Supplemental Digital Appendix 4

Summary Descriptions (Including Common Stimuli, Response Formats, Scoring, Typical Uses, Validity Issues, Feasibility, Advantages, and Disadvantages) of Clinical Reasoning Assessment Methods From a 2016 Scoping Review

Chart Stimulated Recall

Valerie J. Lang, MD, MHPE

Chart stimulated recall (CSR) involves three components: (1) review of the note from an actual patient encounter; (2) an oral exam in which the evaluator probes the learner's underlying thought processes; and (3) feedback that may include action plans to improve future clinical decision making.[1,2] Typically, CSR is used for graduate medical education or assessment of practicing physicians, but it could be utilized with any level of learner.[3] Stimulus: Documentation from an actual patient encounter. Response format: Structured discussion between learner and evaluator, with or without a written form to guide the chart review and oral examination. Scoring: Rating scales, checklists with comment boxes, or none. Typical Use: Formative feedback, often on clinic notes, but can be used in inpatient and other settings.

Validity Considerations

Content: CSR is based on authentic clinical encounters. Breadth and difficulty of patient problems may be limited due to patient availability and selection of charts by the learner or the evaluator. Response Process: Optimal training of raters is unknown. Raters' own clinical reasoning approach and biases which affect any oral examination (e.g. gender, racial) may impact scores. Internal Structure: The number of encounters to achieve reliability for high stakes decisions is unknown. Interrater reliability for chart review alone by investigators with 6 months of training is moderate. Relationship to Other Variables: Oral exam provides additional information not found on chart review alone. Consequences / Outcomes: May encourage learners to focus on their diagnostic performance and documentation.

Feasibility

CSR may be implemented asynchronously, e.g. after a clinic or night float session ends. The main resource cost is time for the evaluator to review the chart, conduct the oral exam, and provide feedback (approximately 25-35 minutes per encounter), and for rater training. Construct underrepresentation (incorrect interpretation of test results due to inadequate sampling) is a risk.[4] To overcome this, an adequate number of cases must be assessed which requires significant time, limiting feasibility.

Advantages / Disadvantages

CSR uses authentic cases in which the learner is directly involved, and the feedback provided may encourage learners to improve their diagnostic reasoning and documentation. However, the subject matter is highly idiosyncratic. Since learners or evaluators must select the cases, the difficulty may be inappropriately high or low. Because CSR is conducted asynchronously, it may identify issues that may need urgent follow-up that should have been addressed earlier.

References:

1. Goulet F, Jacques A, Gagnon R, Racette P, Sieber W. Assessment of family physicians' performance using patient charts: Interrater reliability and concordance with chart-stimulated recall interview. Eval Health Prof. 2007;30:376-392.
2. Schipper S, Ross S. Structured teaching and assessment A new chart-stimulated recall worksheet for family medicine residents. Can Fam Physician. 2010;56(9):958-9.
3. Philibert I. Using chart review and chart stimulated recall for resident assessment. J Grad Med Educ. 2018. Feb:95-96.
4. Reddy S, Endo J, Gupta S, Tekian A, Park YS. A case for caution: Chart-stimulated recall. J Grad Med Educ. 2015;7:531-535.

Clinical or Comprehensive Integrative Puzzle

Steven J. Durning, MD, PhD

The Comprehensive Integrated Puzzle (CIP) takes the form of a "grid" that a learner completes that places a number of findings as the columns and a number of related diagnoses on the rows.[1] The learner is asked to compare / contrast findings within a column (selecting the best "match" for the finding) as well as across all rows (e.g. building a basic illness script for each diagnosis). The assessment addresses both knowledge and how it is organized. Stimulus: The grid can be highly flexible, comprised of written answers, digital images, videos, etc. Response Format: The learner records responses (typically matching five response options) for each domain in medicine tested. Responses can be used once, more than once or not at all within a content area. Scoring: Diseases (rows) as well as findings (columns) are scored. This dual scoring stresses the integrative elements of clinical reasoning, while retaining the ability to discern proficiency in different disciplines.[1] Typical Use: Few studies have explored the CIP and they have primarily involved medical students for low stakes testing.

CIP Example Cardiovascular

Diagnosis	History	Physical	Xray	ECG	Labs	Treatment	Pathophysiology	Pathology
Myocardial Infarction								
Pericarditis								
Hypertrophic Cardiomyopathy								
Infective Endocarditis								
Aortic Dissection								

Validity Considerations

Content: A team of writers with content expertise typically constructs the CIP. The representativeness of the test blueprint to the achievement domain is established by how learners differentiate a group of related diagnoses, integrate their fund of knowledge, and link basic and clinical science content. Response Process: Consideration should be given to orienting examinees to the unique testing format. Scoring is often manual, and passing scores are set by the CIP development team, which may include learners. Internal Structure: Data on reliability are limited, but one study demonstrated high Cronbach alphas.[2] A typical CIP has 20-30 question (e.g. 20-30 total blocks to complete on the CIP "grid"). Odd-even reliability per individual CIP item can vary a moderate amount.[3] Relationship to Other Variables: One study correlated CIP performance with future NBME shelf exam performance.[3] Consequences/ Outcomes: Pass / Fail consequences may be based on conventional scoring (e.g. the average of all correct responses) or criterion referenced.

Feasibility

The CIP is an efficient assessment method. Less time is required to complete a block on CIP "grid" than a standard MCQ. The time to develop a CIP grid, based on number of items, appears far less than the amount of time to construct high quality MCQs (5-10 min vs 1 hour).

Advantages / Disadvantages

The CIP allows exploration of learners' linking of basic and clinical sciences through a series of related questions on a diagnosis (vertical columns). The CIP explicitly forces learners to compare and contrast a group of related diagnoses on multiple domains. Limited data suggest that learners enjoy the format.[3] The main disadvantages of this technique include the limited number of psychometric studies. Resources needed for developing a CIP grid online are unknown (existing CIP grids have used a paper and pencil format).

References:

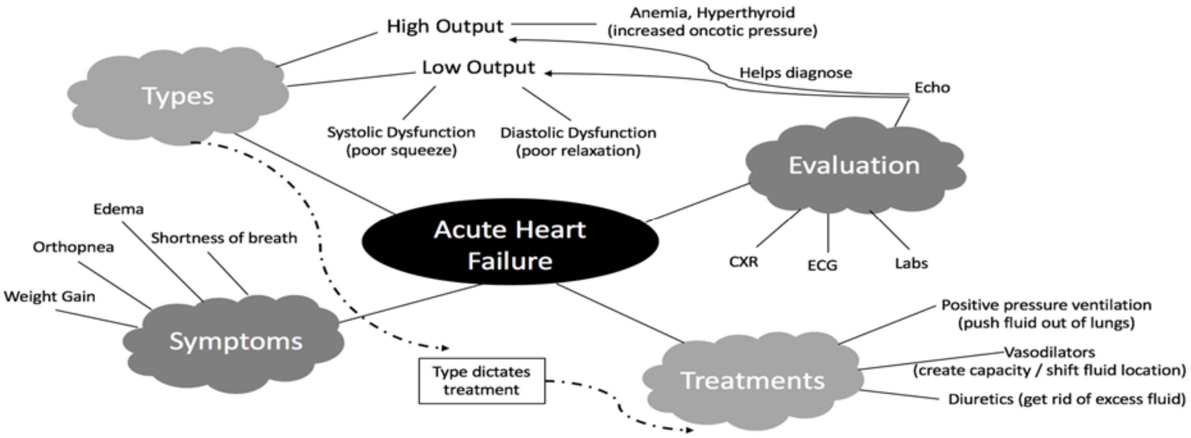
1. Ber R. The CIP (comprehensive integrative puzzle) assessment method. Med Teach. 2003;25(2):171-6
2. Groothoff J, Frenkel J, Tytgat G, Vreede W, Bosman D, ten Cate O. Growth of analytical thinking skills over time as measured with the MATCH test. Med Educ 2008;42:1037-43.
3. Capaldi V, Durning S, Pangaro L, Ber R. The Clinical Integrative Puzzle (CIP) for teaching and assessing clinical reasoning: Preliminary feasibility, reliability and validity evidence. Mil Med. 2015;180(suppl):54-60.

Concept Maps

Ana Da Silva, PhD

Concept maps are a method used to represent relationships between concepts in a visual diagram that has been extensively used in learning, teaching and assessment of factual and procedural knowledge.[1] Its application to the assessment of clinical reasoning is yet to be well established, but some argue that the method holds good potential.[2] Instructions given can be more or less directive depending on the aims of the map, but clear instructions and well-chosen/designed stimuli are essential elements of concept maps.[2] Stimulus: Learners typically create concept maps after being presented with written case scenarios. Past clinical encounters may also be used.[3] Response format: Learners are asked to represent their thought process by identifying both concepts and the relationships involved. Contrary to other assessment methods these maps do not have a standardized answer format. The map types can vary from more 'free-form' to very structured hierarchical representations depending on their aim, purpose, stimulus used, and the underpinning theoretical framework adopted. The maps may be administered via pen and paper, computer platforms or using 'professional writers' to document the thought processes while learners think aloud about the problems. Scoring: Maps are score based on previously defined criteria that can vary from a count of key concepts and valid links[1] to more sophisticated measures such as aggregated scores of reasoning richness and exhaustiveness.[2] Typical use: Concept maps are most often used formatively to support learning or teaching. To date, insufficient studies have evaluated maps as an assessment tool, thus they are not yet suitable for high or medium stakes exams.[1]

Concept Map Example Acute Heart Failure



Validity Considerations

Content: Potential high content validity if the stimuli are the learners' clinical experiences. When using written cases, expert consensus is the most prevalent method. Response Process: Inter-rater (comparison of the scores of two raters/writers on the same map)[2] and intra-rater reliability (comparison rescoring of two random maps by the same rater),[3] is good, but caution is needed to ensure alignment between scoring methods, instructions and the purpose of the map (e.g. teaching, formative assessment, learning from clinical experiences). Internal Structure: Concept maps do not have a standard internal structure. Maps are often scored using a qualitative rubric based on the adopted theoretical frameworks. Cronbach alphas are not frequently reported. Context/case specificity is likely to play an important role as the number of concept maps per individual are often very small. Relationship to Other Variables: Convergence with experience/year of studies is reported by some studies.[2,3]No study reported on convergence with performance across other more standardized examinations such as MCQs, or OSCEs or national examinations. Consequences / Outcomes: May affect learning /cognitive networks.[1] Not used for P/F decisions.

Feasibility

Administration and scoring are very time consuming limiting the number of cases/maps that can be used. This limits the ability to achieve the acceptable psychometric standardization required for higher stakes exams.

Advantages / Disadvantages

Concept maps allow the schematic representation of relationships between concepts. As such, maps may be used to correct possible reasoning errors that may otherwise not be identified. The fast and unconscious nature of pattern recognition would not be captured using this method. Factual and procedural knowledge may be elicited, opening the possibility for gaps or errors in knowledge to be corrected. There is good potential for maps to be used as formative assessments aimed at providing feedback.[2] Any application to high/medium stakes exams would be discouraged until more standardization of this tool as an assessment/evaluation method is achieved.

References:

1. Daley BJ, Torre DM. Concept maps in medical education: an analytical literature review. Medical education. 2010;44(5):440-8.
2. Pottier P, Hardouin JB, Hodges BD, Pistorius MA, Connault J, Durant C, et al. Exploring how students think: A new method combining think-aloud and concept mapping protocols. Med Educ. 2010;44(9):926–35.
3. Vink S, van Tartwijk J, Verloop N, Gosselink M, Driessen E, Bolk J. The articulation of integration of clinical and basic sciences in concept maps: differences between experienced and resident groups. Advances in Health Sciences Education. 2016;21(3):643-57.

Direct Observation

Tiffany Ballard, MD

Direct observation describes the presence of a passive observer (typically faculty) in authentic clinical contexts and is a fundamental method of assessment that can be captured by a variety of assessment tools,[1] such as the mini-clinical evaluation exercise (Mini-CEX).[2,3] It is important to note, that most direct observation tools are not explicitly directed at clinical reasoning, but they often include clinical reasoning as a component. The majority of direct observations are performed by the faculty within a specific discipline (e.g., internal medicine, pediatrics). Stimulus: In almost all situations, faculty observe an actual clinical encounter between an actual patient and learner. These encounters can occur in a wide range of settings, from outpatient clinics to operating theatres. Response Format: A learner's performance with a patient is evaluated. Scoring: A wide variety of scoring mechanisms are associated with direct observation tools, including global ratings with various scale types (norm-referenced, criterion referenced), checklists and even open narrative (e.g. what did the learner do well, etc.). More recently there has been a growing interest in entrustment/ supervision type scales. Typical Use: Direct observations are most commonly used for formative assessment (e.g., 1:1 clinical encounters) during clinical clerkships and residency training.

Validity Considerations

Content: This is provided through the high alignment of direct observation with actual clinical practice and evidence that supports the importance of clinical skills in diagnosis and clinical outcomes. Response Process: Research shows multiple frames of reference used by faculty to judge clinical performance, from self to normative to entrustment. There appears to be some overlap. Criterion-based assessment appears to be infrequent. The effect of faculty training is unclear. Internal Structure: Similar to other methods, it depends on time, sampling and faculty. There is often a wide range of inter-rater reliability.[4] From a generalizability theory perspective, 12-14 Mini-CEX assessments are needed for reliabilities of 0.8 and higher.[2,3] For outliers, 4-5 may suffice using standard errors of measurement (SEM). Construct aligned scales may possess better reliability with fewer observations. Relationship to Other Variables: Direct observations have low to moderate correlations with stage of training. However, data relating the clinical reasoning component to other variables is lacking. Consequences/ Outcomes: Direct observations are often used as a catalyst for remediation and they contribute to Pass / Fail and advancement decisions. Direct observations are valued for their potential to impact learner behavior / performance.

Feasibility

Direct observation is simple to implement. The tools are not difficult to use. The biggest challenge is faculty time.

Advantages / Disadvantages

Direct observations assess the actual care of patients, the top of Miller's pyramid, but direct observations take time, particularly in sufficient quantity to reach reliable conclusions. Faculty variability is a particular challenge and the best approach to prepare faculty raters is unknown. There are also few data on how to best utilize direct observation to assess clinical reasoning, much of which must be inferred by a learner's actions.

References:

1. Kogan J, Holmboe E, Hauer K. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009;302:1316-26.
2. Norcini J, Blank L, Duffy F, Fortna G. The mini-CEX: a method for assessing clinical skills. Ann Int Med. 2003;138(6):476-81.
3. Ansari A, Ali S, Donnon T. The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. Acad Med. 2013;88(3):413-20.
4. De Lima AA, Conde D, Costabel J, Corso J, Van der Vleuten C. A laboratory study on the reliability estimations of the Mini-CEX. Adv Health Sci Educ. 2013;18(1):5-13.

Experimental / Novel Methods

Larry Gruppen, PhD

Investigators continue to do exploratory and applied research in clinical reasoning. Thus, there are a variety of assessment methods that have been developed for research purposes rather than for systematic assessment to make educational decisions.[1-4] Many studies of clinical reasoning use established assessment methods, but novel procedures continue to emerge to study various aspects of the clinical reasoning process. Some of these may eventually become accepted assessment methods, but many are too labor-intensive or narrowly focused to attain widespread application. It is impossible to characterize them as a group because of their diverse purposes, but a common feature is that each is designed to address a specific theoretical question; thus, validity evidence for how well each measures its target construct is essential. Stimulus: Variable stimuli are used, including written cases, images, simulations, etc. Response format: Responses may be open-ended verbal, selected from a set of options, or include unique measures, such as response times. They are often more complex than simple correct answers to questions. Scoring: Some methods may produce a score directly from the responses, but many require analysis and classification of responses, often in relationship to theoretical constructs and predicted relationships. Typical Use: Experimental methods are used primarily for research and innovation.

Validity Considerations

Because experimental methods are not typically used for making educational decisions, explicit validity evidence for the method is rare. However, validity is a core concern to the trustworthiness and utility of the study results; the measures must align with the theoretical constructs or the results are not interpretable. Content: This is an explicit concern for some studies, which address the problem of content/context specificity in clinical reasoning. For others, there is an assumption that the targeted reasoning components are more or less universal. Response Process: Because experimental studies are typically small in size, it may be feasible to have complex response processes that may be well-aligned with the underlying reasoning processes. Internal Structure: This evidence is common in experimental methods. Studies usually focus on an intensive investigation of a small number problems in a specific component of clinical reasoning. Relationship to Other Variables: These relationships are a central focus of experimental studies. The most common relationship is between the target construct and “expertise,” but other external variables are used as reflections of theoretical predictions. Consequences/ Outcomes: Consequential evidence is a negligible concern as these methods are not used for educational decisions (e.g., advancement, grades).

Feasibility

Experimental methods are labor intensive in design, development, implementation, and interpretation. Large-scale application may become a concern if a given method shows promise for measuring a key aspect of clinical reasoning.

Advantages / Disadvantages

Experimental methods are essential tools for augmenting understanding of clinical reasoning and how it relates to other educational phenomena. They are time-consuming, narrowly focused, unproven, often with limited validity or feasibility evidence, and do not have a large implementation base from which to identify best practices.

References:

1) Bordage G, Grant J, Marsden P. Quantitative assessment of diagnostic ability. Med Educ. 1990;24:413-425.

2) Szulewski A, Roth N, Howes D. The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: a new tool for the assessment of expertise. Acad Med. 2015; 90(7):981-7.

3) Durning SJ, Costanzo ME, Artino AR, et al. Neural basis of nonanalytical reasoning expertise during clinical evaluation. Brain Behav. 2015;5(3):e309-19.

4) Durning SJ, Costanzo M, Artino Jr AR, van der Vleuten C, Beckman TJ, Holmboe E, Roy MJ, Schuwirth L. Using functional magnetic resonance imaging to improve how we understand, teach, and assess clinical reasoning. J Cont Educ Health Prof. 2014; 34(1):76-82.

Extended Matching Questions

Eric Holmboe, MD and Brian Heist, MD, MSc

Extended Matching Questions (EMQs) are a written examination format that resemble multiple choice questions (MCQs) in their use of a clinical vignette stem with a single best answer selected from a list of alternatives. EMQS are distinguished from MCQs by significantly longer lists of alternatives applied to multiple questions. The exam is typically organized around themes (e.g. arrhythmias). A list of answer options is provided (e.g. atrial fibrillation, ventricular tachycardia, asystole, etc) followed by a lead-in question (e.g. for each of the following patients, select the most likely arrhythmia.) Stimulus: The stimulus consists of a case vignette, administered either on paper or online. Response Format: A single best response is selected from a list of alternatives, and the same list is used for multiple test items on an examination. Most EMQs contain a minimum of 8 possible answers, though the exact number is dictated by the number of realistic answer options. Scoring: Test items within an exam are generally weighted equally. Typical Use: Most often used for moderate stakes summative assessments in courses or clerkships, however, EMQs have been used on high-stakes national in-service, licensing (e.g. United States Medical Licensing Exams) and certification exams.

Validity Considerations

Content: Validity is established through expert consensus and available evidence-based medicine for each test item. Response Process: Items are produced using a structured process. Item writers typically undergo some training. Quality assurance depends on the stakes of the test. Internal Structure: Some studies report Cronbach alphas of moderate to high acceptability,[1,2] respectively. High internal consistency is achievable, due to the broad range of items that may be tested in a short time frame. Relationship to Other Variables: Included investigations demonstrated construct validity through confirmation of forward reasoning,[3] and concurrent validity through correlation with simultaneous performance on the Diagnostic Thinking Inventory[2] and short-answer questions.[1] There was inconsistency in the correlation between EMQ and MCQ assessments. A study using pre-defined clinical reasoning strategies and categorization of test-takers' think aloud content observed that the number of alternatives does not impact strategy.[4] Other investigators have observed increased difficulty of EMQs over MCQs on test items with the same clinical vignette stem, and hypothesized that the short MCQ alternative list facilitates guessing or hypothetico-deductive review of item alternatives.[5,6] Consequences / Outcomes: These are dependent on purpose of the exam, but can affect pass / fail decisions.

Feasibility

Similar to MCQs, EMQs are straightforward to develop. However, quality vignettes can be challenging to write. The allotted test-taking time for EMQ and MCQ examinations should be similar.[5]

Advantages / Disadvantages

Limited research suggests elimination of the cuing effect that may affect MCQ examination performance.[5,6] The content that can be tested is limited to topics that have a single best answer response. EMQ examinations commonly use the same list of alternatives for multiple test items. The ideal number of alternatives is not established. A long list of alternatives may cause the test-taker to spend excessive time reviewing the list.

References:

1. Brailovsky C, Bordage G, Allen T, Dumont H. Writing vs coding diagnostic impressions in an examination: short-answer vs long-menu responses. Res Med Educ. 1988;27:201-206.
2. Beullens J, Struyf E, Van Damme B. Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. Med Educ. 2006;40(12):1173-1179.
3. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? Med Educ. 2005;39(4):410-417.
4. Coderre S, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Med Educ. 2004;4:23.
5. Case S, Swanson D, Ripkey D. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Acad Med. 1994;69(10 Suppl):S1-3.
6. Heemskerk L, Norman G, Chou S, Mintz M, Mandin H, McLaughlin K. The effect of question format and task difficulty on reasoning strategies and diagnostic performance in Internal Medicine residents. Adv Health Sci Educ Theory Pract. 2008;13(4):453-462.

Global Assessment

Eric Holmboe, MD

Global assessments of clinical reasoning are often included as part of a variety of assessment methods. For example, a global rating of "clinical judgment" is included as part of the mini-clinical evaluation exercise (miniCEX) or as part of an objective structured clinical examination (OSCE) (see direct observation). Global ratings are also commonly included in various forms of faculty evaluation forms completed after various types of curricular activities (i.e. rotations; daily shifts, etc.). Stimulus: In many of these assessment approaches, faculty observe a clinical encounter between an actual or standardized patient(s) and learner. Other stimuli include case presentations as part of clinical rounds and conferences. Response Format: In the case of faculty evaluation forms, the global rating is intended to capture an integrated, composite assessment of a learner performance over some period of time (i.e. longitudinal). The period of time can be as short as a work shift or over weeks to months, as part of clinical rotations, group review of professional development or as a summary judgment during various stages of training. Scoring: A wide variety of scoring formats have been used for global ratings with a variety of scale anchors. Recently, there has been a growing interest in entrustment/supervision type scales. Narrative free text is also common (e.g. what did the learner do well, etc.). Typical Use: Global assessments are utilized in a wide range of settings, from outpatient clinics to operating theatres. They are most often formative but may also be used as summative assessments (e.g. end-of-clerkship evaluations).

Validity Considerations

Content: Validity is typically established through alignment with clinical practice and comparison with an "expert". Response Process: Faculty utilize various frames of reference to judge clinical reasoning performance, ranging from comparisons to self and peers to competency-based or entrustment scales.[1] The use of multiple independent raters can help overcome rater variability. [2] Internal Structure: The length of contact time, case sampling (due to issues of content and context specificity) and faculty characteristics (e.g. seniority) can influence reliability. Inter-rater reliability can be problematic. Training may or may not influence judgements. Construct aligned or entrustment scales, may possess better reliability by reducing rater disagreement and increasing assessor discrimination.[3,4] Relationship to Other Variables: Global ratings appear to correlate with stage of training. Program directors' global ratings of clinical judgment (e.g. on the American Board of Internal Medicine's global rating scale) correlate with resident performance on certification exams.[3] Consequences / Outcomes: Global assessments (often aggregated across multiple assessors) have been used summatively to make remediation, pass / fail and advancement decisions. Individual rater assessments are typically used formatively.

Feasibility

Global assessments are relatively easy to implement, but significant training may be needed to achieve inter-rater reliability. Moving from norm-referenced to construct aligned or entrustment scales requires faculty development. Multiple assessments are needed to achieve higher reliability, which can require a significant amount of time / faculty investment.

Advantages / Disadvantages

Global assessments are often applied to real or standardized patient encounters, and thus have the potential to assess the actual care of patients. Faculty variability in the use of tools is a particular challenge and the best approach on how to prepare faculty raters is unknown.

References:

1. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med. 2010;85(10 Suppl):S25-8.
2. Norcini J, Blank L, Duffy F, Fortna G. The mini-CEX: a method for assessing clinical skills. Ann Int Med. 2003;138(6):476-81.
3. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45(6):560-9.
4. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. Acad Med. 2016;91(2):186-90.

Key-Feature Examinations

Valerie Lang, MD

Key-Feature Examinations (KFE) are written tests designed to assess clinical decision-making. Key features (KFs) are defined as critical or essential elements or step(s) necessary to diagnose or resolve a clinical problem.[1] A case description (the stem) containing KF is followed by one or more questions. KFs may be in the form of clinical clues (e.g. thunder-clap, worst headache of life), diagnostic tests or procedures (e.g. stat CT scan) or therapeutic measures (e.g. neurosurgery consultation). KFs are case-specific. In general, there are 2-3 KFs per case, determined through a consensus process with clinical experts. Patient Management Problems (PMPs) were longer, historical pre-cursors to KFEs. PMPs assess every step in the clinical decision-making process, whereas KFQs focus on just the critical decisions. Stimulus: The stimulus is a written case vignette (paper or online). Response format: Multiple formats can be utilized, including short constructed "free text" entries, selections from short or long lists (from 2 to >500 items).[2] Scoring: Each case within an exam and each KF within a case is weighted evenly. Examinees may select multiple correct options, but only the KFs are scored. For most KFEs, no points are lost for incorrect answers. However, scoring may be structured to penalize dangerous actions (if it is a KF for the problem), or over-ordering (if taking a parsimonious approach to diagnosis or treatment is a KF for solving the problem). Typical Use: KFEs are used for moderate to high-stakes summative assessments (e.g. end-of-clerkship; Canadian Qualifying Exam).

KFE Example Subarachnoid Hemorrhage

A 48-year-old male presents to the emergency department with sudden onset, worst headache of his life that started during sexual intercourse. The pain is located behind his right eye. He has a history of migraine headaches but this headache feels different.

Question 1: What are the diagnoses would you consider at this time? List up to 3

1. ________________________________
2. ________________________________
3. ________________________________

Question 2: With respect to your leading diagnosis, what additional elements of the history and physical would you particularly want to elicit? Select up to 7.

1. Quality of headache
2. Nausea / vomiting
3. Vision changes
4. Photo / Phonophobia
5. Numbness / Tingling
6. Focal Weakness
7. Neck pain
8. Fever
9. Nasal Congestion
10. Ringing in ears
11. Alcohol / Smoking
12. Cocaine abuse
13. Character of prior migraines
14. History of meningitis
15. History of hypertension
16. History of diabetes
17. History of cancer
18. Recent infections
19. Recent surgery
20. Recent trauma
21. Neck tenderness
22. Sinus tenderness.
23. Eye exam
24. Neurologic exam

For a patient with sudden onset, worst headache of life, the student should include subarachnoid hemorrhage (SAH) on the differential, should elicit risk factors for SAH, and perform a neurologic exam.

Validity Considerations

Content: Validity is established through expert consensus on the KFs for each problem. The consensus process is critical and comprises a significant proportion of the exam development process. Alignment with clinical practice, level of learner, and blueprinting are also important considerations.[3] Response Process: Scoring is complex and care must be taken to score each case correctly. If constructed "free text" responses are included, these should be scored using a rubric. Scoring can be performed by hand or automated. Internal Structure: Because cases are distilled down to their critical decisions, more problems can be assessed in a limited testing time, with greater reliability than long cases. However, high stakes exams are still typically 4 hours / ≥40 cases long to reach a reliability of 0.8 or higher.[1] There is low to moderate inter-rater reliability for free text responses. Relationship to Other Variables: There are low to moderate correlations with MCQs that assess application of knowledge.[2] KFE scores correlate with performance in practice.[4] Consequences / Outcomes: KFEs are often used to make high stakes decisions in countries other than the U.S. (e.g. Canadian / Australian licensing exams).[5]

Feasibility

KFEs are challenging to develop. Case writers must be trained and consensus with multiple experts is needed to establish the KFs. Once developed, however, KFEs are easy to administer and scoring may be automated.

Advantages / Disadvantages

KFEs assesses clinical decision-making rather than rote knowledge. Two large studies have demonstrated correlation with clinical practice. Development is resource intensive. The format is largely unfamiliar in the U.S. Acceptable reliability requires a large number of questions.

References:

1. Page G, Bordage G. The Medical Council of Canada Key Features Project: a more valid written examination of clinical decision-making. Acad Med. 1995;70:104-110.
2. Hrynchak P, Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 2014;48:870-883.
3. Farmer E, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 2005;39(12):1188-1194.
4. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Influence of physicians’ management and communication ability on patients’ persistence with anti-hypertensive medication. Arch Intern Med. 2010;170(12):1064-72.
5. Hinchy J. Farmer A. Assessing general practice clinical decision-making skills: the key features approach. Australian Family Physician. 2005;34(12):1059-61.

Modified Essay Questions

Michelle Daniel, MD, MHPE

MEQs comprise a methodology wherein serial information is presented about an evolving clinical scenario.[1] MEQs mimic the temporal sequence of decision-making in clinical practice and provide an alternative to real or standardized patient encounters. The method is situated between MCQs and constructed "free text" essays and is well suited to assess a variety of components of the clinical reasoning process (from data collection to diagnostic justification). Stimulus: A clinical case is presented as a chronologic sequence of items in a booklet or on a computer platform. After each item, the learner must document their decision-making before being allowed to preview subsequent items. Students cannot return to previous test items. Response Format: Each item requires a decision, given in either a free text or multiple-choice format. Cumulative error can significantly impact scores unless the correct answers are provided to the preceding question prior to moving on. Scoring: A score of satisfactory (1) or unsatisfactory (0) is provided for each item. Items are added together to create a score for each case, and multiple cases may be combined for a total score. Scoring may be norm referenced (aimed at differentiating between students) or criterion referenced (aimed at demonstrating students have met competencies). Typical Use: MEQs are used for medium-stakes assessments. They are no longer deemed suitable for high-stakes exams.[2]

Validity Considerations

Content: Validity is typically established through expert consensus and alignment with clinical practice. Items can be designed to measure multiple levels of cognitive processing – including comprehension, analysis, synthesis, and evaluation. Care must be taken to avoid developing items that only test recall and knowledge. Response Process: At least one study demonstrated critical inconsistencies among judges in marking.[2] It was unclear if this was due to the hawk-dove effect of different raters versus flaws in the marking template. Internal Structure: Reliability coefficients showed significant case specificity.[1] Cronbach Alphas were widely variable[1-5] with many if not most being unacceptably low. Estimated Alpha 60s (the predicted reliabilities of the assessment if it contained 60 items) were more reasonable[2,5], but this number of items was almost never administered in actual practice. The length of each case precludes administering an exam with enough cases to reach high reliability. Relationship to Other Variables: Convergence was noted with performance on MCQs on local and national examinations (i.e. NBME clinical subject exams, USMLE Step 1 / 2), OSCEs and clinical performance.[2,5] Consequences / Outcomes: MEQs are used to diagnose student weaknesses, to provide a basis for remediation or to make pass / fail decisions for courses and clerkships.

Feasibility

Highly reliable MEQs are generally not feasible to administer (too many items for the examinee to take and for the examiner to score). Computer based interfaces are more attractive to students, but can take 5-6 times the amount of man-hours to administer and score compared to paper based exams.[3]

Advantages / Disadvantages

Proponents cite the ability of well-constructed MEQs to better assess higher order cognitive skills compared to MCQs. MEQs also avoid the potential cueing effect of MCQs. The administration of a psychometrically reliable MEQ exam is largely limited by feasibility. Automated marking is difficult, if not impossible. Inter-rater reliability is poor. While the potential for MEQs to assess higher order cognitive skills is touted, in practice, they tend to focus on lower order skills such as recall.[1,2] MEQs appear to be susceptible to a more significant number of item writing flaws compared to MCQs.[2] Given that MEQs also do not appear to be as reliable as MCQs, these factors limit the assessment's usefulness.

References:

1. Feletti G. Reliability and validity studies on modified essay questions. Acad Med. 1980;55(11):933-41.
2. Palmer E, Duggan P, Devitt P, Russell R. The modified essay question: its exit from the exit examination? Med Teach. 2010;32(7):300-7.
3. Lim E, Seet R, Oh V, et al. Computer-based testing of the modified essay question: the Singapore experience. Med Teach. 2007;29(9-10), e261-68.
4. Palmer E, Devitt P. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? BMC Med Educ. 2007;7(1):49.
5. Reinert A, Berlin A, Swan-Sein A, Nowygrod R, Fingeret A. Validity and reliability of a novel written examination to assess knowledge and clinical decision making skills of medical students on the surgery clerkship. Amer J Surg. 2014;207(2):236-42.

Multiple Choice Questions

Eric Holmboe, MD and Brian Heist, MD, MSc

Multiple Choice Questions (MCQs) are a form of written assessment comprised of clinical vignette stems followed by up to 5 alternatives including the following described question types: A (single best alternative), M (matching), X (true/false), and combinations of alternatives (e.g. "a,b, and c"). From the 1960s to the 1980s, Patient Management Problems (PMPs) and MCQs were the most common high stakes question formats used to assess clinical competence. While PMPs were initially perceived to assess problem solving skills better than MCQs, research in 1975 and more conclusively in 1984 demonstrated high correlation on items testing clinical judgment, synthesis, and problem solving. Further, MCQs were found more reliable with superior testing efficiency, resulting in their displacement of PMPs.[1,2] Stimulus: The stimulus consists of a case vignette, administered either on paper or increasingly by computer or online. Response Format: The most common response format is a single best answer. Scoring: Final test items within an exam are usually scored equally (i.e. full credit for correct response; no credit for a wrong response). This scoring approach depends on the test blueprint and sample being weighted based on item difficulty prior to test form construction. In other words, each form of the exam may not have exactly the same questions, but the overall difficulty of each form is equivalent. Scoring is often automated but can be completed manually. Typical Use: MCQs are one of the most commonly used testing formats in medical education and are the primary format for high stakes examinations (e.g. United States Medical Licensing Examination and specialty board certification examinations).

Validity Considerations

Content: Validity is established through expert consensus and available evidence-based medicine for each test item. Each item is also judged for difficulty (e.g. using the Angoff method). Response Process: Items are produced using a structured process. Item writers should undergo training. Item performance is examined post-test administration. Attention to exam security in high-stakes applications and how the exam is administered are important considerations. Internal Structure: Reliability is high (typically > 0.85) within typical high-stakes examination testing times (6-8 hours). Fewer items and shorter exam times lower reliability, and the exam becomes more susceptible to case / context specificity. A review of the 1980-82 American Board of Internal Medicine examination results showed adequate internal consistency on items assessing clinical judgment and synthesis.[2] Relationship to Other Variables: Performance on PMPs correlates with performance on problem-solving and clinical judgment components of MCQs.[1,2] Multiple studies demonstrate correlations between quality performance measures and MCQs, but few investigations compare test performance with diagnostic error rates in clinical practice.[5] Consequences / Outcomes: Consequences are significant: MCQ exams are often used to make pass/fail judgments in courses and clerkships, as well as decisions concerning licensing, certification and credentialing to practice medicine. The use of MCQs for progress testing carries more intermediate stakes for learners as they have the opportunity over time to improve performance.[6]

Feasibility

MCQs are straightforward to develop with low resource requirements compared to other testing formats using written clinical vignettes, such as Patient Management Problems and Key Feature Examinations. However, MCQs used in high-stakes examinations can be quite costly to develop and maintain.

Advantages / Disadvantages

Psychometric analysis has demonstrated capacity of MCQ examination to test a wide range of knowledge in a short period of time.[2] Limited research suggests that cueing effects may influence examination performance.[3,4]

References:

1. Joorabchi B, Chawhan A. Multiple choice questions. The debate goes on. Br J Med Educ. 1975;9(4):275-280.
2. Norcini J, Swanson D, Grosso L, Shea J, Webster G. A comparison of knowledge, synthesis, and clinical judgment. Multiple-choice questions in the assessment of physician competence. Eval Health Prof. 1984;7(4):485-499.
3. Case S, Swanson D, Ripkey D. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Acad Med. 1994;69(10 Suppl):S1-3.
4. Heemskerk L, Norman G, Chou S, Mintz M, Mandin H, McLaughlin K. The effect of question format and task difficulty on reasoning strategies and diagnostic performance in Internal Medicine residents. Adv Health Sci Educ Theory Pract. 2008;13(4):453-462.
5. Lipner R, Hess B, Phillips R Jr. Specialty board certification in the United States: issues and evidence. J Contin Educ Health Prof. 2013; 33(Suppl 1):S20-35.
6. Heeneman S, Schut S, Donkers J, van der Vleuten C, Muijtiens A. Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. Med Teach. 2017;39(1):44-52.

Objective Structured Clinical Examination

Joseph Rencic, MD

Objective Structured Clinical Examinations (OSCEs) consist of an adaptable assessment approach that includes various methods that typically focus on what learners do rather than what they know.[1] They use a circuit of timed, standardized "stations" that can assess diverse clinical reasoning content areas and tasks.[2] Reproducibility of the examination situation and predefined grading criteria create the potential for more "objective" assessments as compared with workplace-based assessments using real patients.[2] Stimulus: The stimulus is a diagnostic task accomplished by using various resources (e.g., standardized patients (SPs), ECGs, etc.). Response Format: Learners are assessed on task performance and/or constructed written responses on post-encounter forms. Scoring: Faculty and/or SPs rate examinees using predefined scoring rubrics, typically itemized checklists (i.e., yes/no or behaviorally-anchored rating scales) or a holistic/global rating scale on the overall process. Typical Use: OSCEs are used for formative and/or high-stakes summative assessments (e.g., end of rotation, end of year assessments, certification examinations).

Validity Considerations

Content: Content validity is established by expert consensus and alignment with clinical practice. Systematic and representative sampling with multiple cases is necessary to overcome context specificity (i.e., blueprinting). Whole task exercises may increase use of diagnostic reasoning.[3] Response Process: Highly-organized, standardized, reproducible stations are needed. SP training and monitoring of the accuracy of their character portrayals and ratings are essential.[2] Behaviorally-anchored scoring rubrics should be predefined by subject matter experts. Checklists may fail to recognize expert performance.[4] Faculty and/or SP raters should be trained to use rubrics and provided feedback on their accuracy to improve validity. P/F criteria should be determined through formal standard setting (e.g., Angoff method) by expert assessors. Internal Structure: Reliability varies significantly across studies and can be improved through: 1) increasing number of stations and total testing time (e.g.,10-14 stations over 3-4 hours demonstrated high reliability,[5] 2) use of global/holistic rating rather than checklists,[6] 3) use of a faculty rater or content expert,[6] 4) SP training, and 5) use of evidence-based checklist items.[6] Relationship to Other Variables: Low to moderate correlations exist with the American Board of Internal Medicine examination. Low correlations were observed between OSCE and MCQs, program director ratings, ward ratings, PMPs, NBME scores, clerkship grades and self-ratings. Consequences / Outcomes: High-stakes OSCEs are used for remediation, promotion, graduation, and certification decisions.

Feasibility

OSCEs require high allocations of resources for both development and administration but are used broadly in undergraduate medical education, supporting their feasibility.

Advantages/Disadvantages

OSCEs provide a standardized way to assess a diverse range of clinical reasoning skills in a variety of content domains (Kirkpatrick Model level 3 assessment evidence). The opportunity to assess authentic data gathering, a critical aspect of clinical reasoning, in a standardized setting is a distinct advantage over other non-workplace-based assessment methods. The number of stations required to overcome context specificity may be prohibitive. Clinical reasoning performance on OSCEs likely represents a "best-case" scenario rather than a realistic assessment of actual clinical performance as a result of the behavioral changes that observation causes. No meaningful data exist regarding the predictive validity of OSCEs on future clinical reasoning performance.

References:

1. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437-46.
2. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part II: organisation and administration. Med Teach. 2013;35(9):e1447-63.
3. Lafleur A, Cote L, Leppink J. Influences of OSCE design on students' diagnostic reasoning. Med Educ. 2015;49(2):203-14.
4. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74(10):1129-34.
5. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med. 1993;68:Suppl:S4-S6
7. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181-9

8. Daniels VJ, Bordage G, Gierl MJ, Yudkowsky R. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an internal medicine residency OSCE. Adv Health Sci Educ Theor Pract. 2014;19(4):497-506.

Oral Case Presentation

Michelle Daniel, MD, MHPE

Description

The Oral Case Presentation (OCP) is a structured way to deliver information about a patient to another provider. While ubiquitous in clinical practice, little has been written about OCPs as a clinical reasoning assessment method. Medical students often understand the OCP as a way to organize large amounts of data, while experienced clinicians tend to view the OCP as a flexible way of telling a story to make an argument for particular conclusions. Viewed as the latter, evidence of a learner's diagnostic and therapeutic reasoning can be found throughout the OCP as the learner makes deliberate choices about what to include, what not to include, the order in which data are presented and the structure and content of the assessment and plan. The assessment and plan may contain the most robust evidence of a learner's diagnostic and therapeutic reasoning: This may be presented in the form of a summary (or encapsulation) statement with key features and semantic qualifiers, a prioritized differential diagnosis and treatment decisions with justification of each, however, this format is not as standardized for OCPs as it is for written notes. During or after the delivery of an OCP, raters may probe the learner for additional information and check for understanding. Stimulus: OCPs are typically given after real or simulated patient encounters. Response Format: OCPs are constructed (verbal) "free text" responses similar to written notes. Scoring: OCP scoring varies from informal global assessments, to itemized rating scales with or without behavioral anchors, to discourse analysis with categorical assignment into 1 of 4 semantic classes.[1] Typical Use: OCPs are common to all medical school clerkships and residencies, a component of both OSCEs and the long case (a simulated or real patient encounter wherein a student performs a history and physical, delivers an OCP and writes a post-encounter note.)

Validity Considerations

Content: Content validity is very high as it aligns with clinical practice. Response Process: Rater training on tools is needed. Global assessments are more common in clinical practice, but susceptible to subjectivity of the rater and factors irrelevant to clinical reasoning, such as presenter style, fluency, confidence, etc. OCPs are also subject to context specificity, and multiple observations over time are likely necessary for acceptable reliability (though this is not explicitly discussed in the limited literature on OCPs.) Internal Structure: For the OCP rating tool (PRT),[2] interrater reliability was quite high for the clinical reasoning components. For the long case,[3] factor analysis showed that only a moderate amount of the total variance in data was explained by a candidate's ability, whereas an almost equal amount of the variance was determined by case-specificity. Reliability of a single long case is extremely poor. To achieve dependability of >0.7, 4 long cases with 2 raters each is required, making the format largely unrealistic. Discourse analysis has generally proven infeasible due to poor consistency over judges despite extensive training.[1] Relationship to Other Variables: This is not well described in the literature. Consequences / Outcomes: OCPs often affect grading on clerkships. Long cases have been used for high stakes national board exams (e.g. New Zealand.), but there are challenges with reliability.[2]

Feasibility

The OCP commonly used in clinical practice. Assessment with a structured tool such as the PRT is very feasible with brief (2hr) training. The long case has fallen out of favor due to the significant number of testing hours required to reach reliability, though some still argue its value due to content validity. Discourse analysis or semantic categorization has proven difficult to implement in practice due to the need for extensive rater training.

Advantages / Disadvantages

The OCP is common in clinical practice and could be more deliberately leveraged to assess clinical reasoning using tools such as the PRT. OCPs are still relatively under-studied, and more data are needed concerning reliable rating tools. Long cases and discourse analysis do not have a clear place in the clinical reasoning assessment armamentarium.

Key References:

1. Bordage G, Connell K, Chang R, Gecht M, Sinacore J. Assessing the semantic content of clinical case presentations: studies of reliability and concurrent validity. Acad Med. 1997;72(10), S37-S39.
2. Wilkinson T, Campbell P, Judd S. Reliability of the long case. Med Educ. 2008;42(9):887-893.
3. Lewin L, Beraho L, Dolan S, Millstein L, Bowman D. Interrater reliability of an oral case presentation rating tool in a pediatric clerkship. Teach Learn Med. 2013;25(1):31-38.

Oral Examination

Joseph Rencic, MD

Oral examinations are typically conducted by one or more faculty through unscripted or semi-scripted questions that assess an examinee's clinical reasoning and decision-making ability, as well as the professional values accounting for these decisions.[1] Stimulus: Examiners ask spontaneous and/or structured, standardized questions based on a paper vignette, standardized patient or real patient encounter. Response Format: Examinees respond verbally to questions. Scoring: Scoring is based on an itemized or global rating scale that is criterion-referenced. Typical Use: Oral examinations are used for formative and/or high-stakes summative assessments (e.g., end of rotation, medical school graduation, and postgraduate medical certification).

Validity Considerations

Content: Validity is dependent on systematic and representative sampling of clinical reasoning (i.e., "blueprinting"). Response Process: Subject matter experts should develop behaviorally-anchored scoring rubric and pass/fail criteria through a formal process (e.g., Angoff method). Faculty examiners significantly impact validity, accounting for a fair amount of variance in assessment; therefore, faculty training and monitoring are essential for valid assessment.[2] Structured examination with standardized questions may improve response process.[3] Construct-irrelevant biases can impact scoring (e.g., clothing).[4] Internal Structure: The reliability of oral examinations is low to moderate in general, but they may achieve high reliability depending on the quality and quantity of examiners, number as well as duration of cases (e.g., ten 30-minute cases, or 5 hours of testing), and number of topics discussed per case.[4] Global judgments are typically more reliable than itemized rating scores.[5] Relationship to other variables: Structured oral examinations in surgery had moderate correlations with MCQs and OSCEs;[4] limited data exists in other fields. Consequences / Outcomes: High-stakes oral examinations are used for remediation, promotion, graduation, and certification decisions.[1] However, experts recommend that they should not be used as the sole basis for these decisions given threats to validity.[2]

Feasibility

Oral examinations require significant resource allocation, primarily faculty time, for development and administration. High-stakes examinations require even greater resource allocation because the number of stations and total testing time increases dramatically.

Advantages / Disadvantages

The advantage of oral examination is that it allows for deep probing of clinical reasoning and decision making, including diagnostic and therapeutic justification. It avoids the cueing effects of multiple choice questions and provides invaluable opportunities for formative feedback and clinical reasoning role modeling. The major disadvantage is the high resource utilization necessary to obtain adequate reliability for high-stakes assessment. Performance anxiety can impact scores. Faculty may emphasize recall over higher level thinking questions. Interrater reliability is only moderate. Justification for grades lacks written evidence if students appeal decisions.

References:

1. Royal College of General Practitioners. MRCGP Exam Regulations. http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview/mrcgp-regulations-reviews-appeals-complaints-and-mitigating-circumstances.aspx. Published August 2016. Accessed March 22, 2018.
2. Memon MA, Joughin, GR, & Memon B. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. Adv Health Sci Educ Theory Pract. 2010;15(2):277-89.
3. Wass V, Wakeford R, Neighbour R, Van der Vleuten C. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ. 2003;37(2):126-31.
4. Anastakis D, Cohen R, Reznick R. The structured oral examination as a method for assessing surgical residents. Amer J Surg. 1991;162(1):67-70.
5. Daelmans H, Scherpbier A, Vleuten C, Donker A. Reliability of clinical oral examinations re-examined. Med Teach. 2001;23(4):422-4.
6. Yaphe J, Street S. How do examiners decide? A qualitative study of the process of decision making in the oral examination component of the MRCGP examination. Med Educ. 2003;37(9):764-71.
7. Roberts C, Sarangi S, Southgate L, Wakeford R, Wass V, Esmail A, May C. Oral examinations-equal opportunities, ethnicity, and fairness in the MRCGP. BMJ. 2000;320:370-74.

Patient Management Problems

Jennifer Stojan, MD, MHPE

Patient Management Problems (PMP) are written tests used to assess clinical decision making. Examinees are asked to elicit pertinent information from a patient's history and physical exam, interpret test information, synthesize data obtained, and come up with a diagnosis and treatment/management plan. PMPs assess every step of the clinical decision-making process. At each stage, examinees are presented with a variety of options that could be appropriate, inappropriate, or even contraindicated. New information is discovered as the case unfolds. Responses are presented as realistically as possible. Stimulus: The stimulus consists of an "opening" that introduces the patient, setting, and chief complaint. This is generally followed by sections focusing on history and physical exam data as well as sections focusing on the ordering of diagnostic procedures and treatment and management plans.[1,2] Response Format: Throughout the case, the examinee is given a list of options to choose from, with the number of options varying depending on the PMP. Examinees can often choose more than one option on the list. Once the examinee picks an option, immediate feedback is given on that selected option, which can sometimes change the course of the case. Scoring: Scoring systems can range from subjective global rating scales to more objective weighted scoring algorithms. Scoring generally is determined by comparing the option(s) chosen by the examinee to that agreed upon by an expert panel. Points are assigned to choices made, with points being given for positive choices selected and negative options avoided. Weights are assigned to choices and indices are calculated, determining a clinical reasoning score.[1,2,3] Typical Use: PMPs can be used for formative or summative assessments as well as for teaching purposes, including independent study exercises.

Validity Considerations

Content: PMPs have the potential for high content validity evidence given the cases mimic real-life scenarios and are developed by expert panels. However, content specificity has been shown to be a limitation. Response Process: Scoring systems are developed by expert consensus amongst doctors experienced in the relevant conditions. Reaching consensus as to the "best path" is difficult and weighting schemes introduce difficulties. Calculated scores on one simulation are not predictive of those on another and score correlations between PMPs are low.[4] Internal Structure: There is variability in the literature as to whether the PMP can differentiate test takers at different levels of training. It has been shown that medical students can outperform more experienced physicians on certain PMPs and certain scoring algorithms. Relationship to other variables: It has been found that MCQ are more reliable and efficient than PMPs. It has also been found using regression analysis of scores on validity measures that MCQs and PMPs have a surprisingly high correlation.[4,5] Consequences/Outcomes: PMPs can be used for P/F decisions or to determine student weaknesses. Once popular amongst medical schools and licensing bodies, this method has fallen out of favor for high stakes assessments given its shortcomings.[4]

Feasibility

Developing PMPs can be resource and time intensive, requiring an expert panel to reach consensus on case development and scoring. Once developed, PMPs are not difficult to administer, especially when computerized However, given the difficulty of case development and the need to create multiple long cases to achieve validity, PMPs have largely been replaced by Key Feature Exams.

Advantages / Disadvantages

PMPs can be used to assess clinical decision making in a way that mimics real life situations. Examinees can elicit information and order tests in a simulated environment, where there is no financial cost to the patient or risk to their safety. Although realistic, the list of options presented has the potential to cue the examinee in a way that does not exist in actual clinical situations. Case and scoring development can be time intensive and difficult, requiring the consensus of an expert panel of physicians.[6]

References:

1.Juul DH, Noe MJ, Nerenberg RL.A factor analytic study of branching patient management problems. Med Educ.1979;13(3):199-205.
2.Palchik NS, Wolf FM, Cassidy JT, Ike RW, Davis WK. Comparing information-gathering strategies of medical students and physicians in diagnosing simulated medical cases. Acad Med. 1990;65(2):107-13.
3.Newble DI, Hoare J, Baxter A. Patient management problems. Issues of validity. Med Educ. 1982 May;16(3):137-42.
4.Van der Vleuten CPM, Newble DT. How can we test clinical reasoning? Lancet 1995; 345:1032–34.
5.Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Med Educ. 1985;19:238–47.
6.McCarthy WH, Gonnella JS. The simulated Patient Management Problem: a technique for evaluating and teaching clinical competence. Br J Med Educ. 1967;1(5):348-52.

Script Concordance Test (SCT)

Stuart Lubarsky, MD, MHPE and Carlos Estrada, MD, MS

The Script Concordance Test (SCT) is used in medical education to assess a specific facet of clinical reasoning competence: the ability to interpret medical information under conditions of uncertainty.[1] The response format reflects the way information is processed in authentic clinical problem-solving situations. In contrast to most conventional written assessment tools, the SCT employs a scoring system that acknowledges an important reality in clinical practice: that even experienced clinicians often interpret data, make judgments, and respond to uncertainty in ways that vary. A typical question contains 3 parts: "If you were thinking…, and then you find…, your hypothesis becomes..."[2,3] Stimulus: SCTs consist of short, ill-defined clinical scenarios, followed by a set of independent questions. Part 1 provides a hypothesis in the form of a diagnostic possibility, investigative option, or therapeutic alternative. Part 2 presents new information, such as a physical finding, a pre-existing condition or test result. Response Format: Part 3 is a scale describing the change in probability of the hypothesis. Scoring: SCT scoring uses an aggregate method which assumes that, for each question, the answer provided by the greatest number of expert panelists (i.e. the modal answer) is the 'gold standard'. The responses of the other panelists reflect alternatives that may still be clinically valuable, but worthy of only partial credit. Typical Use: Currently used for assessment across the spectrum, from pre-clinical to continuing professional development.

SCT sample question:

A 19-year-old woman presents to the emergency department with severe lower abdominal pain.			
If you were thinking…	And then you find…	This hypothesis becomes…	Key
Appendicitis	Her temperature is 36.8	**A** **B** **C** **D** **E** -2 -1 0 +1 +2	-2: Much less likely -1: Less likely 0: Neither more/less likely +1 More likely +2: Much more likely
Choledocholithiasis	Her liver function tests are normal	**A** **B** **C** **D** **E** -2 -1 0 +1 +2	
Ruptured Ectopic Pregnancy	Ultrasound shows no intrauterine pregnancy and adnexal mass	**A** **B** **C** **D** **E** -2 -1 0 +1 +2	

Validity Considerations

Content: Validity is derived from clear guidelines for standardizing creation of authentic, ill-defined scenarios. Response Process: Students may struggle with a lack of familiarity with the testing format. Internal Structure: The SCT design has yielded robust indices of internal consistency across a spectrum of medical domains. SCT reliability studies have generally ignored inter-panel, inter-panelist and test–retest measurement error.[4] SCTs featuring 20-25 cases with 3-4 questions nested in each generally provide reliable test scores.[5] Relationship to Other Variables: Relatively weak correlations exist between SCT scores and scores obtained from fact-based exams (supporting the claim that SCTs measure a construct different from tests probing pure recall of knowledge). Evidence exists that SCT scores early in training predict later scores on tests probing similar constructs, but data are limited. Consequences/ Outcomes: SCT is only weakly supported by outcome data.

Feasibility

SCTs can be developed for administration on paper or online. Script concordance tests containing 60–90 questions (nested in 20–25 cases for optimal reliability) can be completed in about 1 hour.

Advantages / Disadvantages

SCT is grounded in established theoretical models of knowledge organization and clinical reasoning (illness scripts). A significant body of evidence supports the validity and feasibility of SCTs in multiple settings. Practical, evidence-based recommendations exist to guide SCT construction. Optimal methods for selecting panel of experts, scoring, and standard setting for SCT have not been established. SCTs are relatively easy to administer and score. The scoring system, however, is a subject of ongoing critique. SCT scores have been shown to present logical inconsistencies and to reflect construct-irrelevant differences in learners' responses.

References:

1. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. Eval Health Prof 2004;27(3):304-19.
2. Dory V, Gagnon R, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Med Educ. 2012;46(6):552-63.
3. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Medical teacher. 2013;35(3):184-93.
4. Lineberry M, Kreiter C, Bordage G. Threats to the validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47:1175–83.
5. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten C. Script concordance testing: A review of published validity evidence. Med Educ. 2011;45:329-338.

Self-Regulated Learning Microanalysis

Timothy Cleary, PhD and Anthony Artino, PhD

Grounded in social-cognitive theory and research, Self-Regulated Learning Microanalysis (SRL-M) is a structured interview protocol designed to gather in-the-moment, task-level information about participants' conscious SRL-M processes as they approach, perform, and reflect on a specific clinical activity.[1] Although relatively brief and straightforward in procedure, microanalytic questions target a wide range of well-defined regulatory, strategic, and motivation processes, such as planning, goal setting, metacognitive monitoring, attributions, and self-efficacy beliefs.[2] Microanalytic questions are purposefully and strategically administered at different time points of a clinical activity to develop a comprehensive account of participants' regulatory processes and beliefs.[3] Further, microanalytic protocols are customized to particular clinical tasks and contexts, and thus can vary based on the needs and interests of clinician educators or researchers. Stimulus: Various formats have been used, including paper and virtual patient cases. Across all formats, microanalytic protocols are structured to optimize alignment between the phase dimensions of SRL-M (e.g. forethought, performance, and self-reflection) and the temporal dimensions of a clinical task (i.e., before, during, and after). Response Format: Free-responses are emphasized, although yes/no or Likert-type formats have been utilized (see examples below).. Responses may be verbal or written. Oral responses are recorded verbatim by the examiner. Scoring: The response format influences the nature of the scoring procedures. Free-response questions are coded into discrete categories using a coding manual. Frequency counts and weighted scoring systems are used to interpret codes. Closed-ended questions have Likert-type scoring, ranging from 0-10 or 0-100. Typical Use: SRL-M is used as a formative assessment tool to diagnose regulatory problems and also to guide instruction or remediation.[4]

SRL-M sample questions:
Strategic planning: In the next section, you will be asked to take the patient's history. Will you employ a particular process or approach for conducting this patient's history? [forethought phase: open-ended item]

- **If NO, please type "No" and then move on to the next question.**
- **If YES, please explain that process/approach in as much detail as possible.**

Self-evaluative (metacognitive) judgment: How confident are you that your leading hypothesis is correct? [performance phase: Likert-type item]

- **Not at all confident**
- **A little confident**
- **Moderately confident**
- **Quite confident**
- **Extremely confident**

Attribution: What do you think is the primary reason why you did not arrive at the correct diagnosis? [self-reflection phase: open-ended question)

Validity Considerations

Content: Phrasing of microanalytic questions is intentionally linked to definitions of specific constructs emphasized in SRL-M theory. Microanalytic protocols tend to represent a diverse array of regulatory processes. Response Process: Simple, brief, non-leading questions should be used. Attempts to minimize response bias are important. Internal Structure: The use of coding schemes is recommended. Calculating inter-rater reliability of free-response questions (Kappa, inter-rater agreement) is key. Relationship to Other Variables: SRL-M questions can predict learning and performance outcomes, however, convergence with other measures has not been well-studied in medical education. Consequences/Outcomes: Information generated can be used to facilitate participant self-reflection, as well as strategic attempts to remediate students' clinical weaknesses.

Feasibility

SRL-M protocols are relatively brief and easy to administer. Although closed-ended questions are simple to score and interpret, development and analysis of free-response questions can be resource and time intensive. Adequate coding requires developing coding schemes, training of coders, and establishing inter-rater agreement.

Advantages / Disadvantages

SRL microanalytic protocols generate context- and task-specific information about participants' regulatory processes as they learn or engage in a clinical activity.[5] Microanalytic protocols are versatile in nature and thus can be customized to a diverse array of clinical activities. The approach is theoretically grounded and aligned with authentic clinical tasks. Thus, microanalytic assessment data can assist in both theory building and in enhancing clinical performance. Microanalytic protocols have been used to shed light onto strategic thought processes that medical students exhibit and whether such processes shift and change over time. A notable drawback of SRL microanalysis is its resource intensiveness for scoring and interpretation (e.g. time spent scoring qualitative responses, training coders, and interpreting results). Other disadvantages include an emphasis on *conscious* regulatory processes (rather than non-conscious processes) and a reliance on participant self-report.

References

1. Cleary T, Durning S, Artino A. Microanalytic assessment of self- regulated learning during clinical reasoning: Recent developments and next steps. Acad Med. 2016;91:1516-21.

2. Artino A Jr, Cleary T, Dong T, Hemmer P, Durning S. Exploring clinical reasoning in novices: A self-regulated learning microanalytic assessment approach. Med Educ. 2014;48:280-91.

3. Cleary T, Dong T, Artino R Jr. Examining shifts in medical students' microanalytic motivation beliefs and regulatory processes during a diagnostic reasoning task. Adv Health Sci Educ Theory Pract. 2015;20:611-26.

4. Durning S, Cleary T, Sandars J, Hemmer P, Kokotailo P, Artino A. Perspective: Viewing "strugglers" through a different lens: How a self-regulated learning perspective can help medical educators with assessment and remediation. Acad Med. 2011;86:488-95.

5. Cleary TJ, Callan GL, Zimmerman BJ. Assessing self-regulation as a cyclical, context-specific phenomenon: overview and analysis of SRL microanalysis protocol. Educ Res Int. 2012; 1-19.

Short and Long Answer (Essay) Tests

Joseph Rencic, MD

Short answer (≤ 3 sentences) and long answer or essay tests (≤ 5 pages) assess written medical knowledge. For clinical reasoning, they allow for a more nuanced assessment of diagnostic and therapeutic decision making because examinees' rationale for their choices can be elicited. Stimulus: The typical stimulus is a clinical vignette which requires the examinee to make diagnosis- and/or management-related decisions with or without justification. Response Format: Examinees respond with short or long written responses. Scoring: Raters use a criterion-referenced, analytic (i.e., a checklist of key components of essay) or a global scoring scale. Typical Use: Short answer tests can be used for formative and summative assessments (e.g., end of course). They can be used alone or combined with other assessment methods if more in-depth assessment of analytical reasoning is desired. Long essay tests were historically used for moderate to high-stakes exams but have fallen out of favor.

Validity Considerations

Content: Systematic and representative sampling of content domain is essential. Response Process: Scoring rubrics should be predefined by subject matter experts. Global rating by expert clinician faculty does not require training. However, non-physician raters or raters using analytic scale rubrics should be trained to use the rubric. They should also be provided feedback on their accuracy to improve validity. P/F criteria should be determined through formal standard setting methods (e.g., Angoff method) by expert assessors. Standardized test administration procedures are essential. Under-sampling is a major threat to validity for long essay tests.[1] Internal Structure: Adequate reliability for high-stakes short answer examinations depends on the number of questions as well as the type and number of raters (e.g., 20 answers administered over 5.5 hours reviewed by 2 physician reviewers can achieve G coefficients > .8).[2] Global scoring enhances reliability, requires no training of physician raters, and correlates well with analytic scoring.[2] Scoring can be impacted by construct-irrelevant factors (e.g., grammatical errors).[3] Long answer essays are impractical because extremely long testing times are necessary for adequate reliability. Relationship to Other Variables: Short essay tests did not appear to add incremental validity beyond that seen with MCQs.[1] Short answer tests of clinical judgment demonstrated modest correlations with MCQ testing and certifying examinations, but no correlation with program director ratings.[1] There was low correlation with standardized patient ratings.[3] Consequences: Essay tests may be used as a component of P/F decisions in courses and clerkships, but are infrequently used for high-stakes decisions.

Feasibility

Written constructed responses are relatively easy to develop but are resource-intensive to grade. They require a statistician who can perform psychometric analyses.

Advantages/Disadvantages

Written constructed responses can assess analytic and synthetic thinking in domains of diagnostic and therapeutic clinical reasoning, including justifications for problem solving and decision making. They avoid cueing effects seen with multiple-choice questions. As compared with certain other non-workplace based assessment methods, they can reduce physician resource utilization because non-physician rater scores correlate highly with physician rater scores (.87) when a specific scoring rubric is provided.[2] Despite these advantages, evidence suggests no significant improvement in validity over MCQ.[1] These data coupled with the significant additional effort required to grade short answer essay questions make their utility in clinical reasoning assessment questionable. Long answer essay questions are rarely used in clinical reasoning assessment because the length of time to produce one response prevents adequate sampling to allow for adequate reliability and generalizability of the assessment to overall clinical reasoning ability.

References:

1. Day S, Norcini J, Diserens D, et al. The validity of an essay test of clinical judgment. Acad Med. 1990;65(9):S39-40.
2. Norcini J, Diserens D, Day S, et al. The scoring and reproducibility of an essay test of clinical judgment. Acad Med. 1990;65(9):S41-2.
3. de Graaff E, Post G, Drop M. Validation of a new measure of clinical problem-solving. Med Educ 1987;21(3):213-18.

Technology-Enhanced Simulation

David Gordon, MD

Technology-enhanced simulation has been defined as an educational tool or device with which the learner physically interacts to mimic an aspect of clinical care.[1] It broadly encompasses a range of instruments from static plastic models to dynamic virtual reality patients. While heavily used for the assessment of procedural skills, technology-enhanced simulation has also been utilized for the assessment of non-technical skills such as communication, leadership, situational awareness, and decision making. High-fidelity mannequins and computer-based simulators lend themselves to the assessment of clinical reasoning.[2] Stimulus: High fidelity mannequins, computer based clinical vignettes with image or video stimuli, virtual patient encounters (interactive computer cases) and avatar-based virtual environments (where a learner interfaces as an avatar) have all been utilized. Response Format: Item selection and action logs are used (history items, physical exam maneuvers, diagnostic tests, treatment, consultations). Diagnoses are often solicited (written or verbal). Constructed "free text" responses and text chats (in virtual environments) are common. Scoring: Attention is given to the sequencing of item selection/ actions and these are typically scored using a point system. Global rating scales and dichotomous checklists are also used. Timed completion of tasks, thematic analysis of case transcripts, and scoring of diagnoses are alternative methods of scoring. Typical Use: Technology enhanced simulation is most commonly used for formative assessments. Some high-stakes assessments use screen-based patient scenarios.

Validity Considerations

Content: Validity is established through the use of assessment blueprints and real patient encounters to model simulated cases. Expert panels may also be utilized for case development. Response Process: Limited data on the response process has been published.[3] Scoring rubrics should be developed when assessing learner performance live or via video review. Raters should receive training on use of the rubric. Learners should be oriented to the virtual environment prior to assessment. Internal Structure: Methods for assessing reliability depend on the type of simulation and whether it involves human raters or computer scoring. Reliability can be estimated through interrater agreement, interstation correlation, and internal consistency. Relationship to Other Variables: Comparisons to a learner characteristic such as training level is the most commonly reported source of validity evidence.[3] Performance correlates with years of training and faculty observations of real clinical performance. Consequences/ Outcomes: Virtual patient cases are used nationally by U.S. medical schools for formative assessment in clinical clerkships. Computer-based case simulation is a component of USMLE Step 3.

Feasibility

The development of technology-enhanced simulation for the assessment of clinical reasoning can be intensive in relation to expensive equipment, support staff, and special technical expertise (e.g., computer programming). Administration costs - human and financial - may vary. Direct observation of mannequin-based simulation can be labor intensive. In contrast, computer-based simulations offer the potential for automated scoring.

Advantages / Disadvantages

Technology-enhanced simulation affords the opportunity to directly observe learner performance in a mock clinical setting (Kirkpatrick Model level 3) and to create dynamic scenarios that respond to learner input. Simulated environments allow learners to perform uninterrupted while avoiding safety concerns that would arise in real clinical settings. Its scoring advantages include the potential to capture not only the learner's answers but also the sequence of actions taken to manage the scenario. In addition, automated delivery and administration enables the assessment of a high volume of learners. Technology-enhanced simulation can involve high financial costs and intensive human capital for the use of high-fidelity mannequins and the development of virtual patients.

References:

1. Ilgen JS, Sherbino J, Cook DA. Technology-enhanced simulation in emergency medicine: a systematic review and meta-analysis. Acad Emerg Med. 2013;20(2):117-27.
2. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ. 2009;43(4):303-11.
3. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med. 2013;88(6):872-83.

Think Aloud

Steven Durning, MD, PhD and Temple Ratcliffe, MD

Think Aloud (TA) is a technique where participants are given a discrete task (e.g. reason through a videotaped clinical case) and, while completing the task, are asked to voice their thoughts in an unfiltered form.[1,2] This technique attempts to access participants' thoughts or "inner speech" and their mental processes as they work through a specific task. In practice, participants are first trained on the TA technique and then prompted to "think aloud". Learners are not typically guided during the TA process, however, they may be encouraged to verbalize their unfiltered thoughts if they don't speak for a period of time while working through the task. In the context of clinical reasoning, TAs are typically administered while completing a task (simultaneous)[1] or immediately following completion of the task (delayed).[3] Stimulus: Formats are highly versatile (e.g. written cases, videos, real or simulated patients, etc). Response Format: Verbal responses are recorded to capture the associated thought processes brought on by the stimulus. Scoring: Recordings are transcribed and then analyzed, typically using qualitative methods. Quantitative scoring is less common. Typical Use: TA is most commonly used in medical education for research purposes or low stakes assessments.

Validity Considerations

Content: Cases (or tasks) are created by drawing on clinical expertise and theoretical predictions of relevant reasoning dimensions. Response Process: TA protocols require training of participants regarding how to think aloud prior to actual task completion. Observers must also be trained not to provide leading prompts and only to encourage the TA process if the learner is silent for some time. Non-physicians may be better suited than physicians to be observers to avoid introducing bias. Attention must be paid to quality control issues with transcription. Internal Structure: Inter-coder agreement during qualitative analysis is challenging. Relationship to Other Variables: TA transcript analysis can be compared against other measures of clinical reasoning performance (e.g., post-encounter notes, diagnostic accuracy, etc). Consequences/ Outcomes: In research settings, TA can be utilized to understand clinical reasoning performance, such as nonanalytic reasoning or to generate theory. In education settings, TA may be employed as part of a remediation program.

Feasibility

TAs are resource intensive, particularly to administer and score. This can limit use. There is a need to train participants in TA as well as observers to avoid bias. TA takes a significant amount of time for subjects to complete. Transcription of sessions and scoring (typically through qualitative analysis) is also resource intense.

Advantages / Disadvantages

TAs can be coupled with simultaneous analysis of behavior, allowing exploration into the process underlying clinical reasoning. This can provide a lens into non-analytic reasoning,[1,4] the shift from analytic to non-analytic reasoning, [1,3] and other clinical reasoning strategies.[5] TAs have also been utilized to build and revise theories and to explore the causes of errors in clinical reasoning. [2,5] There is significant flexibility and versatility in TAs, with an array of possible stimuli. TAs allow close to real time collection of data and can be used in mixed application with other clinical reasoning assessment strategies. As noted above, the main disadvantages of TAs are their resource intensity (e.g. time, training, expertise) and the challenging nature of scoring them (e.g. primarily qualitative analysis). There is also concern about possible biases in verbalizing only conscious reasoning and ignoring non-conscious components,

References:

1. Crowley R, Naus G, Stewart J 3rd, Friedman C. Development of visual diagnostic expertise in pathology: an information-processing study. J Am Med Inform Assoc. 2003;10(1):39-51.
2. Durning S, Artino A, Boulet J, Dorrance K, van der Vleuten C, Schuwirth L. The impact of selected contextual factors on experts' clinical reasoning performance. Adv Health Sci Educ Theory Pract. 2012;17(1):65-79.
3. Sibbald M, de Bruin B. Feasibility of self-reflection as a tool to balance clinical reasoning strategies. Adv Health Sci Educ Theory Pract. 2012;17(3):419-29.
4. Balslev T, Jarodzka H, Holmqvist K, et al. Visual expertise in paediatric neurology. Eur J Paediatr Neurol. 2012;16(2):161-6.
5. Coderre S, Mandin H, Harasym P, Fick G. Diagnostic reasoning strategies and diagnostic success. Med Educ. 2003;37(8):695-703.

Written Notes

Michelle Daniel, MD, MHPE

Written Notes are structured means to communicate information and express clinical reasoning about a patient case to other providers. They are a ubiquitous part of medical student and resident education yet underutilized in the formal assessment of clinical reasoning.[1,2] Post-Encounter Notes are one type of written note used as a component of OSCEs. They may have more specific formats and formal expectations for expressing clinical reasoning, often in the form of a summary statement or encapsulation using semantic qualifiers, a problem list, and differential diagnosis.[3-5] Stimulus: The most common stimuli are real or simulated patient encounters. Response Format: Learners construct written "free text" responses, in the form of clinical documentation or post-encounter notes. Scoring: Global assessments or analytic scoring with checklists or rubrics are used. Typical Use: Written notes are common to all medical school clerkships and residencies. Post encounter notes are part of OSCEs and national clinical skills exams (e.g. USMLE Step 2 CS.)

Validity Considerations

Content: Content validity is high, due to alignment with clinical documentation practice. Checklists and rubrics are typically established by experts. Response Process: Students tend to provide unnecessary details and often lack a concise synthesis using key features and semantic qualifiers, suggesting variability in understanding the goal of written notes. Inter-rater reliability is generally only fair to moderate.[1,2] Categories of disagreement include completeness vs conciseness, relevance, and stringency.[1,4] Internal Structure: Validity evidence for select tools varies. Cronbach's alphas are reasonably high. G coefficients are moderate. Relationship to Other Variables: Studies showed reasonable correlations to other variables, including final clerkship grades,[1] OSCE performance, NBME subject exams, small group grades and course examinations.[3] Consequences / Outcomes: Assessment of written notes often affects evaluations on clerkships and in residency. Post encounter notes have been used to make pass/fail determinations on medium and high stakes exams (e.g. summative OSCEs and USMLE Step 2 CS.)

Feasibility

Training on the various tools is typically brief (1-2 hours) and the actually scoring time is short.[1,3,4] Lay raters can be trained to produce similar results as physicians using analytic rubrics.[5] Exclusive scoring by semantic competence (Bordage) was not feasible, as written notes did not provide robust enough evidence of reasoning.[2]

Advantages / Disadvantages

The organizing frameworks in the tools are well aligned with what naturally occurs in clinical practice. The validity and reliability evidence for select tools is reasonably good. The number of assessments needed to reach generalizability varies by tool. Thus, we recommend caution when post-encounter notes are used in high stakes summative assessments. Of note, much of the work that has been done in this domain is on medicine style notes, and future work is needed in other specialties.

References:

1. Baker EA, Ledford CH, Fogg L, Way DP, Park YS. The IDEA Assessment Tool: assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes. Teach Learn Med. 2015;27(2):163-73.
2. Baker EA, Connell KJ, Bordage GE, Sinacore J. Can diagnostic semantic competence be assessed from the medical record? Acad Med. 1999;74(10):S13-5.
3. Durning SJ, Artino A, Boulet J, et al. The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. Med Teach. 2012;34(1):30-7.
4. Park YS, Lineberry M, Hyderi A, Bordage G, Riddle J, Yudkowsky R. Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. Acad Med. 2013;88(10):1552-7.
5. Berger AJ, Gillespie CC, Tewksbury LR, et al. Assessment of medical student clinical reasoning by "lay" vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. Amer J Surg. 2012;203(1):81-6.