Validity, reliability and feasibility of assessment of clinical reasoning of medical students by observation versus post encounter assessment in a clinical practice setting.

CURRENT STATUS: UNDER REVISION

Catharina Maria Haring
Radboud university medical center

Claudia Klaarwater
Radboud University medical center

Geert Bouwmans
Radboud University Medical Center

Bernadette Cools
Radboudumc

Petra van Gurp
Radboud University Medical Center

Jos van der Meer
Radboud University Medical Center

Cornelis Postma
Radboud University Medical Center

DOI: 10.21203/rs.2.20039/v2

SUBJECT AREAS
Internal Medicine Educational Philosophy and Theory

KEYWORDS
History taking, clinical reasoning
Abstract
Background The assessment of clinical reasoning by medical students in clinical practice is very difficult. Partly this is because the fundamental mechanisms of clinical reasoning are difficult to uncover and when known, hard to observe and interpret.

Methods We developed an observation tool to assess the clinical reasoning ability of medical students during clinical practice. The observation tool consists of an 11-item observation rating form. The validity, reliability and feasibility of this tool were verified among medical students during the internal medicine clerkship and compared to an existing post-encounter rating tool.

Results Six raters assessed each, the same 15 student-patient encounters. The internal consistency (Cronbach’s alfa) for the observation rating tool (ORT) was 0.87 (0.71-0.84) and the 5-item post encounter rating tool (PERT) was 0.81 (0.71-0.87). The intraclass-correlation coefficient for single measurements was poor for both the ORT; 0.32 (p<0.001) as well as the PERT; 0.36 (p<0.001). The generalizability study (G-study) and decision study (D-study) showed that 6 raters are required to achieve a G-coefficient of > 0.7 for the ORT and 7 raters for the PERT. The largest sources of variance are the interaction between raters and students. There was a correlation between the ORT and PERT of 0.53 (p=0.04)

Conclusions The ORT and PERT are both feasible, valid and reliable instruments to assess students’ clinical reasoning skills in clinical practice.

Introduction
Assessment of clinical reasoning of medical students in clinical practice, is a complicated and tricky process. In clinical practice the performance of students is profoundly influenced by context and content specificity of the clinical problems involved. To make it even worse, large inter-rater differences are known to exist, which are due to different frames of reference of the clinical assessors. What complicates matters even more is that there is no consensus on what clinical reasoning exactly comprises and what the driving forces are that determine the process.

It is generally accepted, that in workplace-based assessments, one should not rely on a single measurement to come to a robust conclusion. (1, 2) Every assessment of clinical reasoning by clinical
teachers will contain some level of subjectivity. This is more or less unavoidable because of the nature of this assessment. However, if this could be integrated within a framework of pre-specified carefully defined objective criteria, the subjectivity of the assessors might be corrected enough to make the assessment more reproducible and reliable. It is therefore not surprising that there is consensus that only repeat assessments can give reliable outcomes.

In clinical practice, assessment of clinical reasoning involves either direct observation of a clinical encounter between a student and a patient (either live or video recorded) or an assessment of an oral or written report after completion of such an encounter. Both methods have their advantages and disadvantages. Observation takes time, which has an inhibitory effect in the clinical setting, but it is a very powerful method also for targeted feedback. (3) Assessment of an oral or written report can be less time consuming. Students can explain their analysis and interpretation, but essential information about the data-gathering ability or the diagnostic reasoning during the encounter can easily be missed.

For assessment of a student's clinical reasoning after an encounter took place, tools already exist. An example is the post-encounter form for clinical reasoning by Durning et.al. (4) This form is used to assess a predefined free text post-encounter form that is used by students. Validity, reliability and feasibility is tested in an objective structured clinical examination (OSCE) setting but had not been tested in a setting with real patient encounters.

For assessment of clinical reasoning during observation of a real encounter, we found no formats that were analyzed for validity and reliability. Of course, there is experience with residency training, during which assessment of clinical reasoning in clinical practice is often incorporated in mini clinical evaluations or a related single work-based encounter assessment instrument. The validity and reliability of many of these instruments in basic medical education is often not properly established and clinical reasoning is mostly only superficially itemized in these instruments. (5) This makes them less suitable for more in-depth exploration of clinical-reasoning abilities of students and trainees in the clinical phase of their training.

So, there is a clear need of an easy to use reliable tool to assess the clinical-reasoning skills of
medical students in the clinical-practice setting, in which real patients are involved. It should enable the teaching clinicians to assess a student’s diagnostic reasoning during a clinical encounter with a patient and provide the clinical teacher with information regarding the ability of the student and the student with proper feedback. The first step in the creation of such a tool is the definition of the criteria that can be used to assess the clinical-reasoning process of a student. In a previous study we defined those criteria. (6) With these criteria a clinician can observe a student during a student-patient encounter and assess the student’s skills. Unanswered is the question how reliable, valid and feasible these criteria are if used in an observation tool in clinical practice. This question remains also unclear for the post-encounter assessment in clinical practice. Hence our research question was: What is the validity and reliability and feasibility of a new observation tool and a post-encounter rating tool in assessment of clinical reasoning of medical students in practice during their clerkship internal medicine?

Methods

Setting and participants

The study was done in a real clinical setting, at the Department of Internal Medicine in our university medical centre, in which students are already familiar with video-observation as part of their training. The clerkship internal medicine is the first clinical clerkship in the first master year (M1). The outpatient clinic of internal medicine has two prepared consulting rooms with a fixed camera set-up used for educational purposes.

Participating students were asked to record the history taking during their encounters with real patients, starting in the outpatient clinic from their third encounter. The first encounters could therefore be used to get used to the clinical practice setting. Students were not accompanied by their supervisor during the encounter with the patient, but received the usual feedback from the supervising physician directly after case presentation. Thereafter, the student and supervisor met with the patient together. After the encounter, the supervisor registered a global rating for clinical reasoning as a grade from 1-10 in accordance to usual practice.

For this study, the students would fulfil an extra activity; the completion of a post-encounter form
(PEF). This was done after history taking, before receiving feedback of their supervising physician. Students would send their completed PEF digitally to the researchers.

Participating assessors involved in assessment of the students were six principal lecturers, i.e., clinicians with degrees and assignments in medical education. They were not involved as the supervising physicians of the students at the time the recordings were made. These participants received a short instruction about the two rating tools. They were asked to observe the recordings as long as they deemed needed to complete the observation rating tool (ORT). After completing this form, they also completed the post-encounter rating tool (PERT). The time needed to fulfil the assessment was registered by the participants.

Materials

The Observation-rating Tool (ORT) was composed using data from previous qualitative research (6). It consists of eleven pairs of opposite statements about student behaviour related to clinical reasoning. On 5-point scale participants could rate which of the opposite statements was most applicable.

(additional file 1)

The student post-encounter form (PEF) and the post-encounter rating tool (PERT) were modified after Durning et al (4). (additional file 2&3) The student post-encounter form consisted of 5 items based on essential parts of clinical reasoning; summary statement, problem list, differential diagnosis, most likely diagnosis and support for the most likely diagnosis. Assessment of the 5 items was done using a 5-point Likert-scale.

Case selection

Twenty students were invited to participate in the study. One case per student was used. The first case that did not meet the exclusion criteria was selected. Cases were excluded when a student was not well visible or when audio quality was low, or the patient was not able to communicate easily, e.g., because of language barrier. Cases that would not induce clinical reasoning, for example when a patient presented with an established diagnosis, were also excluded. After inclusion of 15 cases,
selection was stopped. We calculated that 15 cases and 6 observers were needed to reach a power of > 80% to detect an intra-class correlation of 0.30 for the new observation form (7).

Measurements

Feasibility was measured as student completion rate and the time the assessors needed for completing of both the rating instruments and filling out the assessment of satisfaction with the instruments. We included time needed for completion, because it is a limiting factor in clinical practice.

Validity for both instruments was measured as follows: (8) Content validity; For both instruments content validity was measured in two previous studies (4, 6). Internal structure; analysis of internal consistency and generalizability study to explore factors of variance. Response process: This was measured by rater evaluation of both instruments. Relation with other variables: This was done by analysing the association between both instruments.

Reliability of the use of the post-encounter form was already analysed in an OSCE (4), but not for assessment in clinical practice. For both rating tools (PERT and ORT) inter-rater reliability was tested and a generalizability study (G-study) was performed including a decision study (D-study) to assess how many raters would be needed for each instrument to reach an acceptable reliability, since it is known for workplace based assessment tools that reliability, when using one rater, is often low.

Statistics

Cronbach’s alfa was used to calculate internal consistency of scales. Inter-rater agreement was computed using the Intra Class Correlation. We used the ‘two-way mixed model’, because assessments will be done by a selected group of assessors, with measures of ‘consistency’, because the assessment will not be used as a pass or fail test. We used ‘single measures’, since one assessor on the work floor usually performs the assessment during the clerkships.

A Generalizability study was conducted to identify various sources of variance. For evaluation of the observation rating tool a two-facet crossed design with six assessors and 11 items was used. For
evaluation of the post-encounter rating tool a two-facet crossed design with six assessors and 5 items was used. A relative G coefficient was computed since we were mostly interested in the rank order of the measurement objects rather than consistency in raw scores.

A Decision study (D study) was performed to forecast changes in G coefficients with alternate levels of facets (assessors and checklist items).

The association between the two rating tools was calculated using Pearson’s correlation coefficient. We used SPSS version 20 for intra-class correlation and Pearson’s correlation coefficient.

Generalizability study was done using the G1 SPSS program (9)

Ethics
Participation was voluntary for all students. All patients were informed about the study, agreed and signed the informed consent form for recording of history taking for educational purposes. Approval of the institutional ethical review board was not required according to the guidelines of the institutional research board based on the Dutch law (WMO)

Results
Eighteen out of twenty students agreed to participate in the study. In the study period 81 new encounters were planned. 30 recordings were completed including post-encounter form in combination with a qualitative suitable recording of which 15 were selected for the study.

Observation of the student and completing the ORT on average 32 minutes. Completion of the PERT took on average 4 extra minutes. Students needed on average 6.5 minutes to complete the post-encounter form.

The average total score for the ORT was 32.2 (range 25.5-42.8) and for the PERT was 12.8 (8.00-16.6) and the rating of the supervisor 7.38 (range 6.00-8.00). There was a significant correlation between the ORT and PERT of 0.53 (p=0.04), but no significant correlation between the PERT and the supervisor rating or the ORT and the supervisor rating.

Internal structure
The internal consistency (Cronbach’s alfa) for the 11- item ORT was 0.87 (0.71-0.84) and the 5-item PERT was 0.81 (0.71-0.87).

Inter-rater reliability
The intra-class-correlation coefficient was poor for both the ORT; 0.32 (p<0.001) as well as the PERT;
0.36 (p<0.001) for a single measurement.

**G- and D-study**

The G and D-study (figure 1) show that six raters were required to achieve a G-coefficient of > 0.7 for the ORT and seven raters for the PERT. The largest sources of variance (table 1) were caused by the interaction between raters and students and general sources of error (persons x item x rater) that cannot be further unravelled.

|                | PERT % variance | ORT % variance |
|----------------|-----------------|----------------|
| Participant    | 16              | 12             |
| Item           | 1.5             | 5.0            |
| Rater          | 7.8             | 4.1            |
| participant x item | 12             | 5.5            |
| participant x rater | 24             | 23             |
| Item x rater   | 6.5             | 10             |
| participant x item x rater | 33             | 40             |

Variance components from the generalizability study for the PERT and ORT

**Response process:**

Medical students reported that completing the post-encounter form was mostly done after the physical examination for practical reasons. Assessors reported that the assessment procedure was time consuming. The item ‘body language’ in the rating tool was regarded difficult to assess. The assessors all regarded the content of both rating tools comprehensible and adequate.

**Discussion**

In this study we developed and evaluated a new instrument, the Observation Rating Tool (ORT) to assess the clinical-reasoning skills of medical students in the clinical-practice setting. We could demonstrate that the content and construct validity of this instrument were high. We developed the ORT based on the results of a study with experienced clinical teachers (6). When we compared our instrument with a modified post-encounter rating tool that was developed and proved reliable before (4), we found a significant correlation between the two. However, we found no significant correlation of the global supervisor rating with either the ORT or the PERT. We interpret this discrepancy as a sign of the poor standardisation and the subjectivity of the traditional global rating after an
encounter.

Also, the inter-rater reliability of both instruments appeared to be poor for a single measurement.

These findings are in line with previous studies that investigated the characteristics of work-place based assessment methods. (5)

To reach a G-coefficient of > 0.7 with the ORT, 6 raters are required. This is not an uncommon requirement. For example, for a tool like the MINI-cex 5 to as many as 60 raters were needed to reach a G-coefficient of 0.7 in different studies. (10) So as already alluded to, one should realize that the assessment of a complicated task such as clinical reasoning cannot not be captured in one observation. Almost all methods to assess proficiency in clinical skills encounter these kinds of problems. As said, a major reason is personal bias of clinical teachers. Even teachers within the same discipline (e.g., internal medicine) and affiliated with the same institution, will vary regarding their opinions on clinical skills. This is because raters make and justify judgments based on personal theories and performance constructs. (11)

To attain a fair assessment of medical students, it is worthwhile to discuss these diverse ratings in teach-the-teacher sessions. As this study shows, the low inter-rater reliability of these instruments for a single encounter can be tackled by using more raters assessing the same encounter. In practical terms it is preferable to enhance reliability by having different raters assessing different patient encounters of the same student. Indeed, the latter approach has been found more effective to improve reliability than using one rater to rate different encounters. (12). Several investigators who have evaluated the reliability of clinical assessment methods arrive at this recommendation. (13, 14)

An additional advantage is that students will receive feedback on various clinical encounters and problems from different teachers. Ideally, these encounters should deal with a variety of clinical problems to overcome the problem content specificity. In conclusion the ORT and PERT can be used as formative tests for high stakes exams needing 6 to 7 assessors. However, they can also be used as summative tests by involving various clinical encounters and different raters as part of a longitudinal assessment program in clinical reasoning.

The largest source of variance in our study appeared to be the interaction between person x rater for
both instruments (table 1). This implies that the ranking order of the students varied greatly between the raters. Unfortunately, it is known that rater training has little effect on improvement of outcome of workplace-based assessment. (15) Although the teachers involved have to be familiarised with the procedures of course.

It should be noted that the present study used video recordings of clinical encounters and not direct, real-time observation. The obvious reasons for that were of practical nature: in this way the assessors could make the observations at a time and a place that suited them most and the assessor could also be another clinical teacher and not necessarily the supervisor in the clinic.

Next to that real-time observation has the disadvantage that the presence of an observer during the encounter of the student with the patient influences the performance.

Although we have not tested it in this study, it is likely that the ORT instrument can also be used during real-time observations. In further research it should be tested how many cases are needed in that situation to get a reliable assessment of clinical reasoning.

In clinical medicine, most clinicians have a tight schedule, and hence there is a risk that no time will be reserved to observe the video recording of the encounter of the student with the patient. In fact, it is known that observation in clinical practice often does not regularly take place. (16). This means that the observation and the rating (including the feedback to the student) should be part of the planned daily duties and recognized as an important task. If it’s not possible to observe the students in the clinic in real time, because of busy schedules, the debriefing with the student could also take place during out of clinic educational hours.

The relatively weak correlation between the ORT and the PERT suggests that clinical reasoning as is performed during an encounter does not necessarily reflect the contents and the quality of the analysis that will be written down afterwards while completing a post encounter form. This problem has been recognized before: judgment of clinical problem-solving skills by observation of medical students poorly correlates with the judgment of their written analysis. (17) Thus, both assessment methods are not exchangeable. It is clear that quite different skills are involved in showing how to perform clinical reasoning during encounter with a patient and how to put the analysis on paper in a
medical record afterwards. For clinical teachers this means that both elements of clinical practice need attention in the guidance of medical students, and that different instruments are needed to assess these two.

Our study has some limitations. First of all, our study was limited to one clinical encounter per student. Extension to more encounters per student is needed to establish how many encounters should be observed for proper judgment of the clinical reasoning abilities of the student practice. Ratings on the post-encounter form may have been influenced by the observed performance of the student during the encounter. This bias is hard to avoid, since the assessor needs to know which clinical information was used for the post encounter form that was recorded by the student.

Thirdly, because of the study setting and for practical reasons, the assessors gave no direct feedback to the students. When used in clinical practice, combined direct observation and assessment of the post-encounter form will provide an excellent opportunity for meaningful feedback on clinical reasoning.

Conclusions
In conclusion, the rating tool presented in this paper provides clinical teachers with an instrument to assess the quality of the clinical reasoning during a student’s encounter with a patient. in our opinion this instrument fills a niche and is a first step towards building consensus among clinical teachers and towards more objectivity in the assessment of medical students during their practical learning.

Abbreviations
PEF: post-encounter form
PERT: post-encounter rating tool
ORT: observation rating tool

Declarations
Ethics approval and consent to participate

Approval of the Radboud university medical center institutional ethical review board was not required according to the guidelines of the institutional review board. Participation was voluntary.

Consent for publication
Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding

Authors contributions

CH participated in the design of the study, data acquisition, data analysis and manuscript drafting. CK participated in the design of the study, data acquisition and data analysis. GB participated in data analysis. BC and PG participated in the designs of the study and data acquisition. JM participated in the design of the study and manuscript revision. CP participated in the design of the study, data acquisition, study coordination and manuscript revision.

Acknowledgements

Not applicable

Authors ‘information

Catharina M Haring, MD is a nephrologist at Deventer Ziekenhuis, previously employed as a researcher in Medical Education at the Radboud university medical center. Her main research interest is assessment and obtaining of clinical competence in Internal Medicine.

Claudia C.R. Klaarwater is a former medical student at the Radboud university medical center. She
works at the Erasmus University Medical Center in Rotterdam and is training to become a cardiologist.

Geert A. Bouwmans is a psychologist working at the department of medical education at the Radboud university medical center.

Jos WM van der Meer, MD, PHD is emeritus Professor of Medicine, principal lecturer in medical education and former head of the department of Internal Medicine of the Radboud university medical center.

Bernadette M Cools, MD is an internist and junior principal lecturer in medical education at the department of Internal Medicine of de Radboud university medical center. Her main medical education interest is preclinical skills training.

Petra JM van Gurp, MD is an internist and principal lecturer in medical education at the department of Internal Medicine of the Radboud university medical center. Her main medical education interests are professional behavior and medical communication.

Cornelis T Postma, MD PhD is emeritus associate professor of Medicine and principal lecturer in medical education at the department of Internal Medicine of the Radboud university medical center. His main medical education interests are in the training of practical medical education and medical competence and in the field of clinical assessment.

References
1. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. Med Educ. 2012;46(1):38-48.

2. Ilgen JS, Humbert AJ, Kuhn G, Hansen ML, Norman GR, Eva KW, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 2012;19(12):1454-61.

3. Ende J. Feedback in clinical medical education. JAMA : the journal of the American Medical Association. 1983;250(6):777-81.

4. Durning SJ, Artino A, Boulet J, La Rochelle J, Van der Vleuten C, Arze B, et al. The
feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. Med Teach. 2012;34(1):30-7.

5. Pelgrim EA, Kramer AW, Mokkink HG, van den Elsen L, Grol RP, van der Vleuten CP. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ Theory Pract. 2011;16(1):131-42.

6. Haring CM, Cools BM, van Gurp PJM, van der Meer JWM, Postma CT. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. BMC Med Educ. 2017;17(1):147.

7. Hintze J. Power Analysis Sample Size System (PASS) Quick start Manual: Kaysville, Utah USA: NCSS; 2011.

8. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19(2):233-50.

9. Mushquash C, O'Connor BP. SPSS and SAS programs for generalizability theory analyses. Behavior research methods. 2006;38(3):542-7.

10. Mortaz Hejri S, Jalili M, Masoomi R, Shirazi M, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: A BEME review: BEME Guide No. 59. Med Teach. 2019:1-18.

11. Govaerts MJ, Van de Wiel MW, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: raters' performance theories and constructs. Adv Health Sci Educ Theory Pract. 2013;18(3):375-96.

12. Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, et al. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. Acad Med. 2006;81(10 Suppl):S56-60.
13. van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best practice & research Clinical obstetrics & gynaecology. 2010;24(6):703-19.

14. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, et al. 2018 Consensus framework for good assessment. Med Teach. 2018;40(11):1102-9.

15. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. Journal of general internal medicine. 2009;24(1):74-9.

16. Pulito AR, Donnelly MB, Plymale M, Mentzer RM, Jr. What do faculty observe of medical students' clinical performance? TeachLearnMed. 2006;18(2):99-104.

17. Woolliscroft JO, Calhoun JG, Beauchamp C, Wolf FM, Maxim BR. Evaluating the medical history: observation versus write-up review. Journal of medical education. 1984;59(1):19-23.

Figures
Figure 1

Estimated G-coefficients from the D-study as a function of number of raters per assessment format

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

Additional file 3 PERT def.docx
Additional file 2 PEF.docx
Additional file 1 ORT def.doc