simulations has been noted by some experts to be stuck in time.[3-5] The "marvelous medical education machine," a complete simulator for medical education as described by Friedman,[3] has yet to be built. As its potential impact upon medical education and patient care quality is every bit as powerful as the impact of the flight simulator upon aviation, the marvelous computer will likely be built, though probably not all at once.[3-6] The ultimate virtual patient simulator will be high-fidelity—meaning it will faithfully simulate the actual physician-patient encounter. It will also be free-inquiry—meaning users can access data freely without menus or other branching limitations and without cues. Rather than text or verbal descriptions of physical exam and diagnostic test findings, actual visual and auditory responses will be provided, such as visual cues for skin rashes, cardiac and respiratory sounds, and digital images for electrocardiographs (EKGs) and radiographs. While the USMLE® step 3® computer-based case simulation exam has made notable strides in this regard, it is not the ulitmate virtual patient simulator and it still has branching and cueing limitations.[7-10]

Our aims in this study were to implement a high-fidelity, free-inquiry virtual patient simulation (VPS) model into the medical school curriculum to teach clinical reasoning (CR) skills, and then develop a scoring rubric using the VPS model as an assessment tool for measuring data-gathering and decision-making CR competencies. Specifically, we hypothesized that: (1) three hours of VPS practice with feedback would significantly impact CR competency as measured by VPS assessments, (2) CR learning constructs could be identified, and a corresponding scoring rubric of CR indicies developed to detect expected gains in CR competency, (3) Certain CR construct(s) would be case content dependent and represent "medical knowledge" and other CR constructs would be independent of any VPS case content effect, representing underlying CR "process skills", (4) stability of CR constructs (and their corresponding CR indicies) across VPS cases of varying content could be taken as a measure of reliability, (5) construct validity of CR indices would be supported if indices detected expected pretest-posttest gains (e.g., construct validity here refers to whether an index correlates with the theorized learning construct, such as "clinical reasoning proficiency," that it purports to measure), and (6) concurrent validity of CR indices would be supported if indices from the same CR construct correlated more highly than indices from different CR constructs, and the two measures were taken at the same time.[11]

Methods

Study Design:

We used an experimental pretest, posttest control group design to assess expected gains in CR competency after three hours of VPS practice. To address the effects of medical information (content) upon clinical reasoning (process), pretest-posttest and practice-posttest cases of similar and dissimilar content domains were utilized as controls.

Study Setting and Population:

The study qualified for institutional review board (IRB) exemption as a curriculum innovation project and was conducted at the Taubman Health Sciences Library Learning Resource Center of the University of Michigan Medical School. Ninety-seven of 191 post-second-year medical students volunteered without compensation to participate in a computer simulation (CS) elective during a required, four-week problem-based learning curriculum (PBLC). The PBLC occurred between the preclinical and clerkship years with 23-25 CS participants being randomly assigned to each PBLC week from May 7 to June 1 after their second year.

Computer Simulation Elective:

The 6.5 hour CS elective included two sessions (3.0 and 3.5 hours) on Monday-Wednesday, Tuesday-Thursday or Wednesday-Friday mornings during which students worked through six VPSs: one 60-minute pretest (cardiology), three 60-minute practice simulations with corrective feedback (pediatric endocrinology, infectious disease and pulmonary), and two 45-minute posttests (pulmonary and cardiology). No corrective feedback was provided for pretest or posttest assessment simulations. Students were randomly assigned to work in groups of three or individually during practice simulations only. All students completed their pretest and posttest simulations as individuals.

Virtual Patient Simulations:

The multi-problem, network-based VPSs used in the study simulated the actual physician-patient encounter with high fidelity and free inquiry and included 21 patient problems among the six cases.[12] Following an "opening scene," users assumed the role of physicians and moved to and from history, physical examination, diagnostic study, diagnosis and treatment sections without menu-driven cueing or branching limitations.[13] The VPSs were not the ultimate virtual patient, however, as artificial intelligent responses to all history, physical exam and diagnostic test inquiries were provided as text, and not virtual touch, sound or images.

Assessments and Procedure:

Computer transcripts (N=486) were generated by 81 medical students with complete data, and documented student-computer interactions for 243 hours of medical student practice and 202 hours of assessment. Outcome performance scores along nineteen predetermined CR indices (dependent variables) were derived from 243 hard-copy computer assessment transcripts (one pretest and two posttests). To standardize transcript scoring, coding regulations were developed using sample transcripts. Case-specific VPS scoring protocols provided a summary of those expert-recommended critical inquiries that had been made. Diagnosis sections were independently scored by two individuals using case-specific coding regulations that identified acceptable synonyms for diagnoses. Transcripts were scored by at least one rater who was blinded to pretest-posttest classification, and inter-rater agreement was consistently high (r>.90). While therapeutic and management plans were also computer-scored, these were ignored for the purposes of this study.

Development of Scoring Rubric:

Nineteen CR competency indices were defined *a priori* based upon a review of the medical problem-solving literature and were classified into one of four clinical reasoning constructs: proficiency, efficiency, thoroughness, and strategy (See Table 1).

Clinical reasoning proficiency referred to how effectively critical data were gathered and correct diagnoses made. The CR proficiency indices were: percent of critical data-gathering inquiries obtained for history (history proficiency), physical examination (physical examination proficiency), and diagnostic tests (diagnostic test proficiency); percent of correct diagnoses made (diagnosis proficiency); Problem Solving Index (PSI)–an

TABLE 1: Mathematical Descriptions of Nineteen Clinical Reasoning Performance Indices Derived for Use in Multi-Problem Virtual Patient Simulations

Index	Abbreviation	Description[a]
Proficiency		
History Taking	HTP	(Obtained CHT/Total CHT) X 100
Physical Examination	PEP	(Obtained CPE/Total CPE) X 100
Diagnostic Tests	DTP	(Obtained CDT/Total CDT) X 100
Correct Diagnoses	DP	(Obtained CD/Total CD) X 100
Program Solving Index	PSI	(HTP + PEP + DTP + DP) / 4
Proficiency Index	PI	(Obtained CHT + CPE + CDT) X 100 / (Total CHT + CPE + CDT)
Efficiency		
History Taking	HTE	(CHT Obtained/HTT) X 100
Physical Examination	PEE	(CPE Obtained/PET) X 100
Diagnostic Tests	DTE	(CDT Obtained/DTT) X 100
Thoroughness		
History Taking	HTT	Total HT
Physical Examination	PET	Total PE
Diagnostic Tests	DTT	Total DT
Total Data-Gathering	TDG	(HTT + PET + DTT)
Diagnosis	DT	Total D
Strategy		
History Taking	HTS	[HTT/(HTT+PET + DTT)] X 100
Physical Examination	PES	[PET/(HTT+PET + DTT)] X 100
Diagnostic Tests	DTS	[DTT/(HTT+PET + DTT)] X 100
Focused Strategy Index	FSI	(HH + PP+ DD +1) / (HP + HD + PH + PD + DH + DP + 1)
Invasiveness/Cost Index	ICI	[DTT/(HTT + PET)] X 100

[a]*Symbol Key: HT*= history taking inquiries, *PE*= physical examination inquiries, *DT*= diagnostic test inquiries, *D*= diagnoses indicated, *C*= critical inquiry or diagnosis (e.g. CHT=critical history taking inquiries), *HH*= history to history transition, *PP*= physical exam to physical exam transition, *DD*= diagnostic test to diagnostic test, *HP*= history to physical exam, HD= history to diagnostic test, *PH*= physical exam to history, *PD*= physical exam to diagnostic test, *DH*= diagnostic test to history, and *DP*= diagnostic test to physical exam transition.

average of data-gathering and decision-making proficiencies; and Proficiency Index (PI)–the percent of data-gathering critical information obtained.

Clinical reasoning efficiency was defined as the percentage of data-gathering inquiries that were critical in making the diagnosis of a patient's problem(s). Higher scores represented greater efficiency in making medical inquiries. Clinical reasoning efficiency indices included history, physical examination and diagnostic test efficiencies.

Clinical reasoning thoroughness reflected the frequency of data-gathering inquiries made or diagnoses indicated. Clinical reasoning thoroughness indices included: total number of history inquiries (history thoroughness), physical examination inquiries (physical examination thoroughness), and diagnostic test inquiries (diagnostic test thoroughness); total number of history, physical examination and diagnostic test inquiries combined (total data-gathering thoroughness); and total number of diagnoses hypothesized at the completion of each simulated case (diagnosis thoroughness).

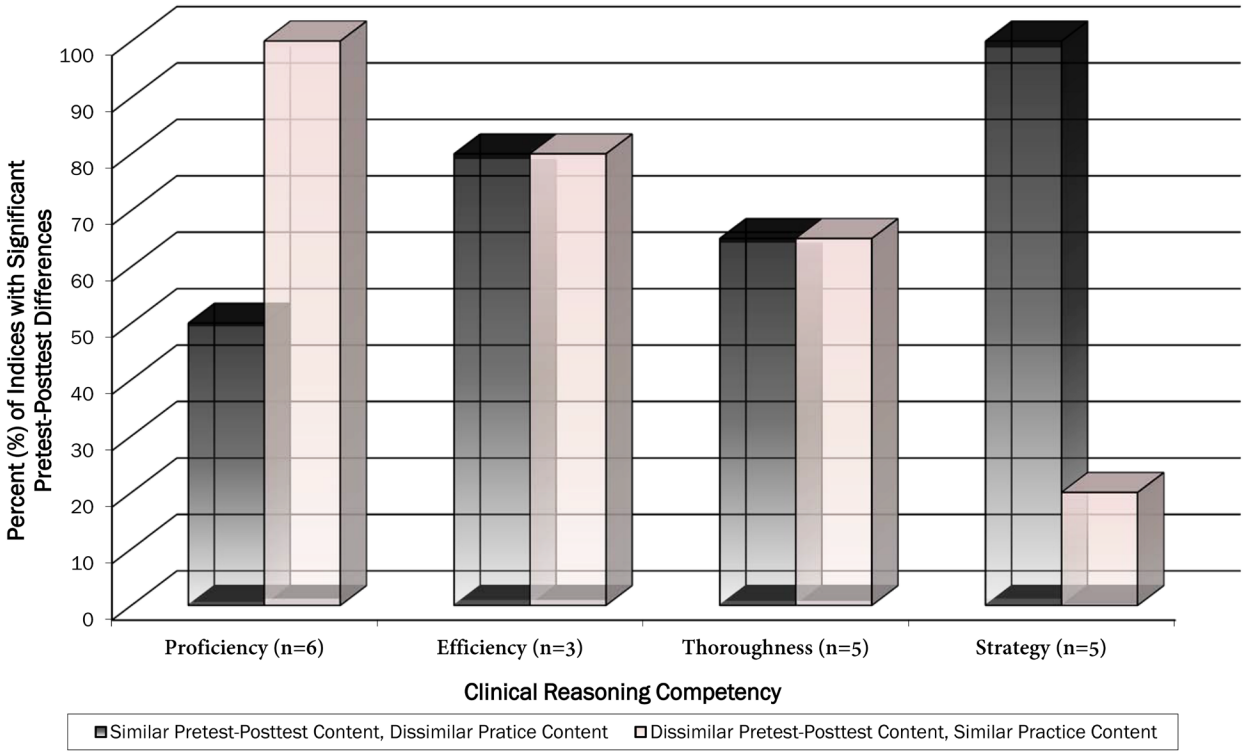
Clinical reasoning strategy referred to the cognitive strategies used to arrive at correct diagnoses. It reflected individual preference for certain data-gathering techniques (e.g. to use either a focused inquiry approach or a "shot gun" or haphazard approach). CR strategy indices included: percent of total data-gathering inquiries that relate to history taking (history strategy), physical examination (physical examination strategy), or diagnostic test (diagnostic test strategy); Focused Strategy Index—the ratio of data-gathering inquiry transitions of similar type (e.g. history to history) to all other combinations of possible inquiry transitions from one type of inquiry to another (e.g. history to physical examination, diagnostic test to history, etc), where high scores reflect a more focused and systematic data-gathering approach; and Invasiveness/Cost Index—the ratio of diagnostic test inquiries (relatively invasive and costly) to the sum of history-taking and physical examination inquiries (relatively non-invasive and less costly), where higher scores

TABLE 2: Pretest-Posttest Means (SD) for Nineteen Clinical Reasoning Indices Across Virtual Patient Simulations of Similar and Dissimilar Case Content (N=81)[a]

Index	Case 1 Pretest (Cardiology)	Case 5 Posttest (Pulmonary)	Case 6 Posttest (Cardiology)
Proficiency			
History Taking	45.1 (24.8)	56.2 (19.1)[c]	49.0 (15.0)
Physical Examination	50.9 (22.9)	67.1 (17.9)[c]	62.2 (19.5)[c]
Diagnostic Tests	58.6 (26.0)	51.7 (12.2)[b]	52.5 (16.7)[b]
Correct Diagnoses	37.0 (26.4)	46.9 (21.4)[c]	30.2 (17.5)[b]
Program Solving Index	47.9 (14.9)	55.5 (10.4)[c]	48.5 (9.7)
Proficiency Index	51.6 (16.1)	56.2 (10.8)[b]	53.1 (10.1)
Efficiency			
History Taking	8.2 (4.5)	19.3 (8.5)[c]	25.1 (10.9)[c]
Physical Examination	17.7 (8.5)	17.0 (10.2)	21.9 (10.9)[c]
Diagnostic Tests	35.4 (22.7)	43.0 (16.1)[c]	36.1 (13.6)
Thoroughness			
History Taking	23.7 (9.7)	26.1 (10.9)[b]	22.4 (10.1)
Physical Examination	12.2 (5.9)	15.0 (7.5)[c]	15.9 (6.2)[c]
Diagnostic Tests	8.4 (4.6)	9.4 (3.7)[b]	14.5 (6.4)[c]
Total Data-Gathering	44.4 (12.8)	50.5 (16.2)[c]	52.8 (14.2)[c]
Diagnosis	3.0 (1.2)	3.1 (1.1)	4.9 (1.8)[c]
Strategy			
History Taking	52.7 (12.4)	51.0 (9.9)	41.3 (13.5)[c]
Physical Examination	27.5 (8.8)	28.8 (8.6)	30.2 (9.6)[b]
Diagnostic Tests	19.8 (11.0)	20.2 (8.6)	28.4 (11.7)[c]
Focused Strategy Index	4.8 (3.3)	8.9 (5.7)[c]	9.0 (6.0)[c]
Invasiveness/Cost Index	27.6 (21.1)	26.9 (14.8)	44.3 (29.4)[c]

[a]Repeated-Measures ANOVA for Pretest - Posttest Comparisons for 81 medical students with complete data; [b]p = .050; [c]p = .010

FIGURE 1: Percent of Clinical Reasoning Indicies (N=19) with Significant Pretest-Posttest Differences Following Three Hours of Multi-Problem Virtual Patient Simulation Practice in Post-Second Year Medical Students (N=81)(ANOVA, F Test)



reflect a more invasive and costly data-gathering approach.

Data Analysis:

BMDP multivariate, factorial, repeated measures ANOVA statistics were used to determine any overall effect of three hours of VPS practice upon the four clinical reasoning constructs (proficiency, efficiency, thoroughness, and strategy), while controlling for CS weeks, group or individual practice, and a justifying strategy. CR constructs with significant multivariate effects were further defined using univariate ANOVA or ANCOVA of the individual indices of the constructs. Expected pretest-posttest gains along the nineteen CR indices were also taken as a measure of construct validity. Correlation analyses were used to determine the reliability (stability) of nineteen CR indices across cases of similar and dissimilar content and the concurrent validity of these indices in measuring one of the four CR constructs. (See also http://www.statistical-solutions-software.com/bmdp-statistical-software/bmdp/).

Results

The study sample (N=97) appeared to be representative of the entire medical school class (N=191) as VPS students did not differ significantly from other class members in ethnicity, sex, prior clinical experience on the hospital wards, or independent PBLC CR assessments (P>.050, ANCOVA). Approximately two-thirds of the VPS enrollees had never previously participated in computer-based instruction and almost one-fourth had never interacted with a computer in any capacity at the time of the original study,[13] making a selection bias, which favored students who were more comfortable with using computers for learning, unlikely.

Effect of VPS Practice:

Repeated measures factorial ANOVA analyses revealed significant pretest-posttest differences between the pretest and first posttest (13/19 indices) and between the pretest and second posttest (14/19 indices), supporting the instructional effectiveness of only three hours of VPS practice (See Table 2). It is unlikely that VPS pretest-posttest differences were a result of the PBL curriculum alone as there was no difference in PBLC CR assessments between VPS enrollees and the remainder of the medical school class during each week of the PBLC (P>.050, ANCOVA).

Effect of VPS Case Content:

Proficiency indices demonstrated pretest-posttest gains that were most notable when practice and posttest content were similar (pulmonary-pulmonary). Efficiency and thoroughness indices demonstrated significant pretest-posttest differences

TABLE 3: Correlations[a] of Nineteen Clinical Reasoning Performance Indices Across Computer Simulations of Similar and Dissimilar Case Content (N=81)[b]

Index	C1 Cardiology C5 Pulmonary	C1 Cardiology C6 Cardiology	C5 Pulmonary C6 Cardiology
Proficiency			
History Taking	.11	.23[c]	.21
Physical Examination	-.01	.23[c]	-.06
Diagnostic Tests	.10	.28[d]	-.03
Correct Diagnoses	-.03	.23[c]	.10
Program Solving Index	.03	.58[d]	.02
Proficiency Index	.11	.52[d]	.08
Efficiency			
History Taking	.06	.09	.35[d]
Physical Examination	.06	.29[d]	.19
Diagnostic Tests	.11	-.16	.24[c]
Thoroughness			
History Taking	.52[d]	.50[d]	.70[d]
Physical Examination	.44[d]	.43[d]	.63[d]
Diagnostic Tests	.42[d]	.43[d]	.63[d]
Total Data-Gathering	.53[d]	.50[d]	.70[d]
Diagnosis	.11	.31[d]	.13
Strategy			
History Taking	.35[d]	.32[d]	.54[d]
Physical Examination	.24[c]	.18	.53[d]
Diagnostic Tests	.50[d]	.40[d]	.67[d]
Focused Strategy Index	.41[d]	.35[d]	.50[d]
Invasiveness Index	.50[d]	.40[d]	.65[d]

[a]Pearson Product-Moment Correlations; [b]C1=Case 1 (Card. Pretest), C5=Case 5 (Pulm. Posttest), C6=Case 6 (Card. Posttest); [c]p = .050; [d]p = .010

regardless of case content, suggesting their stability across cases and their relation to underlying CR process skills. Strategy indices demonstrated the greatest pretest-posttest differences when posttest content was different from practice content (pulmonary-cardiology) (See Figure 1). Students became more focused in their problem-solving approach from pretest to posttest simulations as evidenced by significant improvements on the Focused Strategy Index (p=.010) regardless of case content. However, when content was unfamiliar—had not been taught during a virtual patient practice session—students used a significantly more invasive and costly problem-solving approach and relied less upon history taking and physical examination as evidenced by the Invasiveness/Cost Index (p=.010) (See Table 2).

Construct Validity, Concurrent Validity, and Strategy (Reliability):

Construct validity of the four CR constructs (proficiency, efficiency, thoroughness, and strategy) was supported by expected pretest-posttest gains after three hours of VPS practice. Concurrent validity[11] of the four CR constructs was suggested, as indices from each contruct tended to behave similarly with regard to case content and pretest-posttest effect. Higher correlations were noted among proficiency indices as expected when case content was similar (Case 1: Cardiology and Case 6: Cardiology) and were greater than pretest-postetest correlations (Case 1 and Case 5; Case 1 and Case 6; see Table 3). Concurrent validity of efficiency, thoroughness, and strategy indices was supported by generally higher correlations between the two posttests than between either posttest and the pretest (See Table 3). Concurrent validity is demonstrated when a test correlates well with a measure that has been (previously or simultaneously) validated for the same construct, or for different, but presumably related, constructs, and the two measures are taken at the same time. This is in contrast to predictive validity, where one measure occurs earlier and is meant to predict some later measure.[12] Between case correlations remained moderate to high, regardless of case content, for thoroughness and strategy indices, suggesting higher reliability (stability) of these indices across cases. Reliability of

efficiency indices was less well-supported as correlations were inconsistent across case content.

Discussion

The free-inquiry VPS model, including its validated CR constructs (proficiency, efficiency, thoroughness and strategy) and nineteen CR indicies, has proven useful as both a teaching and assessment tool. We found the teaching utility of the model so profound that even as little as three hours of VPS practice resulted in significant pretest-posttest differences for many of the CR indices. This is not to say that our novice pre-clerkship students had achieved CR competency. Their mean scores remained far below expected competenecy even if defined at a 70-percent cutoff for CR proficiency indices. These results help to elucidate those aspects of CR that can be taught as process skills independent of knowledge content, and may help to resolve some of the CR teaching and assessment chaos described by Norman[14] and Elstein.[15]

Considerable CR occurs in the earliest stages of the patient presentation. Generating correct diagnostic hypotheses (i.e. hypothesis generation) has been shown to be significantly related to the patient's chief complaint and history, while physical examination and diagnostic studies contributed less to generating correct hypotheses than to eliminating alternatives (i.e. hypothesis confirmation/exclusion).[16] Moreover, students who failed to list the correct diagnosis in the differential diagnosis after obtaining the history were significantly less likely to reach the correct diagnosis at the end of the case, suggesting the critical importance of the history in medical problem solving.[17] The fact that our novice, pre-clerkship medical students had relatively low diagnosis proficiency scores compared to their data-gathering proficiencies is consistent with this finding. With their heads full of isolated, unassimilated medical facts, not organized around clinical scenarios or schemata, students did not have the key concepts or clinical features of disease patterns assimilated sufficiently to prompt their history inquiries. Still, the VPS model and CR indices were sensitive enough to detect pretest-posttest gains in both history-taking proficiency and diagnosis proficiency when content was familiar to students. These results are consistent with previous research demonstrating that medical decision-making expertise is related to one's ability to recognize content-specific disease patterns ("illness scripts") and to perceptual and cognitive skills, and that expertise is more dependent upon hypothesis generation through history taking than upon hypothesis confirmation through physical examination and diagnostic testing.[14-16]

Our results confirm that some CR skills can be enhanced or learned independent of case content, namely CR efficiency, thoroughness, and strategy. However, it is less clear which efficiency, thoroughness, or strategy adjustments would be most rewarding in terms of improved diagnostic decision making. Wolf et.al.[18] found that learning to use a competing hypothesis strategy enhanced medical problem-solving performance independent of case content.

In training clinical decision makers, medical schools and residency training programs typically emphasize thoroughness. However, the more thorough physician is not always the most expert (i.e. accurate or proficient) at clinical decision-making.[19] Increasingly, thoroughness has been taken to mean "ordering more diagnostic tests" rather than being thorough in history taking or in conducting a thorough physical examination. David Sklar,[20] in his editorial "Beginning the Journey" as the new editor-in-chief of Academic Medicine, has noted that "CT scans and ultrasounds have virtually replaced the traditional physical examination, and computers have invaded the consultation room, interposing themselves between the clinician and the patient, diverting the clinician's attention from conversations with the patient to the documentation requirements demanded by payers and employers." This is a worrisome trend that threatens our professional identity as health care providers. The relationship "between the healer and the sick, the most sacred, core responsibility and privilege in medicine" is being threatened.[20]

In our attempt to teach and assess core competencies through VPSs, we must be on guard not to lose the sacred trust of our patients. It seems contradictory to teach physician-patient interactions using computer-based technologies that may be the very cause of our eroding physician-patient relationships. However, if properly designed, VPS could be useful in teaching and assessing professionalism and the other core competencies identified by the Accreditation Council on Graduate Medical Education (ACGME).[21] The VPS model and CR indices could also be implemented to study intervention effects upon CR competency. It has been suggested that decision making could be enhanced and its teaching facilitated if disease-specific, data-gathering elements were identified and characterized as most consistent and predictive of each competing diagnostic hypothesis. Understanding the optimal disease-differentiating pivotal elements, key concepts, features[14-16] and knowledge structures[14] would seem to significantly augment acquisition of clinical reasoning skills—especially when programmed into virtual patient simulators.[3-5, 22-23] Developers of newer generation, virtual patient simulators would also do well to incorporate the free-inquiry approach, without cueing or branching limitations. Such cognitive-based simulators would also be most useful if they incorporated an artificial intelligence function that responded to user treatments in disease-predictable ways, such that users are able to perform "what if" inquiries as they learn.[3-5, 22-24]

This study has limitations. It was conducted nearly three decades ago as part of a PhD dissertation,[13] and was never formally published. With recent developments in the ACGME

core competencies,[21] new accreditation system (NAS) and milestones,[25] the *a priori* development and validation of CR constructs with a scoring rubric using free-inquiry VPSs has greater relevance now than thirty years ago. VPSs have changed in some ways that might impact study results. However, one could argue that the free-inquiry capability of the VPS model in this study is the gold standard which has yet to be achieved by the USMLE® step 3® computer-based or OSCE-based exam.[5-10] Further study is needed to apply generalizability analysis of the scoring rubric to better understand inter-case variability. Generalizability refers to external validity and is limited when the cause or independent variable (e.g., three hours of VPS practice) is influenced by other factors—all threats to external validity or generalizability interact with the independent variable.[26-27] Although this study was conducted at a single institution at a single point in time some years ago, more than half of a large medical school class participated, and the results of this study would be expected to generalize to other post-second year medical students with similar aptitudes and experiences. It is less clear whether results would generalize to medical students in their clinical years or to residents and physicians.

In summary, four clinical reasoning constructs of proficiency, efficiency, thoroughness and strategy were defined *a priori* and validated using a high-fidelity, free-inqiry, computer-based virtual patient simulation model. With ever-changing protocols and increasing medical knowledge, VPS may be helpful in positioning medical students and trainees for life-long learning as part of their daily clinical practice.[21, 24-25] If the ultimate goal for incorporating VPSs into all levels of medical education is to promote improved quality of care for patients,[1-2] while regaining a new sense of commitment to the clinician-patient relationship,[20] then we will ultimately succeed in building the marvelous medical education machine. After thirty years of processing and assimilation, the VPS machine may be capable of both teaching CR skills and producing a scoring rubric that can detect subtle differences in clinical data-gathering and decision-making core competencies.

Received for publication: August 20, 2012. Revisions received: February 20, 2013. Accepted for publication: February 20, 2013. Available online: February 21, 2013.

Presented in part at the Society for Academic Emergency Medicine (SAEM) Annual Meeting, San Antonio, Texas, May 21-24, 1995; A National Symposium on Microcomputers in Health Care Education, Omaha, Nebraska, April, 1985.

Contributions of Authors: Conception (DMC, JGC, APV, WKD), design (DMC, JGC), analysis and interpretation (DMC, JGC), drafting (DMC, JGC, WKD), and revising (DMC) the article. DMC takes responsibility for the paper as a whole.

Acknowledgements: The authors are indebted to Robert H. Bartlett, MD and Richard G. Judge, MD, course directors of the Problem-Based Learning Curriculum, for their enthusiastic endorsement of the medical problem solving computer simulation elective; the ninety-seven University of Michigan medical students who enrolled in the elective; the Taubman Medical Library Learning Resource Center staff for their essential role in making the computer simulation elective an overwhelming success, namely: Cindy A. McBride, Patricia Tomlin, Ellen S. Hoffman, Achla B. Karnani, Chris M. Chapman, Denise M. Axelrod, Patricia W. Martin, Dennis M. Poupart, Jackie E. McKenzie, Lois L. Katon, Mark D. Heidt, William A. Vicini, and Sharon J. Love; and, Fredric M. Wolf, PhD, Elaine M. Hockman, PhD, Nancy P. Allen, PhD, C. Victor Bunderson, PhD, Russell T. Osguthorpe, PhD and Sally H. Cavanaugh, PhD for their insightful review and recommendations to improve this work; and, to Becky Bluett and Christine Downs for their assistance in preparing the manuscript.

Funding: In part by a University of Michigan curriculum innovation grant.

Conflicts of Interest: No personal, commercial, political, academic or financial conflict of interest was reported by any of the authors. DMC is editor-in-chief of JCRPC.

Ethical Approval: Authors report the study qualified for IRB exemption.

Correspondence: Dane M. Chapman, MD, PhD, Director, Center for Clinical Reasoning and Procedural Competency, Department of Emergency Medicine, University of Missouri-Columbia School of Medicine, 1 Hospital Drive, Columbia, MO, 65201, e-mail: <chapmandan@health.missouri.edu>

References

1. Gallagher AG, Cates CU. Approval of virtual reality training for carotid stenting: what this means for procedure-based medicine. *JAMA* 2004 Dec 22;292(24):3024-26.
2. AAMC Report of the Ad Hoc Committee of Deans. Educating doctors to provide high quality medical care. 2004: 1-12. Available from: https://members.aamc.org/eweb/upload/Educating%20Doctors%20to%20Provide%20July%202004.pdf.
3. Friedman CP. The marvelous medical education machine or how medical education can be unstuck in time. *Acad Med.* 2000;75(10):S137-142.
4. Cook DA, Triola MM. Virtual patients: A critical literature review and proposed next steps. *Med Educ*. 2009;43(4):303-11.
5. Courteille O, Bergin R, Stockeld D, Ponzer S, Fors U. The use of a virtual patient case in an OSCE-based exam—a pilot study. *Med Teach.* 2008;30(3):e66-76.
6. Noble C. The relationship between fidelity and learning in aviation training and assessment. *J of Air Transport.* 2002 7(3):33-54.
7. Feinberg RA, Swygert KA, Haist SA, Dillon GF, Murray CT. The impact of postgraduate training on USMLE® step 3® and its computer-based case simulation component. *J Gen Intern Med.* 2012 Jan;27(1):65-70.
8. Harik P, Cuddy MM, et al. Assessing potentially dangerous medical actions with the computer-based case simulation portion of the USMLE step 3 examination. *Acad Med.* 2009 Oct;84(10 Suppl):S79-82.
9. Andriole DA, Jeffe DB, Hageman HL, Whelan AJ. What predicts USMLE Step 3 performance? *Acad Med.* 2005 Oct;80(10 Suppl):S21-4.
10. Margolis MJ, Clauser BE, Harik P. Scoring the computer-based case simulation component of USMLE Step 3: a comparison of preoperational and operational data. *Acad Med.* 2004 Oct;79(10 Suppl):S62-4.
11. McIntire SA and Miller LA, *Foundations of Psychological Testing*, 2nd Edition, Thousand Oaks, CA:Sage Publishing Co., 2005.
12. Harless W. CASE: A computer-assisted simulation of the clinical encounter. *J of Med Educ.* 1971;46:443-448.
13. Chapman DM. Teaching and evaluating clinical reasoning through computer-based patient management simulations. *Dissertation Abstracts International.* 1985;46:784-B.
14. Norman GR. The epistemology of clinical reasoning. *Acad Med.* 2002;75-S127-133.

15. Elstein AS. Clinical problem solving and decision psychology: Comment on "The epistemology of clinical reasoning." *Acad Med*. 2002;75(10):S134-36.

16. Gruppen LD, Woolliscroft JO, Wolf FM. The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. *Proc Annu Conf Res Med Educ.* 1988;27:242-7.

17. Gruppen LD, Palchik NS, Wolf FM, et al. Medical student use of history and physical information in diagnostic reasoning. *Arthritis Care Res*. 1993;6(2):64-70.

18. Wolf FM, Gruppen LD, Billi JE. Use of the competing-hypotheses heuristic to reduce "pseudodiagnosticity'. *J Med Educ.* 1988 Jul;63(7):548-54.

19. Voytovich AE, Rippey RM, Copertino L. Scorable problem lists as measures of clinical judgment. *Eval Health Prof.* 1980;3:159-171.

20. Sklar D. Beginning the journey. *Acad Med.* 2013;88(1):1-2.

21. Chapman DM, Hayden S, Sanders AB et al. Integrating the Accreditation Council for Graduate Medical Education core competencies into the Model of the Clinical Practice of Emergency Medicine. *Ann Emerg Med.* 2004;33(6):756-769.

22. Stevens SM, Goldsmith TE, Summers KL et. al. Virtual reality training improves students' knowledge structures of medical concepts. *Medicine Meets Virtual Reality* 13, James D. Westwood et al. (Eds.), Amsterdam:IOS Press, 2005, pp. 519-525.

23. Deterding R, Milliron C, Hubal R. The virtual pediatric standardized patient application: formative evaluation findings. *Medicine Meets Virtual Reality* 13, James D. Westwood et al. (Eds.), Amsterdam:IOS Press, 2005, pp. 105-107.

24. Jarrell BE. Simulation for teaching decision making in medicine: The next step. Presented at Medicine Meets Virtual Reality 13, Long Beach, CA, Jan 28, 2005.

25. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system — Rationale and benefits. NEJM Special Report. http://www.acgme-nas.org/assets/pdf/NEJMfinal.pdf

26. Shadish W, Cook T, Campbell D. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston:Houghton Mifflin, 2002.

27. Campbell DT, Stanley JC. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally College Publishing Company, 1963.