The validity of performance-based measures of clinical reasoning and alternative approaches

Clarence D Kreiter & George Bergus

CONTEXT The development of a valid and reliable measure of clinical reasoning ability is a prerequisite to advancing our understanding of clinically relevant cognitive processes and to improving clinical education. A record of problem-solving performances within standardised and computerised patient simulations is often implicitly assumed to reflect clinical reasoning skills. However, the validity of this measurement method for assessing clinical reasoning is open to question.

OBJECTIVES Explicitly defining the intended clinical reasoning construct should help researchers critically evaluate current performance score interpretations. Although case-specific measurement outcomes (i.e. low correlations between cases) have led medical educators to endorse performance-based assessments of problem solving as a method of measuring clinical reasoning, the matter of low across-case generalisation is a reliability issue with validity implications and does not necessarily support a performance-based approach. Given this, it is important to critically examine whether our current performance-based testing efforts are correctly focused. To design a valid educational assessment of clinical reasoning requires a coherent argument represented as a chain of inferences supporting a clinical reasoning interpretation.

DISCUSSION Suggestions are offered for assessing how well an examinee's existing knowledge organisation accommodates the integration of new patient information, and for focusing assessments on an examinee's understanding of how new patient information changes case-related probabilities and base rates.

*Medical Education 2009: **43**: 320–325*
doi:10.1111/j.1365-2923.2008.03281.x

Department of Family Medicine, University of Iowa, Iowa City, Iowa, USA

Correspondence: Clarence D Kreiter, College of Medicine, University of Iowa, OCRME 1204 MEB, Iowa City, Iowa 52246, USA.
Tel: 00 1 319 335 8906; Fax: 00 1 319 335 8904;
E-mail: clarence-kreiter@uiowa.edu

INTRODUCTION

The development of a valid and reliable measure of clinical reasoning ability is a prerequisite to advancing our understanding of clinically relevant cognitive processes and to improving clinical education. Although such a measure would promote the accurate evaluation of medical education's instructional methodologies and improve the relevance of in-course and licensure assessments, researchers have discouragingly concluded that it is 'probably close to the truth, to say that we have no method of assessing clinical reasoning which stands up to critical scrutiny'.[1] In the 8 years since these authors offered up this rather pessimistic assessment, significant resources have been invested in constructing performance-based clinical reasoning tests that have been shaped by a number of larger trends within medical education.

BACKGROUND TO PERFORMANCE-BASED ASSESSMENTS OF PROBLEM SOLVING

The general trend in medical education has been to rely less on written testing formats and to increasingly employ performance-based assessment (PBA) methods to assess student achievement. This is also true for the assessment of clinical reasoning where case simulations are used as a form of performance assessment. Although there is little direct evidence to suggest that clinical reasoning PBAs are more valid than written assessments, the motivation to develop the current generation of computerised and standardised patient-based simulations (simulated patients [SPs]) was strengthened by the belief among many medical educators that written testing formats were incapable of assessing the application of medical knowledge (clinical reasoning). With the implicit assumption that performances of clinical problem solving reflect clinical reasoning, PBAs have documented problem-solving behaviour within simulated case encounters. The early performance-based simulations consisted of written patient management problems[2] and the still widely used SP-based examinations; with the advent of new technologies, these have evolved over the last 40 years into the more sophisticated computer-based patient simulations used today.[3]

The assumption that the record of performances of problem solving will reflect clinical reasoning has not been explicitly tested. Although the underlying clinical reasoning construct has important validity implications related to test design, developers of case simulation assessments have not clearly acknowledged or defined this aspect of the test. In fact, simulation-based performance assessment scores are often characterised as simply reflecting patient management skills or success at solving specific clinical problems. Rather than defining an underlying construct, this validity orientation uses only the observable attributes of the task, or a task-centred approach, to establish validity. An observable attribute approach to validity requires that performance on the test's tasks generalises to a universe of similar test tasks and does not use the preferred construct-centred approach that utilises and references an underlying skill or ability that enables clinical competence.[4] Unfortunately, even with this restricted validity framework, generalisability studies suggest problem-solving PBAs are inefficient.[5]

THE IMPORTANCE OF SPECIFYING AND DEFINING THE CONSTRUCT

In developing any assessment, it is important to carefully specify and define the underlying construct or ability that the test is designed to measure. In the case of a clinical problem-solving PBA, specification of the clinical reason construct is important for evaluating validity. For example, one obvious implication of recognising the clinical reasoning construct is that an appropriately designed PBA must present medical cases for which the examinee population has already acquired the background knowledge. If a significant proportion of the examinee population has failed to acquire the background knowledge required for solving the case problems, the variance in problem-solving performance scores will reflect variation in examinees' knowledge, not their skilfulness in applying that knowledge to the clinical problem (the clinical reasoning construct). Performance-based assessments are not an efficient means of assessing clinical knowledge, but without carefully designing such assessments around the target construct, they can easily revert to a test of simple factual recall.

Carefully defining the target construct for an assessment allows researchers to critically examine whether various implied inferences related to scores generated by a test are correct. In the case of medical education's current clinical reasoning PBAs, it is useful here to reconsider validity-related arguments and whether the assessment of clinical problem solving is indeed the most strategic approach for

assessing examinees' clinical reasoning ability. Currently, there is limited validity evidence to establish a statistical or substantive relationship between clinical problem-solving performance measures and the ability to reason within the clinical context. In addition, psychometrically defensible methods for scoring clinical reasoning PBAs have not been developed. Major issues related to scoring the record of behaviours elicited during engagements with clinical problem-solving PBAs need to be resolved, and can be highly problematic without objective consideration of the relative costs and benefits of various 'correct' and 'incorrect' decisions.[6]

THE ROLE OF CASE SPECIFICITY

A key assumption that has led to the development of medical education's current clinical reasoning PBAs relates to the popular interpretation of the so-called 'case-specific' finding. The finding that the statistical correlation between cases in problem-solving assessment scores is very low has led many medical educators to assume that clinical reasoning is highly dependent on the particular case encountered.[7] The case-specific assumption has been ubiquitous in the PBA research literature since the 1970s and has led test developers to model clinical reasoning as dependent on case characteristics and cognitive theorists to model the clinical reasoning construct as a multi-dimensional composite of highly independent or weakly correlated skills. A case-specific interpretation of performance data has additionally provided the rationale for test specialists to record and score case-based clinical problem-solving behaviours in order to make inferences regarding the strength of an underlying multi-dimensional clinical reasoning construct. If one assumes that clinical reasoning is highly case-dependent, or that each medical case emphasises a different composite of moderately independent clinical reasoning skills, the decision to assess examinees over a large number of realistic medical cases makes logical sense. However, the classic interpretation of the case-specific finding appears questionable, and a reanalysis of the evidence has provided an alternative statistical perspective that points to other sources of measurement error as more influential causal agents for low case correlations.[8]

Measurement error unrelated to case characteristics may offer a more accurate explanation for low case correlations and may also imply that clinical reasoning ability is more uni-dimensional than previously believed. A re-interpretation of the so-called 'case-specific' finding requires that we reassess our current thinking regarding why clinical reasoning measures need to be acquired within multiple simulated case contexts, and whether each case actually requires a unique composite of weakly associated clinical reasoning abilities. Although clinical reasoning clearly requires relevant clinical knowledge, it appears debatable whether the application of that knowledge is highly case-dependent, whether there is a strong person–case interaction and, ultimately, whether the assessment of clinical reasoning need occur in the context of measured behaviour within realistic clinical cases. The issue of low across-case generalisation is a reliability issue with validity implications and reflects the test's efficiency. Given this, it is important to critically examine whether our current best developed testing efforts are correctly focused, or whether alternative assessment methodologies might yield a more practical measure capable of displaying a more robust relationship with a clinical reasoning construct.

MEDICAL EDUCATION'S MOST ADVANCED PROBLEM-SOLVING PBA

Primum Clinical Case Simulation® (CCS) software was developed by the National Board of Medical Examiners and is the best developed computerised case simulation PBA testing program in use today.[9] In 1999, after over three decades of research and development and very large financial investment, the CCS was included as part of the United States Medical Licensing Examination (USMLE) Step 3; it has since proven capable of generating modestly reliable scores ($G = 0.61$ for nine cases) with a half-day of testing.[10] The software that runs the testing process is programmed to recognise and evaluate over 2500 actions that can be indicated by the examinee's use of free text commands in simulated time. The scoring is automated with the use of regression-based equations that model a large number of expert ratings.[11] The validity of the CCS score is primarily supported by findings that the automated scores are successful in predicting expert ratings of the same performances ($r = 0.84$), that the cases evoke simulated clinical management behaviours that appear to be of a similar variety to those observed within real clinical encounters, and that the disattenuated correlation with traditional written item formats is < 1.0 ($r_c = 69$).[10,12,13] Although the Primum CCS format represents a unique and positive contribution to the library of measurement techniques used to assess clinical reasoning, there is little evidence to support the usefulness of the resulting

scores. In addition, its programming complexity, lack of strongly supportive validity evidence, low reliability, expensive rater-modelling scoring algorithms, extensive testing time requirements and high development costs render it impractical in most testing contexts. As further validity evidence is needed, such as that documenting the relationship between scores and level of experience and expertise, it is premature to recommend that medical education abandon this problem-solving PBA approach to assessing clinical reasoning. However, it does seem reasonable for medical educators to reassess the allocation of future test development resources targeted at creating a valid and practical test of clinical reasoning ability.

FUTURE RESEARCH AND DEVELOPMENT OF INSTRUMENTS DESIGNED TO MEASURE CLINICAL REASONING

Because attempts to utilise hypothesis generation and data-gathering actions evoked by the interaction with a simulated clinical problem have not proved highly successful in distinguishing excellent diagnosticians from those less skilled,[14,15] it is important to consider other approaches. Cognitive research may provide a basis for the design of alternative methodologies. For instance, qualitative and quantitative research studies utilising both direct and indirect observations have established the interdependence of knowledge organisation and clinical reasoning. Although a review of this research literature is beyond the scope of this discussion, measurement methods, such as the think-aloud protocol, multi-dimensional scaling and the quantitative concept map, have provided key insights that have elucidated the relationship between the level of knowledge organisation and the ability to reason clinically.[16–19] Although research has demonstrated that knowledge organisation increases with experience, the structure of that knowledge is likely to be quite heterogeneous and hence no single structure can be logically regarded as the best representation of a particular topic. The fact that experts display higher levels of knowledge organisation and that its structure is highly idiosyncratic has important implications for those seeking to assess clinical reasoning. Specifically, it suggests that directly assessing an examinee's understanding of specific relationships related to complex clinical knowledge structures is unlikely to provide an objective assessment of clinical reasoning skill. As no 'gold standard' or best knowledge structure is likely to exist, the descriptive measures used in cognitive research are unlikely to be appropriate for objective educational assessment.

To design a valid educational assessment of clinical reasoning requires a coherent argument represented as a chain of inferences supporting a clinical reasoning interpretation.[20] All definitions and models of clinical reasoning suggest that it is a cognitive activity that integrates information from a clinical encounter with an existing system of knowledge organisation. This integration enables diagnostic and patient management decisions and might be a reasonable starting point for defining an explicit chain of supporting validity-based inferences. This very general definition of clinical reasoning suggests three potentially observable and measurable aspects related to the clinical reasoning process. Firstly, one might assess whether important information was collected and retained by the clinician. Secondly, one might assess diagnosis and patient management outcomes resulting from the process of integrating new clinical information with pre-existing knowledge structures. The third, largely ignored, measurable aspect of the clinical reasoning process entails assessing the development of an examinee's pre-existing knowledge organisation as revealed by how efficiently the examinee is able to integrate new patient information within that existing structure. Figure 1 provides a graphic representation of the primary inferences related to both clinical problem-solving PBAs and alternative assessment methodologies that rely on measures reflecting an examinee's level of knowledge organisation and the extent to which it allows the useful integration of new case-related data. This representation of the elements required for successful clinical problem solving conveys the dependence between the elements of the clinical reasoning process and highlights opportunities for their measurement. When establishing score validity, such models are crucial because a valid score need not directly assess the target construct.[21] This is especially true when more direct performance-based measures of the target construct are unreliable and when less direct approaches are capable of producing a more reliable score that is closely related (highly correlated) to the trait or ability of primary interest.[22] The central section of Fig. 1 (between the dotted lines) diagrammatically displays the relationships between knowledge organisation and information integration, clinical reasoning and problem solving. The arrow at the top of the figure (above the dotted line) conveys the validity-related inferences relevant to validating current problem-solving PBAs. The bottom arrow (below the dotted line) conveys a validity inference that may facilitate the design of more efficient assessments that employ new and innovative formats for the assessment of clinical reasoning. As measurement instruments designed to

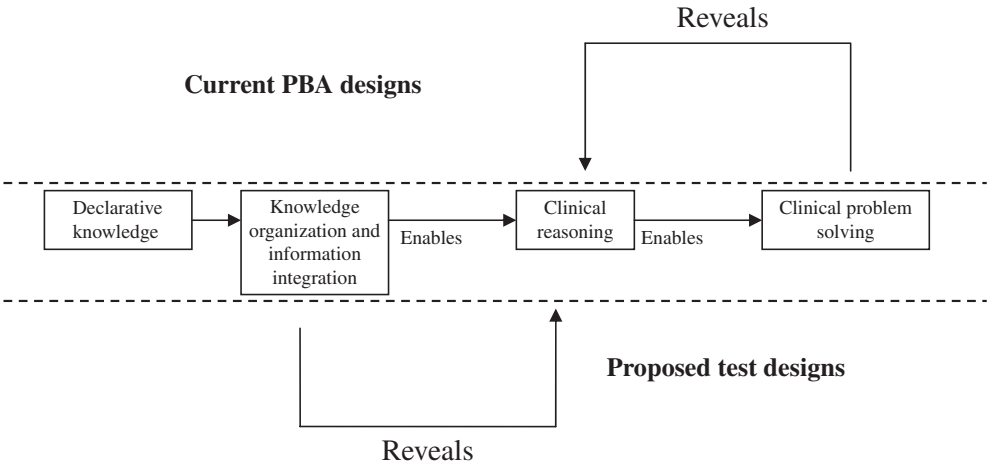

Figure 1 Important relationships for assessing clinical reasoning

characterise clinical knowledge structures are unlikely to yield objective or reliable scores, it may be useful for researchers to examine whether other testing formats are capable of reflecting the general level of an examinee's knowledge organisation and their ability to integrate new information. Already existing item formats, such as the script concordance test, appear to provide the opportunity to reliably, but indirectly, assess whether an examinee generally understands important relationships between the components of a complex clinical problem.[23] Performance on clinically-related items that require examinees to provide estimates of change in the likelihood or probability of certain conditions resulting from the integration of new findings related to a complex medical case might reflect clinical reasoning ability. Although formats such as the script concordance item have been shown to be quite reliable, current scoring procedures do not yet yield objective scores appropriate for educational assessment.[24] There are, however, other promising formats utilising computerised case presentations that allow the sequential assessment of an examinee's proficiency at similar tasks requiring the integration of clinical information to generate and modify diagnostic judgements and treatment decisions. Response formats that capture an examinee's intuitions regarding underlying case probabilities and base rates are also likely to display a positive statistical relationship with clinical reasoning ability and provide an appropriate method for aptitude and achievement testing within medical education. The probabilistic nature of these new tests provides a strategic methodology against which reasoning precursors might be measured. Bayesian-based reasoning models are proving increasingly important in a number of cognitive research areas,[25] and formats that are sensitive to an examinee's understanding of the probabilistic relationships displayed in complex clinical problems are also likely to be predictive of clinical reasoning ability and clinical decision performance. Clearly, additional research and development of new clinical reasoning assessment formats is needed. A valid and reliable measure of clinical reasoning, informed by the cognitive sciences and modern validity theory, would significantly advance the goals of health science education, and should become a research priority.

Contributors: CDK conceived and developed the main ideas described in this paper and wrote the manuscript. GB contributed a number of ideas related to clinical reasoning and helped to edit, critically review and write the manuscript. Both authors had final approval of the manuscript.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Newble D, Norman G, van der Vleuten C. Assessing clinical reasoning. In: Higgs J, Jones M, eds. *Clinical Reasoning in the Health Professions*, 2nd edn. Oxford, UK: Butterworth-Heinemann 2000;156–65.
2 McGuire CH, Babbott D. Simulation technique in the measurement of problem solving. *J Educ Meas* 1967;**4**:1–10.
3 Melnick DE, Clauser BE. Computer-based testing for professional licensing and certification of health

professionals. In: Bertram D, Hambleton RK, eds. *Computer-based Testing and the Internet*. Hoboken, New Jersey: John Wiley & Sons. 2006;164–85.
4 Kane MT. Current concerns in validity theory. *J Educ Meas* 2001;**38** (4):319–42.
5 Swanson DB, Clauser BE, Case SM. Clinical skill assessment with standardised patients in high-stakes tests: a framework for thinking about score precision, equating and security. *Adv Health Sci Educ Theory Pract* 1999;**4**:67–106.
6 Wainer H, Mee J. On assessing the quality of physicians' clinical judgement. *Eval Health Prof* 2004;**27** (4):369–82.
7 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem-solving: an Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press 1978;292–4.
8 Kreiter CD, Bergus GR. Case specificity: empirical phenomenon or measurement artifact? *Teach Learn Med* 2007;**19** (4):378–81.
9 Margolis MJ, Clauser BE, Harik P. Scoring the computer-based case simulation component of USMLE 3: a comparison of preoperational and operational data. *Acad Med* 2004;**10** (Suppl):62–4.
10 Clauser BE, Margolis MJ, Swanson DB. An examination of the contribution of computer-based simulations to the USMLE Step 3 examination. *Acad Med* 2002;**10** (Suppl):80–2.
11 Clauser BE, Subhiyah RG, Piemme TE, Greenberg L, Clyman SG, Ripkey D, Nungester RJ. Using clinician ratings to model score weights for a computer-based clinical simulation examination. *Acad Med* 1993;**10** (Suppl):64–6.
12 Clauser BE, Swanson DB, Clyman SG. A comparison of the generalisability of scores produced by expert raters and automated scoring systems. *Appl Meas Educ* 1999;**12**:281–99.
13 Clyman SG, Melnick DE, Clauser BE. Computer-based case simulations from medicine: assessing skills in patient management. In: Tekian A, McGuire CH, McGaghie WC, eds. *Innovative Simulations for Assessing Professional Competence*. Chicago, IL: University of Illinois, Department of Medical Education 1999;29–41.
14 Newble DI, Hoare J, Baxter A. Patient management problems: issues of validity. *Med Educ* 1982;**16**:137–42.
15 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assess Eval High Educ* 1987;**12**:220–46.
16 McGaghie WC, McCrimmon DR, Mitchell G, Thompson JA, Ravitch MM. Quantitative concept mapping in pulmonary physiology: comparison of student and faculty knowledge structures. *Adv Physiol Educ* 2000;**1**:72–81.
17 Edmonson KM. Concept maps and the development of cases for problem-based learning. *Acad Med* 1994;**69**:108–10.
18 Mahler S, Hoz R, Fischl D, Tov-ly E, Lernau OZ. Didactic use of concept mapping in higher education: applications in medical education. *Instr Sci* 1991;**20**:25–47.
19 McGaghie WC, Boerger RL, McCrimmon DR, Ravitch MM. Learning pulmonary physiology: comparison of students and faculty knowledge structures. *Acad Med* 1996;**71** (Suppl):13–4.
20 Crooks TJ, Kane MT, Cohen AS. Threats to the valid use of assessments. *Assess Educ Princ Pol Pract* 1996;**3** (3):265–86.
21 Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;**50** (9):741–9.
22 Wainer H, Thissen D. Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Appl Meas Educ* 1993;**6**:103–18.
23 Charlin B. Standardised assessment of reasoning in the contexts of uncertainty. *Eval Health Prof* 2004;**27** (3):305–19.
24 Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. *Acad Med* 2005;**80** (4):395–9.
25 Oaksford M, Chater N. *Bayesian Rationality: the Probabilistic Approach to Human Reasoning*. Oxford, UK: Oxford University Press 2007;3–60.

Received 30 June 2008; editorial comments to authors 25 September 2008; accepted for publication 20 October 2008