

Daily Encounter Cards—Evaluating the Quality of Documented Assessments

Warren J. Cheung, MD, MEd, FRCPC

Nancy Dudek, MD, MEd, FRCPC

Timothy J. Wood, PhD

Jason R. Frank, MD, MA(Ed), FRCPC

ABSTRACT

Background Concerns over the quality of work-based assessment (WBA) completion has resulted in faculty development and rater training initiatives. Daily encounter cards (DECs) are a common form of WBA used in ambulatory care and shift work settings. A tool is needed to evaluate initiatives aimed at improving the quality of completion of this widely used form of WBA.

Objective The completed clinical evaluation report rating (CCERR) was designed to provide a measure of the quality of documented assessments on in-training evaluation reports. The purpose of this study was to provide validity evidence to support using the CCERR to assess the quality of DEC completion.

Methods Six experts in resident assessment grouped 60 DECs into 3 quality categories (high, average, and poor) based on how informative each DEC was for reporting judgments of the resident's performance. Eight supervisors (blinded to the expert groupings) scored the 10 most representative DECs in each group using the CCERR. Mean scores were compared to determine if the CCERR could discriminate based on DEC quality.

Results Statistically significant differences in CCERR scores were observed between all quality groups ($P < .001$). A generalizability analysis demonstrated the majority of score variation was due to differences in DECs. The reliability with a single rater was 0.95.

Conclusions The CCERR is a reliable and valid tool to evaluate DEC quality. It can serve as an outcome measure for studying interventions targeted at improving the quality of assessments documented on DECs.

Introduction

Work-based assessments (WBAs) have garnered renewed attention as medical training programs have adopted competency-based medical education (CBME) frameworks.^{1–3} Daily encounter cards (DECs) are a form of WBA widely used to evaluate the performance of trainees in settings where they are assigned by schedule (eg, shifts) rather than to a specific preceptor.^{4,5} Unlike in-training evaluation reports (ITERs), which are aggregate end-of-rotation assessments, DECs are completed immediately following a clinical encounter, therefore minimizing recall bias and facilitating assessments based on observation of actual performance.^{5–8}

Unfortunately, DECs, along with other forms of WBAs, do not always reflect the assessor's actual judgment of the trainee's performance.^{9–12} Interventions to improve the quality of completion of WBAs have been designed, including faculty development workshops and rater training initiatives.^{13–15} To evaluate the effectiveness of these interventions, a

tool is needed that can measure outcomes beyond participant satisfaction.^{16,17} This tool should be capable of measuring changes in assessor behavior that correspond to Kirkpatrick's third level of program evaluation.¹⁸

The completed clinical evaluation report rating (CCERR) is a 9-item instrument designed to assess the quality of a completed ITER.⁹ It has repeatedly demonstrated excellent reliability and validity in determining ITER quality and has been used to evaluate interventions designed to improve ITER completion.^{9,13,14} The items on the CCERR appear to be applicable to DECs, which share a similar structure to ITERs (list of items with a rating scale and written comments). However, the CCERR has thus far only been used to assess ITER quality.^{9,13,14,19} Therefore, the purpose of this study is to provide validity evidence to support using the CCERR to assess the quality of DEC completion.

Methods Procedures

This study was conducted at the University of Ottawa in Canada. Six clinical supervisors with expertise in resident assessment from a variety of specialties were purposefully recruited from the University of Ottawa,

DOI: <http://dx.doi.org/10.4300/JGME-D-15-00505.1>

Editor's Note: The online version of this article contains exemplar daily encounter cards in the low- and high-quality groups.

TABLE

Variance Components of Generalizability Study

Facet	VC	% Variance	Explanation
d	1.15	56	The variance attributable to differences in DEC's
r	0.05	3	The variance attributable to differences in raters
i	0.28	14	The variance attributable to differences in items
dr	0.04	2	The variance attributable to the interaction between DEC's and raters (ie, Did raters rate some DEC's differently than other raters?)
di	0.23	11	The variance attributable to the interaction between DEC's and items (ie, Did item means vary by DEC?)
ri	0.07	3	The variance attributable to the interaction between raters and items (ie, Did raters differ in how they rated particular items?)
dri	0.21	11	The variance attributable to the DEC-rater-item interaction plus random error

Abbreviations: VC, variance component; d, DEC; DEC, daily encounter card; r, rater; i, item.

Faculty of Medicine. They were asked to review 60 DEC's representing assessments of residents at various levels of training and performance, completed by supervisors within the Department of Emergency Medicine during the 2012–2013 academic year. They were then asked to determine the quality of these DEC's based on their perception of how informative each DEC was for reporting judgments of the resident's performance, and to group each DEC into 1 of 3 quality categories: high, average, or poor. The experts were subsequently instructed to individually rank the DEC's within each quality category according to which DEC most represented that group. The rankings from all 6 experts were combined, and the top 10 ranked DEC's in each of the 3 quality groups were selected (provided as online supplemental material). This number of DEC's was based on the original validation study.⁹ Assuming a level of significance of $P = .05$, power of 0.80, and a standard deviation of 6.51,⁹ 10 DEC's would be required per quality group to show a significant difference of 8 points on the CCERR (the difference seen between quality groups in the original study).⁹

A second group of 8 clinical supervisors from the University of Ottawa Department of Emergency Medicine with experience using DEC's and blinded to the expert groupings were asked to score the quality of the 30 selected DEC's using a modified version of the CCERR. The only modification was that the term *ITER* was replaced with *DEC* (FIGURE 1).⁹

Research ethics board approval was obtained from the Ottawa Health Science Network.

Analysis

Data analysis was conducted using SPSS Statistics version 23 (IBM Corp, Armonk, NY). The mean total CCERR scores for each of the 3 expert-rated quality groups (high, average, and poor) were calculated. An

analysis of variance and subsequent pairwise comparisons were performed using a 2-tailed level of significance of $P < .05$ to determine to what degree the CCERR could discriminate DEC quality as judged by experts. Reliability for the CCERR scores was determined using a generalizability analysis with raters and items treated as within-subject variables. Individual DEC's were treated as the object of measurement for this analysis.

Results

Mean CCERR scores for the high (37.3, SD = 1.2), average (24.2, SD = 3.3), and poor (14.4, SD = 1.4) quality groups differed ($F_{2,27} = 270$; $P < .001$; FIGURE 2). A subsequent pairwise comparison demonstrated that these differences were statistically significant between the 3 quality groups ($P < .001$), indicating that the CCERR was able to discriminate DEC quality as judged by experts.

The TABLE displays the variance components derived from the generalizability study. The majority (56%) of the variance was attributable to differences in DEC's. Items and DEC by items accounted for 14% and 11% of the variability, respectively, indicating there was some variation in scores across items. Facets involving raters (r, dr, and ri) contributed very little to the variability, indicating that raters scored the DEC's in a similar manner. The reliability of the CCERR with 8 raters was 0.99, and even with a single rater the reliability was 0.95.

Discussion

Statistically significant differences in mean CCERR scores between all 3 quality groups were demonstrated, indicating that the CCERR is able to discriminate DEC quality as judged by experts. Reliability, determined using a generalizability analysis, was high

The purpose of this scale is to evaluate the quality of how a clinical evaluation form, such as a Daily Encounter Card (DEC), has been **filled out**. With this in mind, please use this scale with regard to **how the form has been completed** rather than the design of the form.

"Ratings" refers to the checklists or global ratings used on the evaluation form.

If there is no space for comments on the form, this scale cannot be used. However, if there is a comments section but no comments have been included (ie, the comments section is blank), please indicate a score of "1" for the questions referring to the comments.

Please rate the following by checking the appropriate box.

		1 Not at all	2	3 Acceptable	4	5 Exemplary
1.	Checklist/numeric ratings show sufficient variability to allow identification of relative strengths and weaknesses of the trainee.					
2.	Comments are balanced providing both strengths and areas for improvement.					
3.	The trainee's response to feedback and/or remediation during the rotation is described in the comments.					
4.	Comments justify the ratings provided.					
5.	Clearly explained examples of strengths using specific descriptions (not generalizations) are provided in the comments.					
6.	Clearly explained examples of weaknesses using specific descriptions (not generalizations) are provided in the comments.					
7.	Concrete recommendations for the trainee to attain a higher level of performance are provided.					
8.	Comments are provided in a supportive manner .					
9.	Overall, this DEC provides enough detail for an independent reviewer to clearly understand the trainee's performance on the rotation.					

Adapted from Dudek NL, Marks MB, Wood TJ, et al. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816–822.

FIGURE 1
Modified Completed Clinical Evaluation Report Rating

and raters accounted for very little variability in the scores.

The CCERR has been used as a program evaluation tool to assess the impact of faculty development initiatives aimed at improving the quality of ITERs.^{13,14} Our study suggests that it can be used in a similar manner for DEC. Additionally, because reliable scores can be obtained with a single rater and DEC quality can be evaluated relatively quickly (approximately 15 DECs per hour in this study), a program director could use the CCERR to provide individual feedback to clinical supervisors about the quality of their assessments documented on DECs. The tool may also be used to systematically identify those high-quality DECs that are most informative for making decisions about a trainee's clinical competence.

The primary limitation of this study is that it was conducted at a single center with specialty-specific DECs. A follow-up study providing validity evidence involving raters and DECs from various specialties and training centers, as was done in the original validation of the CCERR,⁹ would increase the generalizability of these results.

Conclusion

The study found evidence for the validity of CCERR scores when applied to DECs, including high reliability

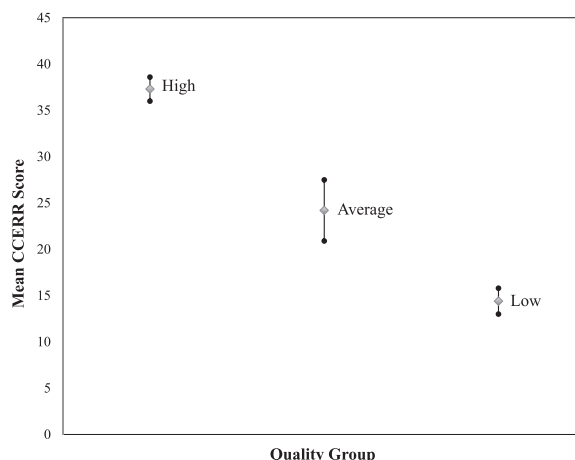


FIGURE 2
Mean CCERR Score $\pm$ Standard Deviation by Expert-Determined Quality Group
Abbreviation: CCERR, completed clinical evaluation report rating.

and discrimination among DECs of varying quality, suggesting that the CCERR can be used as a tool to provide feedback to supervisors to improve documentation of assessments on DECs. The CCERR also offers a quantitative measure of change in assessor behavior when utilized as a program evaluation instrument for determining the quality of completed DECs.

References

- Holmboe ES, Sherbino J, Long DM, et al. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676–682.
- Iobst WF, Sherbino J, Cate OT, et al. Competency-based medical education in postgraduate medical education. *Med Teach*. 2010;32(8):651–656.
- Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach*. 2007;29(9):855–871.
- Sherbino J, Kulasegaram K, Worster A, et al. The reliability of encounter cards to assess the CanMEDS roles. *Adv Health Sci Educ Theory Pract*. 2013;18(5):987–996.
- Sherbino J, Bandiera G, Frank JR. Assessing competence in emergency medicine trainees: an overview of effective methodologies. *CJEM*. 2008;10(4):365–371.
- Al-Jarallah KE, Moussa MA, Shehab D, et al. Use of interaction cards to evaluate clinical performance. *Med Teach*. 2005;27(4):369–374.
- Humphrey-Murto S, Khalidi N, Smith CD, et al. Resident evaluations: the use of daily evaluation forms in rheumatology ambulatory care. *J Rheumatol*. 2009;36(6):1298–1303.

8. Richards ML, Paukert JL, Downing SM, et al. Reliability and usefulness of clinical encounter cards for a third-year surgical clerkship. *J Surg Res.* 2007;140(1):139–148.
9. Dudek NL, Marks MB, Wood TJ, et al. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ.* 2008;42(8):816–822.
10. Speer AJ, Solomon DJ, Ainsworth MA. An innovative evaluation method in an internal medicine clerkship. *Acad Med.* 1996;71(suppl 1):76–78.
11. Cohen G, Blumberg P, Ryan N, et al. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med.* 1993;5(1):10–15.
12. Bandiera G, Lendrum D. Daily encounter cards facilitate competency-based feedback while leniency bias persists. *CJEM.* 2008;10(1):44–50.
13. Dudek NL, Marks MB, Bandiera G, et al. Quality in-training evaluation reports—does feedback drive faculty performance? *Acad Med.* 2013;88(8):1129–1134.
14. Dudek NL, Marks MB, Wood TJ, et al. Quality evaluation reports: can a faculty development program make a difference? *Med Teach.* 2012;34(11):725–731.
15. Albanese MA. Challenges in using rater judgements in medical education. *J Eval Clin Pract.* 2000;6(3):305–319.
16. Steinert Y, Mann K, Centeno A, et al. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. *Med Teach.* 2006;28(6):497–526.
17. Gofton WT, Dudek NL, Wood TJ, et al. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med.* 2012;87(10):1401–1407.
18. Kirkpatrick DL, Kirkpatrick JD. *Evaluating Training Programs: The Four Levels.* 3rd ed. San Francisco, CA: Berrett-Koehler Publishers; 2006.
19. Bismil R, Dudek NL, Wood TJ. In-training evaluations: developing an automated screening tool to measure report quality. *Med Educ.* 2014;48(7):724–732.



All authors are with the University of Ottawa, Ontario, Canada. **Warren J. Cheung, MD, MMed, FRCPC**, is Assistant Professor, Department of Emergency Medicine; **Nancy Dudek, MD, MEd, FRCPC**, is Associate Professor, Division of Physical Medicine & Rehabilitation, Department of Medicine; **Timothy J. Wood, PhD**, is Assistant Professor, Department of Innovation in Medical Education; and **Jason R. Frank, MD, MA(Ed), FRCPC**, is Associate Professor, Vice-Chair of Education, and Director of Education Research and Development, Department of Emergency Medicine, and Director of Specialty Education, Strategy, and Standards, Royal College of Physicians and Surgeons of Canada.

Funding: This study received grant funding from the University of Ottawa, Department of Innovation in Medical Education.

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Warren J. Cheung, MD, MMed, FRCPC, University of Ottawa, Department of Emergency Medicine, F-Main, Room EM-206, 1053 Carling Avenue, Ottawa, Ontario K1Y 4E9 Canada, 613.798.5555, ext 17196, fax 613.761.5488, wcheung@toh.on.ca

Received October 6, 2015; revision received April 8, 2016; accepted April 20, 2016.