# Resident Evaluations: The Use of Daily Evaluation Forms in Rheumatology Ambulatory Care

SUSAN HUMPHREY-MURTO, NADER KHALIDI, C. DOUGLAS SMITH, ELZBIETA KAMINSKA, CLAIRE TOUCHIE, ERIN KEELY, and TIMOTHY J. WOOD

***ABSTRACT.*** ***Objective.*** The in-training evaluation report (ITER) is widely used to assess clinical skills, but has limited validity and reliability. The purpose of our study was to assess the feasibility, validity, reliability, and effect on feedback of using daily evaluation forms to evaluate residents in ambulatory rheumatology clinics.

***Methods.*** An evaluation form was developed based on the Royal College of Physicians and Surgeons of Canada CanMEDS roles. There were 12 evaluation items including overall clinical competence. They were rated on a 5-point scale from unsatisfactory to outstanding. All internal medicine residents rotating on rheumatology were strongly encouraged to provide the form to their preceptor at the end of each clinic. A questionnaire was administered to residents and faculty.

***Results***. Seventy-three internal medicine residents completed a 1-month rotation at University of Ottawa (n = 26) and McMaster University (n = 47). Faculty members completed a total of 637 evaluation forms. The number of evaluation forms ranged from 2 to 16 (mean 8.73) per resident. At an average of 8.73 forms per resident the reliability was 0.71 for the composite score. Fourteen forms would be required for a reliability of 0.8. The correlation between the objective structured clinical examination scores and the forms was 0.48 (p = not significant). Faculty and residents reported increased feedback following implementation of the forms.

***Conclusion***. The use of daily evaluation forms is feasible and provides very good reliability. Use of the evaluation forms increases feedback to residents on their performance. The forms were well received by faculty and residents. (J Rheumatol First Release April 1 2009; doi:10.3899/jrheum.080951)

*Key Indexing Terms:*
EVALUATION POSTGRADUATE EDUCATION

In the present model of medical education, students and postgraduate trainees learn their clinical skills by rotating in various clinical settings. The evaluation of trainee performance during these rotations has been challenging. Although there are several performance-based evaluation methods, the in-training evaluation report (ITER) is frequently used to document student or resident performance in day to day practice. An ITER is typically a rating form that is completed at the end of a rotation. ITER allow ongoing assessment of clinical practice performance, but have been noted to have limited reliability and validity. The ITER are often completed by rotation supervisors who have had very little direct contact with the trainee, and are often completed weeks after the end of a rotation[1]. The retrospective character of the evaluation often leads to a lack of specific examples of students' strengths and weaknesses with inadequate provision of feedback.

Various evaluation tools such as encounter cards, portfolios, and the mini-clinical evaluation exercise (mini-CEX) have been developed to overcome the deficiencies inherent in the ITER. Encounter cards or interaction cards have been used in various inpatient or combined inpatient and ambulatory rotations. This method involves repeated documentation of performance by multiple observers[2]. Several articles have reported on the reliabilities of encounter cards. Hatala and colleagues report a reliability of 0.79 for 7.9 cards collected in an inpatient internal medicine 8-week rotation[3]. Another study from Kuwait found a reliability of 0.91 for 184 cards collected over multiple rotations (internal medicine, surgery, obstetrics and gynecology, and pediatrics) over 12 months[4]. A study of residents in obstetrics and gynecology determined the reliability was 0.73 for 8 encounter cards[5]. In this latter study, all encounters were directly observed, as opposed to only 60% in Hatala's study.

*From the Faculty of Medicine, University of Ottawa, Ottawa; and the Faculty of Medicine, McMaster University, Hamilton, Ontario, Canada.*

*Supported by The Ottawa Hospital Academic Enrichment Fund. Drs. Humphrey-Murto and Khalidi acknowledge the support received as recipients of The Arthritis Society Clinician Teacher Award.*

*S. Humphrey-Murto, MD, Associate Professor, University of Ottawa; N. Khalidi, MD, Associate Professor, McMaster University; C.D. Smith, MD, Associate Professor, University of Ottawa; E. Kaminska, MD, Assistant Clinical Professor, McMaster University; C. Touchie, MD, Associate Professor; E. Keely, MD, Full Professor; T.J. Wood, MD, Adjunct Professor, University of Ottawa.*

*Address reprint requests to Dr. S. Humphrey-Murto, The Ottawa Hospital–Riverside Campus, 1967 Riverside Drive, Box 37, Ottawa, Ontario K1H 7W9, Canada. E-mail: shumphrey@ottawahospital.on.ca*

*Accepted for publication December 9, 2008.*

Other studies have evaluated the effect of encounter cards on feedback. Feedback plays an essential role in medical education. Effective feedback should be immediate, specific, and corrective, and occur regularly. The use of encounter cards has the potential to improve the delivery of feedback. Paukert, *et al* documented that student satisfaction with the feedback process improved significantly with the use of 40 encounter cards during a 12-week surgery rotation[6].

In several ambulatory rotations including rheumatology, trainees interact with multiple faculty members over a 4-week rotation. In many centers this differs from inpatient rotations where trainees may work with only 1 or 2 preceptors. Several teaching programs still rely primarily on the ITER as an evaluation method. To our knowledge, there are no studies to date that review the reliability of evaluation cards in a relatively short, strictly ambulatory setting and evaluate both resident and faculty perception on the effect the forms have on feedback.

The purpose of our study was to assess the reliability and effect on feedback of using daily evaluation forms to evaluate internal medicine residents in ambulatory rheumatology clinics in several Canadian universities. Other outcomes included the feasibility and validity of this evaluation method.

## MATERIALS AND METHODS

Potential collaborators in Canadian university rheumatology programs were approached. Five programs agreed to participate, but 2 dropped out early for logistical reasons and 1 university dropped out because of an interruption in secretarial support. The medical schools at the University of Ottawa and at McMaster University completed the study. Research ethics board approval was received at all centers.

An evaluation form was developed based on the Royal College of Physicians and Surgeons of Canada CanMEDS roles[7]. CanMEDS is a framework of essential physician competencies needed for medical education and practice. They are part of the objectives of training and accreditation standards for postgraduate education in Canada. Based on these competencies, the form was developed by the authors, by consensus.

Figure 1 displays the version of the form that was used. There were 11 evaluation items (history, physical examination, clinical judgment, verbal communication, written records, humanistic qualities, collaborator, organization, scholar, advocate, procedural skills) plus a rating of overall clinical competence. A 5-point rating ranging from unsatisfactory to outstanding was associated with each item.

In addition to the 12 items on the form, faculty were asked to record the percentage of their evaluation that was based on direct observation, case review, or written note review. Also, at the top of the form residents were asked to list the diagnoses as well as any procedure performed during the encounter.

For 6 months prior to the implementation of the evaluation forms, a 6-item questionnaire was administered to internal medicine residents rotating through rheumatology and teaching faculty to determine their perceptions of provision of feedback and direct observation on a 5-point scale. Following implementation, a questionnaire was given to all rotating residents and faculty, which included questions from the pre-implementation questionnaire but also questions assessing the usefulness, perceived fairness, and effect on feedback resulting from implementation of the evaluation forms (Tables 2 and 3).

To assess the validity of the forms, the ratings from the evaluation forms were compared to the scores from an annual objective structured clinical examination (OSCE) that residents must complete. This OSCE is for formative purposes but is mandatory for all internal medicine residents. The OSCE consists of 10 stations that test physical examination skills, communication skill, procedural skills, and ability to manage typical general internal medicine scenarios. Measures of validity also included the perceptions of residents and faculty.

Most internal medicine residents complete a 1-month rotation in rheumatology during their core internal medicine training. During the rotation, residents work with multiple faculty members. All internal medicine residents rotating on rheumatology at both medical schools were strongly encouraged to provide the evaluation form to their preceptor at the end of each clinic. Faculty members were encouraged to complete all categories on the forms and to hand in to the rotation coordinator. Clinical faculty were introduced to the form but not formally trained; however, all had many years of teaching experience with internal medicine residents and were familiar with the CanMEDS roles. Forms were collected over an 18-month period at the 2 universities. Residents continued to receive the end of rotation ITER, as these are a requirement of the respective programs.

The number of forms collected from each resident over the month at both sites was recorded. For each form, a composite score was created by averaging the ratings assigned to the 11 evaluation items, and a generalizability analysis using the composite score and the overall rating was conducted to determine the reliability of the forms and the number of forms per resident required to achieve a reliability of 0.80. An independent t–test was used to compare the 6 ratings to 6 items that were identical on the pre- and post-questionnaires.

## RESULTS

Seventy-three internal medicine residents completed a 1-month rotation in rheumatology at the University of Ottawa (n = 26) and at McMaster University (n = 47). At the University of Ottawa the percentage of first-year residents was 6.5%, second-year 36.5%, and third-year 57%. At McMaster the breakdown was 32.9%, 27.7%, and 39.5%, respectively. Faculty members completed a total of 637 evaluation forms for the 73 residents. The number of evaluation forms per resident ranged from 2 to 16. The mean number of forms collected at the University of Ottawa was 10.46 and at McMaster 7.76, for an overall mean of 8.73 forms per resident.

Table 1 displays the descriptive statistics for all the cards. As shown, not all items were filled out for each form. Most of the forms involved the assessment of the first 6 items on the form, with fewer forms having ratings for scholarly activity/literature reviews or procedural skills. The mean ratings for the items on the form ranged from 3.7 to 4.1, indicating that on average, the supervisors thought the residents' performance was above expectations. That said, there was some variation within the items. As shown in Table 1, the ratings for each item ranged from either 2 to 5 or 3 to 5. More importantly, there was considerable variability in the correlations between ratings on the items. These results indicate that raters were willing to give relatively independent ratings for each of the 12 items.

To determine the reliability of the ratings on the forms, a composite score for each form was created by averaging the ratings on the first 11 items. A generalizability analysis was then conducted. For this analysis, the composite rating on

**Ambulatory Clinic Evaluation Form**

| | | |
|---|---|---|
| **U= unsatisfactory** | **BE= Below Expectations** | **N/A= Not Assessed** |
| **ME=meets expectations** | **AE=Above Expectations** | **O= Outstanding** |

| EVALUATION ITEM | U | BE | ME | AE | O | NA |
|---|---|---|---|---|---|---|
| History | | | | | | |
| Physical exam | | | | | | |
| Clinical Judgement | | | | | | |
| Verbal communication | | | | | | |
| Written records | | | | | | |
| Humanistic Qualities/ Professionalism | | | | | | |
| Collaborator MD/Team | | | | | | |
| Organizational/Efficiency | | | | | | |
| Scholar/ lit.searches | | | | | | |
| Advocate (consider patient's social situation) | | | | | | |
| Procedural Skills | | | | | | |
| Overall Clinical Competence | | | | | | |

*Figure 1.* The ambulatory clinic evaluation form.

each form was nested within resident, with resident treated as a between-subject factor. At an average of 8.73 forms per resident, the generalizability coefficient (g-coefficient) for the forms was 0.71. A decision study showed that it would require an average of 14 forms per resident to achieve a g-coefficient of 0.80. The generalizability analysis was repeated for the single item "overall rating of clinical competence." The g-coefficient for this item was 0.50, and an average of 33 forms per resident would be required to achieve a g-coefficient of 0.80.

Faculty reported that only 10% of the evaluation was based on direct observation, with 80% resulting from case review and the remaining 10% from a review of the written note.

Resident pre-questionnaire responses (n = 27) were compared to resident post-responses (n = 70) for the first 6 items (see Table 2). Resident responses before the institution of the evaluation forms (pre-) and post-responses were not statistically different for any of the items. Table 1 displays the percentage of residents who agreed or strongly agreed with the comments listed in the post-survey. It appears that residents felt the form was a fair evaluation of their skills and should continue to be used. Over 80% of residents felt that they received more feedback and more timely feedback as a result of the form. Only a small percentage of the residents felt the form was intimidating.

Resident comments on the forms also supported the improved feedback. A few illustrative comments include: "timely feedback on the same day was very helpful"; "probably more accurate assessment, timely feedback and faculty forced to consider evaluation immediately"; "liked most to receive feedback at the end of each clinic and to see the progress"; "I truly appreciate feedback being given in an immediate and constructive fashion... allowed me to improve over course of rotation, as opposed to getting a generic ITER 1–2 month later which has no direct relevance"; "the forms provided an avenue for constructive feedback on an ongoing basis so changes could be implemented during the rotation."

Faculty pre- and post-questionnaires were compared for the first 6 items that appeared on both forms (see Table 3). There were no statistically significant differences noted except for the statement, "I provide feedback to residents on their clinical skills on a regular basis," with a score of 3.45

*Table 1.* Scores on individual items of the evaluation form.

| Category | Number of Ratings | Mean Rating | SD | Range of Scores* |
|---|---|---|---|---|
| History | 637 | 3.9 | 0.60 | 2–5 |
| Physical examination | 625 | 3.7 | 0.63 | 2–5 |
| Clinical judgment | 631 | 3.8 | 0.63 | 2–5 |
| Verbal communication | 636 | 3.8 | 0.64 | 2–5 |
| Written records | 588 | 3.8 | 0.66 | 2–5 |
| Humanistic qualities/professionalism | 593 | 3.9 | 0.61 | 3–5 |
| Collaborator MD/team | 347 | 3.8 | 0.67 | 3–5 |
| Organizational/Efficiency | 498 | 3.8 | 0.69 | 2–5 |
| Scholar/lit. searches | 121 | 4.1 | 0.62 | 3–5 |
| Advocate (consider patient's social situation) | 382 | 3.9 | 0.66 | 3–5 |
| Procedural skills | 114 | 3.8 | 0.71 | 3–5 |
| Overall clinical competence | 608 | 3.8 | 0.62 | 2–5 |

* Scores: unsatisfactory = 1, below expectations = 2, meets expectations = 3, above expectations = 4, outstanding = 5.

*Table 2.* Means of pre and post-resident questionnaires and percentage of residents that agreed or strongly agreed to the following statements. No statistically significant difference between the pre- and post-questionnaire responses. Only the first 6 items were included in the pre-questionnaire.

| Statement | Mean Score* (maximum 5) of Pre-implementation, n = 27 | Mean Score* of Post-implementation, n = 70 | Percentage of Residents that Agreed or Strongly Agreed with Statement Pre-implementation, n = 27 | Percentage of Residents that Agreed or Strongly Agreed on Post-implementation, n = 70 |
|---|---|---|---|---|
| I was observed on a regular basis performing a history | 2.59 | 2.47 | 14.8 | 17.2 |
| I was observed on a regular basis performing a physical examination | 3.59 | 3.36 | 55.5 | 50.7 |
| I was observed on a regular basis performing a procedural skill | 3.83 | 3.86 | 73.9 | 71.2 |
| I was given helpful feedback on my clinical skills | 4.15 | 4.24 | 88.9 | 90 |
| I was given timely feedback on my clinical skills | 4.15 | 4.14 | 88.9 | 85.9 |
| The ITER is a fair evaluation of my clinical skills | 3.71 | 3.53 | 66.6 | 53.3 |
| The evaluation form provided a fair evaluation of my clinical skills | | 3.94 | | 79.1 |
| The evaluation form was a better evaluation tool than the ITER | | 3.78 | | 61.8 |
| I received more feedback during my rotation as a result of the evaluation form | | 4.14 | | 84.1 |
| I received more timely feedback as a result of the evaluation form | | 4.22 | | 86.7 |
| The evaluation form was intimidating | | 2.33 | | 11.5 |
| The evaluation form should continue to be used | | 4.07 | | 84.1 |
| Logging patient diagnoses and procedures was useful | | 3.74 | | 73.9 |

* Mean score of rating scale, where 1 = strongly disagree and 5 = strongly agree. ITER: in-training evaluation report.

before the institution of the evaluation forms (pre-) versus 4.27 post ($p = 0.02$). For the statements on the post-form only, the percentage of faculty that agreed or strongly agreed with the statements is shown in Table 3. Sixty-four percent of faculty felt they provided more feedback to the residents as a result of the form and 82% felt the form was well suited to the outpatient setting.

Eight residents involved in our study also completed an OSCE. The number of residents participating in the OSCE was low because many of the residents were completing electives outside the academic center when the OSCE was administered. The Pearson correlation coefficient comparing the overall OSCE score to the composite score from the evaluation forms was 0.48. Although this correlation was

*Table 3.* Means of pre- and post-faculty questionnaires and percentage of faculty that agreed or strongly agreed to the following statements. Only the first 6 items were included in the pre-questionnaire.

| Statement | Mean Score* of pre-questionnaire n = 12 | Mean Score* of post-questionnaire, n = 11 | Percentage of Faculty that Agreed or Strongly Agreed on pre-questionnaire, n =12 | Percentage of Faculty that Agreed or Strongly Agreed on post-questionnaire, n = 11 |
|---|---|---|---|---|
| I observe residents performing histories on a regular basis | 1.75 | 2.36 | 16.7 | 9.1 |
| I observe residents performing physical examinations on a regular basis | 3.08 | 3.64 | 50 | 45.5 |
| I observe residents at the bedside performing procedures | 3.1 | 3.55 | 40 | 54.6 |
| I provide feedback to residents on their histories and physical exams on a regular basis | 3.45** | 4.27** | 63.6 | 90.0 |
| I provide timely feedback to residents on their histories and physical exam | 3.67 | 4.0 | 75 | 72.8 |
| I find the in-training evaluation form a valid measure of resident competency | 3.36 | 3.55 | 36.4 | 54.6 |
| I observed residents at the bedside performing histories more often after using the evaluation card | | 2.09 | | 0 |
| I observed residents at the bedside performing physicals more often after using the evaluation card | | 2.55 | | 18.2 |
| I observed residents at bedside performing procedural skills more often after using the evaluation card | | 2.55 | | 27.3 |
| I provided more feedback to residents after using the evaluation card | | 3.55 | | 63.6 |
| I provided more timely feedback to residents after using the evaluation card | | 3.55 | | 63.6 |
| Overall, I feel the evaluation card is a valuable teaching tool | | 3.64 | | 63.6 |
| Overall, I feel the evaluation card is a valuable tool for evaluation | | 3.82 | | 72.7 |
| I feel the ITER are more objective when based on the evaluation cards | | 4.1 | | 80 |
| The evaluation card is time consuming | | 2.45 | | 27.3 |
| The evaluation card lends itself well to use in the outpatient clinic | | 4.18 | | 81.9 |
| I believe our division should continue to use the evaluation card | | 4.0 | | 72.8 |

* Mean score of rating scale, where 1 = strongly disagree and 5 = strongly agree. ** Pre- vs. post was significant (p = 0.02), all other comparisons of pre- and post-ratings were not statistically different. ITER: in-training evaluation report.

not significant ($p > 0.05$), it is moderately high in magnitude despite the small number of residents and would suggest the encounter cards may be somewhat predictive of performance on the OSCE.

## DISCUSSION

Our study is unique because the data were obtained from an ambulatory 1-month rotation of internal medicine residents and the evaluation was primarily based on case review and not direct observation. Direct observation is generally encouraged and increases validity, but requires more faculty time. We believe our study reflects a more realistic representation of what is feasible and actually occurring with the use of encounter cards in many universities.

Several positive findings were discovered. First, the evaluation forms demonstrated a high degree of reliability (0.71), with as few as 8.73 forms per resident collected over a 1-month rotation. It would take 14 forms per resident to achieve a reliability of 0.80. These findings are similar to previous reports that have used encounter cards[3-5]. Feasibility has also been demonstrated in our study. The number of forms required to be collected per resident is a realistic goal for most programs. In addition, faculty felt the forms were well suited to the ambulatory setting and both universities involved in the study have continued to use the forms although the study has ended. There were some cautions in terms of feasibility, however. Approximately 27% of faculty did report the form as time-consuming, so they may require ongoing encouragement. In addition, to ensure the use of the cards continues, we suggest that a dedicated

administrative assistant and faculty representative are important factors for programs that wish to implement a similar program. That said, once the system is established, it requires minimal time from administrative assistants and rotation coordinators.

The evaluation forms also appear to have face validity, as demonstrated by the favorable ratings provided by residents and faculty. For example, 79.1% of residents agreed or strongly agreed that the evaluation form was a fair evaluation of clinical skills. From the faculty perspective, 72.2% agreed or strongly agreed that the evaluation form overall was a valuable tool for evaluation. Over 70% of faculty and over 80% of residents agreed or strongly agreed that the evaluation forms should continue to be used. Although not statistically significant, the correlation between the composite score on the evaluation form and a formative OSCE score was moderate and could indicate a degree of criterion validity.

Finally, both residents and faculty reported increased feedback as a result of the forms. This concurs with data from a previous study[6]. This formative aspect is important and was reflected in resident comments. The form did not increase direct observation, indicating that if this is a primary objective, then another method of evaluation such as the mini-CEX should be considered[8].

Our study does have limitations. It was completed on internal medicine residents completing an ambulatory rotation in rheumatology. It is not clear if the results will be transferable to other outpatient settings and other trainees. The forms, although useful, did not increase observed clinical skills and should only be one of several methods of evaluation for any trainee. Faculty training would likely improve the performance of the forms.

Our study has demonstrated that the use of evaluation forms in a 1-month ambulatory clinical rotation is feasible, valid, and reliable and that it improved feedback on clinical performance. Evaluation forms provide an important method of evaluation for ambulatory rotations.

## REFERENCES


1. Turnbull J, van Barneweld C. Assessment of clinical performance: In-training evaluation. In: Norman G, van der Vleuten CPM, Newble DI, editors. International handbook of research in medical education. Dordrecht: Kluwer Academic Publishers; 2002:793-810.
2. Rhoton MF. A new method to evaluate clinical performance and critical incidents in anaesthesia: quantification of daily comments by teachers. Med Educ 1989;23:280-9.
3. Hatala R, Norman GR. In-training evaluation during an internal medicine clerkship. Acad Med 1999;74 Suppl 10:S118-20.
4. Al-Jarallah KF, Moussa MAA, Shehab D, Abdella N. Use of Interaction card to evaluate clinical performance. Medical Teacher 2005;27:369-74.
5. Brennan BG, Norman GR. Use of encounter cards for the evaluation of residents in obstetrics. Acad Med 1997;72 Suppl 10:S43-4.
6. Paukert JL, Richards ML, Olney C. An encounter card system for increasing feedback to students. Am J Surg 2002;183:300-4.
7. Frank JR, editor. The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.
8. Norcini JJ, Blank LL, Duffy DD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138:476-81.