A Case for Caution: Chart-Stimulated Recall

Shalini T. Reddy, MD
Justin Endo, MD
Shanu Gupta, MD
Ara Tekian, PhD, MHPE
Yoon Soo Park, PhD

The Accreditation Council for Graduate Medical Education (ACGME) has called for improved assessment systems that better prepare residents for practice in the 21st century.[1–3] As part of the milestone initiative, graduate medical education (GME) programs must convene clinical competency committees (CCCs) to synthesize assessments collected from evaluators in various clinical settings.[4–6] These changes have prompted the GME community to critically review current assessment methods in order to identify workplace-based assessments that would provide meaningful data for CCC deliberations.[7]

This article reviews theoretical advantages of chart-stimulated recall (CSR), explores threats to validity due to construct underrepresentation and construct irrelevant variance using Messick's framework, and discusses possible solutions. The results can inform the GME community on considerations and potential solutions when implementing CSRs as part of an assessment system. We also identify areas for future research studies.

Chart-Stimulated Recall

CSR is a hybrid assessment format that combines chart review and an oral examination, with both based on a clinician's documented patient encounter. Faculty or the learner selects the clinical chart for a learner's patient to be used as a stimulus for questioning.[8–16] Using the learner's own clinical chart situates the examination within a realistic context, adding to the authenticity and value of the exercise.[17] Through a series of probing questions designed to inquire into the learner's clinical decision-making skills, the examinee is asked to reflect on and explain his or her rationale for clinical decisions. CSR has been used extensively in the United Kingdom and in Canada for the assessment of practicing physicians; in the United States it is predominantly used to assess trainees.

A variety of scoring forms have been developed for CSR, ranging from checklists with comment boxes to ordinal rating scales.[11,18,19] Feedback usually is given to the learner at the end of the encounter[11,20] and may include action plans to improve future clinical decision making.[11,18,20–22] Despite evidence to support the use of CSR in assessing the competence of practicing physicians, its use for certification of physicians has diminished due to practical concerns, such as cost, time, and the need for experienced assessors.[10,11,16,23]

In the context of the new accreditation system and the milestones, CSR provides 2 meaningful contributions to the assessment of residents. First, inquiry focused on the specific case allows assessment of the learner's clinical decision making in a controlled, yet authentic, setting.[24] Second, the formative feedback that a learner receives on a one-to-one basis provides individualized learning opportunities. CSR can fill a gap in the systematic assessment of clinical decision making; thus, a critical analysis of this assessment method is warranted.

Validity Threats Due to Construct Underrepresentation

Construct underrepresentation refers to the incorrect interpretation of test results based on inadequate sampling of that which is being measured.[25] Examples of construct underrepresentation issues as they relate to CSR are outlined in the TABLE. Construct underrepresentation is common in all clinical performance assessments,[26] and may be overcome by increasing the representativeness of cases relative to an assessment blueprint.[27] However, simply administering larger numbers of CSR sessions may not be feasible due to practical limitations. For example, if an average-sized internal medicine residency program has 64 residents who are examined 3 times a year with 20-minute encounters, then administering CSR

DOI: http://dx.doi.org/10.4300/JGME-D-15-00011.1

TABLE
Threats to Validity of Chart-Stimulated Recall (CSR)

Threat to Validity	Problem	Illustration of Problem in CSR
Construct underrepresentation	Inadequate numbers of cases	Time needed to administer and prepare for an assessment limits the feasibility of examining the learner using multiple cases
	Inconsistent case difficulty	Case selection by examinee may result in low or high case difficulty
	Low reliability of ratings	Inconsistency of follow-up prompts based on answers to prior questions Examinees' or examiners' misinterpretation of question(s) Lack of rating instruments with sufficient validity evidence for milestone-based assessment for postgraduate trainees
	Mismatch of sample to domain	Poor chart documentation focuses examiner's attention on data gathering and presentation rather than clinical reasoning
Construct irrelevant variance	Verbal and nonverbal communication	Examinee with limited English language proficiency Examiner questioning style Nervous behaviors in examinee such as fidgeting
	Timing of CSR	Duration of time that has passed since the patient was seen A trainee may underperform on an encounter if it is scheduled when a resident has just completed an overnight shift
	Cognitive errors affecting examination administration and scoring	Inadequate clinical knowledge in the content of the case Content expertise or interest biasing toward or away from a particular diagnosis Examiner using own clinical reasoning as frame of reference for scoring, thus conflating the quality of patient care with the quality of chart documentation
	Examiner's bias toward the learner	Examiner biased by examinee's sex, race, or age

will require approximately 64 hours of personnel time annually, excluding time for preparation and feedback.

Case difficulty may additionally contribute to construct underrepresentation. Selection of straightforward cases that pose minimal challenges to clinical decision making, or complex cases selected for the purposes of receiving corrective instruction, may result in higher or lower ratings.

The format of the examination may also contribute to construct underrepresentation. Although open-ended questions provide evaluators some autonomy to probe examinees, such questions are also subject to interpretation, resulting in potential discrepancies between test administrations.[28,29] In addition, there is considerable variation in available rating instruments and a paucity of recommendations for how to conduct rater training. There are no CSR rating instruments with validity evidence to use for generating scores in a milestone framework. Thus, this limits the ability of CCCs to interpret the results in the context of milestone-based assessments.

Validity Threats Due to Construct Irrelevant Variance

Construct irrelevant variance refers to external factors that contribute to systematic error of a measurement. While some construct irrelevant variance is unavoidable in any workplace-based assessments, CSR appears to be more prone to such errors because of the interactive nature of the examination. In particular, verbal and nonverbal communication may affect assessor scoring.[30] For example, a non–native English-speaking resident may struggle to answer a question rapidly because of language challenges. An evaluator may misinterpret this delay as an indication of weak clinical decision-making

BOX FUTURE DIRECTIONS FOR INVESTIGATION

- Determining the optimal frequency and timing of chart-stimulated recall assessments to ensure adequate inter-rater reliability and to minimize construct underrepresentation
- Developing standardized prompts to minimize construct irrelevant variance due to variations in rater questioning
- Developing a defensible scoring rubric and composite score interpretation
- Studying the impact of evaluator training
- Identifying how cases should be selected
- Improving the feasibility of the examination process
- Measuring the impact of feedback on trainees' performance

skills.[28,31,32] Furthermore, styles of questioning vary between evaluators. Learners may view some styles as overly aggressive, leading to heightened anxiety and nervous behaviors.[33–35] Finally, an evaluator may harbor subconscious biases toward the learner based on age, sex, or ethnicity, among others.

Evaluators' cognitive biases and clinical knowledge may affect both administration of the CSR and examination scoring.[36,37] Evaluators must have high levels of competence and familiarity with the clinical subject matter. Moreover, the evaluator's area of clinical expertise may alter the examination. For example, when faced with a case of dyspnea, a cardiologist may gravitate toward a diagnosis of congestive heart failure and lead the questioning in this direction, while a pulmonologist may gravitate toward emphysema. Therefore, it is possible that raters' markings of the learner are influenced by a bias toward a particular diagnosis.

The reliance on chart documentation creates another potential source of construct irrelevant variance. Poor chart documentation may divert an evaluator's attention away from clinical decision making and toward clinical documentation, effectively converting CSR into an assessment of the resident's documentation skills.

Recommendations

The various threats to validity we have described prevent the use of CSR as a single assessment measure for high-stakes summative assessment decisions. However, CSR can play a useful role as part of multiple sources of assessment for CCC decisions regarding resident performance. CSR's contribution, by facilitating the assessment of learners' clinical decision-making skills and allowing the provision of individualized feedback, is important and may not be captured in other assessment methods. CSRs are interactive, decoupled from the daily time pressures of clinical care, allowing for structured reflection on one's practice. Furthermore, CSR provides a venue for trainees to receive individualized face-to-face instruction, feedback, and assessment from an experienced clinician. In order to fully realize the potential of CSR, it is essential to pay close attention to the development of the instrument, the training of evaluators, and the preparation of examinees.[38]

An important step toward improving the quality of CSR assessment is robust faculty development in 2 areas: (1) *how* to conduct the examination, and (2) how to select the *content* to be examined. To mitigate rater (evaluator) cognitive errors, faculty development should include measures to ensure that raters' clinical knowledge is up-to-date. Recommendations for future research to enhance the quality of CSR are displayed in the BOX.

Summary

CSR is a promising assessment method that provides important feedback to learners and can inform CCC deliberations, yet additional research is needed before it can be used for summative assessments in GME. Clarity and recommendations for mitigating threats to the validity of CSR are still largely lacking—answers to these challenges will help health profession educators determine how CSR should fit into an assessment system in the new accreditation system and its relative benefit to opportunity cost with respect to other assessment methods. Faculty rater development for CSR assessment is an important element of improving the validity and utility of this tool.

References

1. Nasca TJ. ACGME initiatives in concert with Institute of Medicine recommendations. *J Grad Med Educ*. 2014;6(4):809–810.
2. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
3. Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing resident assessment in graduate medical education. *J Grad Med Educ*. 2009;1(2):278–286.
4. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. From the educational bench to the clinical bedside: translating

the Dreyfus developmental model to the learning of clinical skills. *Acad Med*. 2008;83(8):761–767.
5. Dreyfus HL, Dreyfus SE. *Mind Over Machine*. New York, NY: Simon and Schuster; 2000.
6. Carraccio C, Sectish TC. Report of colloquium II: the theory and practice of graduate medical education—how do we know when we have made a "good doctor"? *Pediatrics*. 2009;123(suppl 1):17–21.
7. Schuwirth L. From structured, standardized assessment to unstructured assessment in the workplace. *J Grad Med Educ*. 2014;6(1):165–166.
8. Goulet F, Jacques A, Gagnon R, Bourbeau D, Laberge D, Melanson J, et al. Performance assessment: family physicians in Montreal meet the mark! *Can Fam Physician*. 2002;48:1337–1344.
9. Hayden SR, Dufel S, Shih R. Definitions and competencies for practice-based learning and improvement. *Acad Emerg Med*. 2002;9(11):1242–1248.
10. Miller PA, Nayer M, Eva KW. Psychometric properties of a peer-assessment program to assess continuing competence in physical therapy. *Phys Ther*. 2010;90(7):1026–1038.
11. Schipper S, Ross S. Structured teaching and assessment: a new chart-stimulated recall worksheet for family medicine residents. *Can Fam Physician*. 2010;56(9):958–959, e352–e354.
12. Maatsch JL, Krome RL, Sprafka S, Maclean CB. The Emergency Medicine Specialty Certification Examination (EMSCE). *JACEP*. 1976;5(7):529–534.
13. Maatsch JL. Assessment of clinical competence on the Emergency Medicine Specialty Certification Examination: the validity of examiner ratings of simulated clinical encounters. *Ann Emerg Med*. 1981;10(10):504–507.
14. Maatsch JL, Huang R. An evaluation of the construct validity of four alternative theories of clinical competence. *Res Med Educ*. 1986;25:69–74.
15. Munger BS, Krome RL, Maatsch JC, Podgorny G. The certification examination in emergency medicine: an update. *Ann Emerg Med*. 1982;11(2):91–96.
16. Salvatori P, Baptiste S, Ward M. Development of a tool to measure clinical competence in occupational therapy: a pilot study? *Can J Occup Ther*. 2000;67(1):51–60.
17. Carraccio C, Burke AE. Beyond competencies and milestones: adding meaning through context. *J Grad Med Educ*. 2010;2(3):419–422.
18. Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. *JAMA*. 1993;270(9):1046–1051.
19. Academy of Medical Royal Colleges. Improving Assessment, 2009. http://www.aomrc.org.uk/doc_view/49-improving-assessment. Accessed May 11, 2015.
20. Williamson J, Osborne A. Critical analysis of case based discussions. *BJMP*. 2012;5(2):a514.
21. Mehta F, Brown J, Shaw NJ. Do trainees value feedback in case-based discussion assessments? *Med Teach*. 2013;35(5):e1166–e1172.
22. Jennett P, Affleck L. Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *J Cont Educ Health*. 1998;18(3):163–171.
23. Cunnington JP, Hanna E, Turnhbull J, Kaigas TB, Norman GR. Defensible assessment of the competency of the practicing physician. *Acad Med*. 1997;72(1):9–12.
24. Chaudhry SI, Holmboe E, Beasley BW. The state of evaluation in internal medicine residency. *J Gen Intern Med*. 2008;23(7):1010–1015.
25. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837.
26. Norcini JJ. Current perspectives in assessment: the assessment of performance at work. *Med Educ*. 2005;39(9):880–889.
27. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327–333.
28. 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):277–289.
29. Turnbull J, Danoff D, Norman G. Content specificity and oral certification examinations. *Med Educ*. 1996;30(1):56–59.
30. Burchard KW, Rowland-Morin PA, Coe N, Garb JL. A surgery oral examination: interrater agreement and the influence of rater characteristics. *Acad Med*. 1995;70(11):1044–1046.
31. Roberts C, Sarangi S, Southgate L, Wakeford R, Wass V. Oral examinations-equal opportunities, ethnicity, and fairness in the MRCGP. *BMJ*. 2000;320(7231):370–375.
32. Wakeford R, Southgate L, Wass V. Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. Royal College of General Practitioners. *BMJ*. 1995;311(7010):931–935.
33. Weingarten M, Polliack M, Tabenkin H, Kahan E. Variations among examiners in family medicine residency board oral examinations. *Med Educ*. 2000;34(1):13–17.
34. Thomas CS, Mellsop G, Callender K, Crawshaw J, Ellis PM, Hall A, et al. The oral examination: a study of academic and non-academic factors. *Med Educ*. 1993;27(5):433–439.
35. Rowland-Morin PA, Burchard KW, Garb JL, Coe NP. Influence of effective communication by surgery

students on their oral examination scores. *Acad Med*. 1991;66(3):169–171.

36. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068.
37. Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med*. 2014;89(5):721–727.
38. Raj JM, Thorn PM. A faculty development program to reduce rater error on milestone-based assessments. *J Grad Med Educ*. 2014;6(4):680–685.

Shalini T. Reddy, MD, is Professor of Internal Medicine, University of Chicago Pritzker School of Medicine, and Associate Program Director, Internal Medicine Residency, Mercy Hospital Chicago; **Justin Endo, MD,** is Assistant Professor, Department of Dermatology, University of Wisconsin; **Shanu Gupta, MD,** is Assistant Professor and Director of Education, Rush University Hospitalists; **Ara Tekian, PhD, MHPE,** is Professor and Director of the International Program, Department of Medical Education, University of Illinois at Chicago; and **Yoon Soo Park, PhD,** is Assistant Professor, Department of Medical Education, University of Illinois at Chicago.

Corresponding author: Shanu Gupta, MD and Director of Education, Rush University Hospitalists, 10 Kellogg, 1717 West Congress Parkway, Chicago, IL 60612, 312.942.4200, shanu_gupta@rush.edu