INNOVATIONS IN EDUCATION AND CLINICAL PRACTICE

Clinical Work Sampling

A New Approach to the Problem of In-Training Evaluation

J. Turnbull, MD, J. MacFadyen, MD, C. van Barneveld, M Ed, G. Norman, PhD

OBJECTIVE: **Existing systems of in-training evaluation (ITE) have been criticized as being unreliable and invalid methods for assessing student performance during clinical education. The purpose of this study was to assess the feasibility, reliability, and validity of a clinical work sampling (CWS) approach to ITE. This approach focused on the following: (1) basing performance data on observed behaviors, (2) using multiple observers and occasions, (3) recording data at the time of performance, and (4) allowing for a feasible system to receive feedback.**

PARTICIPANTS: **Sixty-two third-year University of Ottawa students were assessed during their 8-week internal medicine inpatient experience.**

MEASUREMENTS AND MAIN RESULTS: **Four performance rating forms (Admission Rating Form, Ward Rating Form, Multidisciplinary Team Rating Form, and Patient's Rating Form) were introduced to document student performance. Voluntary participation rates were variable (12%–64%) with patients excluded from the analysis because of low response rate (12%). The mean number of evaluations per student per rotation (19) exceeded the number of evaluations needed to achieve sufficient reliability. Reliability coefficients were high for the Ward Form (.86) and the Admission Form (.73) but not for the Multidisciplinary Team (.22) Form. There was an examiner effect (rater leniency), but this was small relative to real differences between students. Correlations between the Ward Form and the Admission Form were high (.47), while those with the Multidisciplinary Team Form were lower (.37 and .26, respectively). The CWS approach ITE was considered to be content valid by expert judges.**

CONCLUSIONS: **The collection of ongoing performance data was reasonably feasible, reliable, and valid.**

KEY WORDS: **In-training evaluation; clinical work sampling.**
J GEN INTERN MED 2000;15:556–561.

The principal goal of undergraduate medical education is to prepare medical graduates to competently care for patients. An important part of this process is the observation, assessment and documentation of performance in the care of patients through intraining evaluation (ITE). Other objective measures of competency have been utilized during training programs; ITE has the potential to measure actual practice performance. ITE fulfills dual accountabilities: from the learner perspective, it provides a focus for improving skills and knowledge; from the societal perspective, evaluation discharges the institution's responsibility to ensure that the student has met or exceeded an expected performance level.[1,2] To ensure that graduates are ready to move onto residency, medical schools need effective systems to monitor students' progress through their clinical clerkship experiences.[3,4]

Too frequently, the "system" of evaluation in a clinical clerkship consists of a single global rating scale completed at infrequent intervals by a supervisor who may have had minimal contact with the student. Not surprisingly, research has shown that this approach has serious deficiencies. The reliability of scores generated through current approaches to ITE is close to zero[5,6] because students are generally rated "above average," resulting in limited real variation between students and because of the presence of random and systemic rater biases.[5,7–9] Additionally, there is an insufficient level of structured, documented feedback from supervisors to students during the student's rotation, possibly related to the limited direct observation of student performance by evaluators.[10,11] Finally, the current approach to in-training evaluation is costly in terms of the time and effort required for the minimal utility of the information.[12]

From a psychological perspective, a central issue is that the task of assessing the average performance of an individual student over a period of weeks or months places extreme demands on memory. *"In such instances, subjects must not only recall multiple events, but must also summarize them into a presumed 'average state,' introducing additional opportunities for bias."*[13]

With existing approaches to ITE, we cannot limit the bias in evaluations, or even be assured that the assessor had any opportunity for observation on which to base these evaluations. As a result of the difficulties of feasibility, reliability and validity, existing approaches to ITE are neither effective, accountable, nor educational.

A frequent response to the identification of these problems with current ITE approaches is to blame the evaluation form, and to revise it. The ongoing difficulties

Received from the Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada (JT, JM, CB); and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (GN).

Address correspondence to Dr. Turnbull: Room LM-14, 501 Smyth, Ottawa General Hospital, Ottawa, ON K1H 8L6, Canada.

experienced with ITE suggest that improvements must come, not from a revision of the procedures of global summative ratings, but from a major reconceptualization of the evaluation problem.

The Clinical Work Sampling (CWS) Strategy

It is evident that an effective system must, as a central feature, overcome the primary obstacle of retrospective recall, and must capture evaluation information as the opportunity arises, resulting in multiple observations, from multiple observers, in live time.[10,11] Many opportunities for informal evaluation of students occur in the daily interactions among health professionals in all settings. However, most of these interactions go unreported. The challenge is to devise a system which can capture a reasonable number of these clinical encounters while imposing a minimal additional administrative load on the individuals involved.

One model for this is the work sampling approach used in industry, where observers monitor and record activities at regular intervals (for example, every 15 minutes).[13] While logistically impossible to maintain on an ongoing basis without massive resources, the basic idea of obtaining multiple samples with minimal information from each is attractive. We applied these concepts of CWS during an inpatient internal medicine, unit rotation for the ITE of clinical clerks. Forms were developed to be easily used by evaluators which captured performance data during regular encounters over the course of the work day in a standardized fashion. In this manner, we hoped to avoid the memory decay and subjective averaging resulting from summarizing at the end of the rotation.

The present study was initiated to establish the effectiveness of a CWS approach to ITE. We addressed 3 specific research questions: (1) Is the CWS approach to ITE feasible? (2) Is this approach a reliable measure of student performance on the ward? and (3) Is the CWS approach to ITE a valid measure of student performance?

METHODS

Subjects

Sixty-two third-year University of Ottawa clinical clerks participated in the study during their 8-week internal medicine inpatient rotation between September 1996 and June 1997. During their medical clerkship, students perform admission assessments and assume the role of the primary medical caregiver for their assigned patients.

Instruments

A series of evaluation forms to assess the performance of students were developed for the purposes of the research project. Items were drawn from existing evaluation forms and were modified based upon those competencies considered necessary for the practice of medicine.[13]

Evaluation forms were incorporated as part of the day-to-day activities of the health care team. All evaluation data collected on the forms were computer scannable and utilized 5-point rating scales (ranging from "unsatisfactory" to "excellent") for each content domain. The forms are described in Table 1 and included:

1. Admission Rating Form: To capture data on student skills related to the admission of patients, an Admission History and Physical Rating Form was developed such that a copy of the summary sheet of the form contained a student evaluation section. Four content domains were assessed: communication skills, physical examination skills, diagnostic acumen, and management skills as well as a global rating of overall performance. Formative feedback on each performance was accommodated by a comments section on the form. Students were evaluated by their attending faculty supervisor upon the bedside presentation of each new patient under their care. This provided an opportunity for verbal feedback on student performance as well.
2. Ward Rating Form: To capture data on a student's performance in patient management, an evaluation form was developed and was included as part of the process of patient billing. Seven content domains were assessed; therapeutic strategies, communication skills, consultation skills, management skills, interpersonal behaviors, continued learning skills, and health advocacy skills as well as a global rating of overall performance. Formative feedback on each performance was accommodated by a comments section on the form. Students were evaluated by their supervisor at the time of patient discharge. Providing an opportunity for verbal feedback on student performance as well.
3. Multidisciplinary Team Rating Form: This evaluation form assessed six content domains; thera-

Table 1. Study Raters and Forms Utilized

Form	Rater	Timing	Desired Number of Responses
Admission Rating Form	Supervisor	Admission	1 per patient admitted
Ward Rating Form	Supervisor	Discharge	1 per patient cared for
Multidisciplinary Team Rating Form	Multidisciplinary team	1 per month	1
Patient's Rating Form	Patient	Discharge	1 per patient cared for

peutic strategies, communication skills, consultation skills with nurses and other health care providers, management of resources, discharge planning, and interpersonal relations as well as a global rating of overall performance. While student performance information was gathered daily using notes in a nursing log, the results for each student assessment were summarized by a nursing supervisor on the evaluation form every fourth week during multidisciplinary rounds for those students completing the rotation. They were asked to evaluate only those students with whom they felt they had a significant amount of contact. No physicians were involved. A comments section provided opportunity for formative feedback.

4. Patient's Rating Form: To capture input from patients, a seven-item evaluation form outlined four content domains; communication skills, collaboration skills, health advocacy skills and professionalism as well as a global rating of overall performance. A comments section provided opportunity for formative feedback. This form was administered through an interview format by a Research Associate.

Procedure

Following an initial piloting of the forms, the formal project was implemented over the following 9-month period on 2 general medicine clinical teaching units at 2 separate hospitals. A 2-hour faculty development workshop was arranged for the 18 attending staff, followed by monthly communications reminding evaluators of the basis of the project and the need to complete evaluations in all circumstances described.

Students were oriented at the beginning of their rotation, and subsequently met with the research associate on a weekly basis to discuss issues or questions related to the project. Students participated voluntarily (with consent). Completed evaluations were potentially formative, as written and verbal feedback were frequently given to students during the course of their rotation. This more formative element of the evaluations were used at the discretion of the evaluator. This system ran in parallel with the existing system of ITE which consisted of a single clinical rating form at the end of the rotation. Patients were oriented to the project at the time of interview.

Evaluation forms were collected on a per patient basis for each student as described. A variable number of forms of each type were available on each student. In order to assess compliance (and therefore feasibility), students were asked to keep an ongoing record of the patients they admitted, cared for, and discharged during the project. This record was updated weekly with the research associate.

Students would solicit evaluations of their performance directly from their supervisors. This provided an opportunity for verbal feedback as well as documentation of performance through the rating forms. Once completed and reviewed, rating forms were submitted by the students in several conveniently located drop-off boxes throughout each hospital during the clerkship rotation. Forms were cross-checked to assess compliance and then analyzed. While not the purpose of this study a final summary profile report was provided to the student at the end of each rotation.

Equipment

Teleform 4.0 and a Fugitsu ScanPartner 10C scanner were used to develop and scan the evaluation forms. BMD software was used to conduct statistical analysis of the data.

ANALYSIS

Feasibility

The feasibility of the CWS approach to ITE was measured by the return rate of completed evaluation forms and the average number of evaluations completed per student. For the purposes of this study, a feasible system would collect sufficient data to meet the appropriate requirements for reliability and validity. All potential opportunities where information could be captured were tracked, and contrasted with those interactions actually documented. The Spearman Brown Prophecy formula[15] was used as a method to determine the number of forms necessary to achieve an acceptable, within-method reliability.

Reliability

Reliability was assessed assuming each student was assessed by different evaluators using the formulas of Shrout and Fleiss.[14] While this is not strictly correct, as some assessors will have assessed more than one student, the result is a conservative estimate of reliability. The analysis was conducted using the student as a grouping factor, and item and rater as repeated measures, using a generalizability theory framework. Thus, for each form, we estimated the interrater reliability for each subscale and the internal consistency (α). Interrater reliability was computed for the average score based on the number of forms collected, since the system was designed to utilize a variable number of observations. As a secondary analysis, we reanalyzed the Ward and Admitting Rating Forms by evaluator to determine whether there were significant systematic differences among examiners as an indicator of examiner bias.

Validity

Content validity was ensured through a thorough process of internal and external expert reviews. Concurrent validity was assessed by examining the correlations of scores on the different measures. It was postulated that

in a valid system a modest correlation should exist (especially among raters completing the same forms) assuming that these forms documented overall performance even though from different perspectives.

RESULTS

Feasibility

During the 1-month clerkship rotation, students were evaluated on average, 19 times by a series of different evaluators at the time of the interaction in an objective fashion. The number of potential rating forms that could be collected per student from all sources, on average, ranged from 4.0 for the multidisciplinary team rating form to 35.2 for the ward rating form. On average, 8.4 admission rating forms were submitted per student, 8.1 ward rating forms were submitted per student, 1.7 multidisciplinary team rating forms were submitted per student and 1.6 patient rating forms were submitted per student. Response rates for each form were 64%, 23%, 43%, and 12%, respectively. The number of Ward Rating Forms completed per student exceeded the number of Admission Rating Forms, as students do not necessarily admit all the patients they ultimately care for. The average number of ratings obtained per student from the multidisciplinary team was 1.7 because students frequently cared for patients on more than one ward and they consequently had ratings completed by the teams on each ward. Finally, the return rate for Patient's Rating Forms was only 12%, an unacceptably low figure, and no further analysis of this form was conducted.

Reliability

The average number of forms completed per student, and the interrater and interitem (Cronbach's α) reliability for the different assessments are shown in Table 2. The reliability of the Multidisciplinary Team Rating Form by category (based on those students where 2 or more forms were collected) was low, ranging from .00 to .22. By contrast, for the Admission Rating Form, all reliability coefficients were in the range of .64 to .73, and for the Ward Rating Form, all coefficients exceeded .8. These reliability coefficients are high, and comparable with many high stakes examinations.

The Spearman Brown Prophecy formula was used to determine the number of assessments (completed forms) necessary to achieve an acceptable within method reliability. On the single item "overall impression" which concluded each form, 3.2 Ward Rating Forms, 7.4 Admission Rating Forms, and 18.9 Multidisciplinary Rating Forms would be required to achieve a reliability of .70.

Because a number of students had multiple assessments from the same rater, we were concerned that the observed reliable differences between students may actually reflect confounding with rater leniency. To assess this, we repeated the analysis for the overall rating category for the Ward and Admission Rating Forms using the rater as the grouping factor. For the Admission Rating Form, significant differences between raters were present ($F = 5.66$, $P < .01$); however, the average correlation across ratings of each examiner was .29. Similarly, for the Ward Rating Form, there was a significant difference in examiner ratings ($F = 7.56$, $P < .01$), but again the correlation was only moderate ($r = .19$). Thus, although some of the variation in scores was attributable to systematic differences in rater leniency, this was apparently small relative to real differences among students.

The high item-total correlations and the very high α coefficients for all forms (all greater than .96) suggested that raters were not able to differentiate among the behaviors in the different categories and were rather basing their assessment on a global impression of the clerk.

Validity or Relationship Between Measures

The correlations between the Admission and Ward Rating Forms were moderately high (.47). Correlations between the Ward and History and Physical Forms and the Multidisciplinary Team Rating Forms were lower (.37 and .26, respectively), reflecting the lower reliability of the Multidisciplinary Team Rating Form.

DISCUSSION

Feasibility

While response rates were low, the mean number (19) of evaluations submitted per student exceeded the minimum number needed to achieve an adequate level of reliability for the Ward and Admission Rating Forms, thereby meeting our requirement for feasibility. Compliance to this approach was facilitated by utilizing existing structures within the clinical clerkship rotation and data was captured in an easily scannable form.

Patient evaluations were not considered to be helpful in view of their very low return rate. Reasons for this include rapid patient discharges from the internal medicine floor prior to the arranged interview with the research assistant. When interviewed, patients were not often mentally able or were too unwell to complete the questionnaire, and finally, they often could not recognize their attending clinical clerk, even with prompting.

Reliability

Ratings provided by the multidisciplinary team were of limited reliability, despite the fact that ongoing notes were kept on student performance. The marginal reliability may be because they resembled the traditional one final clinical assessment with its attendant deficiencies.

By contrast, the Admission and Ward assessments demonstrated acceptable reliability with as few as 4 to 8 as-

Table 2. Reliability Analysis

ITEM	Mean* (SD)	Item/Total Correlation (corrected)	Interrater Reliability†	Standard error of measurement
Admission Rating Form (α = .92)				
Diagnostic/therapeutic plan	3.7 (0.54)	0.73	0.71	2.00
Differential diagnosis	3.8 (0.57)	0.69	0.64	2.38
Physical examination	3.9 (0.55)	0.76	0.71	2.10
Communication skills (written and verbal)	4.0 (0.52)	0.61	0.67	2.30
Overall impression	3.8 (0.51)	0.85	0.73	1.97
Ward Rating Form (α = .96)				
Diagnostic/therapeutic plan	3.3 (0.52)	0.73	0.81	1.44
Communication skills	3.6 (0.54)	0.76	0.83	1.48
Consultation skills	3.5 (0.51)	0.74	0.81	1.52
Management of resources	3.4 (0.57)	0.79	0.86	1.27
Health advocacy skills	3.4 (0.51)	0.68	0.82	1.44
Interpersonal skills	3.5 (0.59)	0.81	0.88	1.21
Fund of knowledge	3.3 (0.57)	0.81	0.81	1.44
Overall impression	3.4 (0.54)	0.87	0.86	1.27
Multidisciplinary Team Rating Form (α = .98)				
Diagnostic/therapeutic plan	3.7 (0.72)	0.74	0.14	3.4
Communication skills	3.8 (0.68)	0.74	0	3.8
Consultation skills	3.7 (0.80)	0.80	0	3.7
Management of resources	3.6 (0.66)	0.78	0	3.6
Discharge planning	3.7 (0.71)	0.75	0.18	3.3
Interpersonal skills	3.8 (0.72)	0.83	0	3.8
Overall impression	3.7 (0.72)	0.90	0.22	3.3

**Rating Scale: 1 = unsatisfactory, 2 = meets expectations, 3 = good, 4 = very good, 5 = excellent.*

†Interrater reliabilities were computed using the mean number of rating forms returned per student (i.e., 8.4 Admission Rating Forms per student, 8.1 Ward Rating Form per student, and 1.7 Multidisciplinary Team Rating Forms per student). This represented form response rates of 64%, 23%, and 43%, respectively. Patient evaluation forms were not included in the analysis due to low response rates (12%).

sessments. The reliability observed in these assessments is comparable to large scale objective assessments utilized for licensure and certification and differs substantially from traditional end of rotation clinical ratings, where reliability is generally unacceptable. Potential sources of an artificially high reliability were examined. There was a statistically significant rater bias; however, this only accounted for a small degree of the overall variance.

The high item-total correlations suggested that raters were unable to differentiate between behaviors. As a consequence, in this setting, the overall rating may provide as much information as the detailed category ratings, and the form could potentially be reduced to a single rating.

Validity

The CWS approach to ITE was considered to have content validity as these detailed forms did reflect the necessary domains of practice; the high item-total correlations suggest that evaluators may not be reflecting these competencies. The modest correlations between different forms provides further supportive evidence of validity. As expected, this was highest between physician raters (Admission and Ward Rating Forms) and modest between the Ward Rating Form and that of the Multidisciplinary Team Rating Form. Studies are underway to see if a correlation exists with other objective examinations throughout clerkship and beyond as a measure of predictive validity.

A direct comparison of the CWS results and the existing traditional approach to ITE using a single summary rating form was not made as it was felt that the lack of reliability of this approach had been clearly established. Poor return rates and delayed submission of traditional ITERs, along with questionable reliability and validity of information collected on them, did not allow for a meaningful comparison of the two approaches.

The reliability and validity of the CWS approach to ITE must be viewed as promising. While feasible in the context of this study, this approach has been slow to be accepted because of the lack of administrative assistance and the time required on a daily basis to provide effective evaluation. Alternate versions of the CWS approach (e.g., using computer based evaluations instead of paper-based) have been implemented in other disciplines where there is sufficient administrative support. However, in disciplines where there is limited support, the approach has not been implemented. The implications for other educators who may consider adopting this evaluation strategy is that the successful implementation of this approach depends, at least in the beginning, on having an administrative assistant that has dedicated time towards it.

While not a part of this study, additional input could be sought from other sources (such as peers), and it is

postulated that this approach is potentially formative as it provides an opportunity for verbal and or written feedback from evaluators who opt to use them, in a timely way. Further studies will be necessary to assess the predictive validity of this assessment tool and its applicability to different settings (such as ambulatory care) and levels of training (such as residency). This study does compliment existing work that supports input from multiple observers,[16,17] such as allied health professionals,[18] and that evaluates the essential components[19] of performance in practice in a valid fashion.

CONCLUSIONS

A system of evaluation of individual patient-based performances (as opposed to rotation-based evaluation), when measured at the time of the behavior by multiple observers in a standardized fashion is both feasible and reliable. Evidence is provided to support validity. The CWS approach to ITE has the potential to be a psychometrically defensible alternative to existing traditional methods of ITE.

With support from the Educating Future Physicians for Ontario Project, the Royal College of Physicians, and Surgeons of Canada, and the Medical Council of Canada.

REFERENCES

1. Hull AL, Hodder S, Berger B, et al. Validity of three clinical performance assessments of internal medicine clerks. Acad Med. 1995; 70:517–22.
2. Short JP. The importance of strong evaluation standards and procedures in training residents. Acad Med. 1993; 68:522–5.
3. Phelan S. Evaluation of the noncognitive professional traits of medical students. Acad Med. 1993; 68:799–803.
4. Hunt DD. Functional and dysfunctional characteristics of the prevailing model of clinical evaluation systems in North American medical schools. Acad Med. 1992; 67:254–9.
5. Gray JD. Global rating scales in residency education. Acad Med. 1996;71(1suppl): S55–S63.
6. Kaplan CB, Centor RM. The use of nurses to evaluate house-officers' humanistic behavior. J Gen Intern Med. 1990;5:410–4.
7. Dauphinee D. Assessing clinical performance: Where do we stand and what might we expect? JAMA. 1995;274:741–3.
8. van der Vleuten CPM, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ. 1991;25:110–8.
9. Maxim BR, Dielman TE. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. Med Educ. 1987;27:130–7.
10. Stillman PL. Positive effects of a clinical performance assessment program. Acad Med. 1991;66:481–3.
11. Turnbull J, Gray J, MacFadyen J. Improving in-training evaluation programs. J Gen Intern Med. 1998;13:317–23.
12. Irby DM, Milam, S. The legal context for evaluating and dismissing medical students and residents. Acad Med. 1989;64:639–43.
13. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Beh Med. 1994;16:199–202.
14. Fleiss J, Shrout, PE. Approximate interval estimation for a certain inter-class correlation coefficient. Psychometrika. 1978;43:259–62.
15. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use. Oxford, U.K.: Oxford University Press; 1995.
16. Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med. 1996;71:364–70.
17. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, Logerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655–60.
18. Butterfield PS, Mazzaferri EL. New rating form for use by nurses in assessing residents' humanistic behavior. J Gen Intern Med. 1991;6:155–61.
19. Societal Needs Working Group, CanMEDS 2000 Project. Skills for the new millennium. Ann RCPSC. 1996;29:206–16.