Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: informahealthcare.com/journals/imte20

# Clerkship evaluation–what are we measuring?

Dr. Kevin Mclaughlin, George Vitale, Sylvain Coderre, Claudio Violato & Bruce Wright

**To cite this article:** Dr. Kevin Mclaughlin, George Vitale, Sylvain Coderre, Claudio Violato & Bruce Wright (2009) Clerkship evaluation–what are we measuring?, Medical Teacher, 31:2, e36-e39, DOI: 10.1080/01421590802334309

**To link to this article:** https://doi.org/10.1080/01421590802334309

Published online: 03 Jul 2009.

2009; 31: e36–e39 MEDICAL TEACHER

WEB PAPER

# Clerkship evaluation – what are we measuring?

KEVIN McLAUGHLIN, GEORGE VITALE, SYLVAIN CODERRE, CLAUDIO VIOLATO & BRUCE WRIGHT
University of Calgary, Canada

## Abstract

**Background:** As society's expectations of physicians change, so must the objectives of training. Professional organizations involved in training now emphasize multiplicity of roles. But how well do we evaluate these multiple roles?

**Aims:** To investigate the principal components of evaluation in the Internal Medicine clerkship rotation at the University of Calgary.

**Methods:** We performed factor analysis on all evaluation components in the Internal Medicine clerkship rotation, including the in-training evaluation report (ITER), objective structured clinical examination (OSCE), and multiple choice questions (MCQ) examination.

**Results:** We identified three principal components: information processing, professionalism, and declarative knowledge. Both the OSCE and MCQ loaded on a single factor, declarative knowledge. The nine items on the ITER loaded on two factors – information processing and professionalism.

**Conclusions:** Despite using 11 evaluation items on three tools, we identified only three principal components of evaluation. Both our MCQ and OSCE appeared to measure declarative knowledge. The latter may be due to the fact that we use standardized patients without clinical findings – such that evaluations are primarily based upon the demonstration of examination routines. Reasons for the lack of discriminant validity of our ITER include overlapping attributes and constant errors, including a halo effect and an error of leniency.

## Introduction

As society's expectations of physicians change, so must the objectives of medical training. Recently, there have been calls for medical schools to develop, teach and assess the various attributes that are associated with being an effective physician – including personal characteristics such as altruism, empathy, integrity, and compassion (Cohen 2001; Ferguson et al. 2002; Albanese et al. 2003). Professional organizations, such as the Accreditation Council for Graduate Medical Education, the American Board of Medical Specialties, and the Royal College of Physicians and Surgeons of Canada, now emphasize the multiplicity of physician roles, such as those articulated in the CanMEDS framework (Medical School Objectives Writing Group 1999; Accreditation Council for Graduate Medical Education 2001; CanMEDS 2005). But how should we evaluate these multiple roles?

Compared to knowledge and clinical skills, personal attributes that contribute to 'professionalism' are more difficult to define. Consequently, their measurement is more subjective and prone to sources of constant error. For example, an inclination to rate every student in a certain direction may produce an error of severity, leniency, or central tendency (Kerlinger & Lee 2000; Daelmans et al. 2005). Another source of constant error is the 'halo effect', where the perception of a particular attribute is influenced by the perception of the former attributes in a sequence of interpretations: the first attributes we recognize influence our perception and interpretation of latter attributes (Thorndike 1920; Iramaneerat & Yudkowsky 2007). Sources of constant error can reduce discriminant validity of measurement tools so that we may not be measuring what we think we are. So what are we really measuring?

In this study our objective was to identify the principal components of assessment on the Internal Medicine clerkship rotation at the University of Calgary and to explore potential sources of constant error in our evaluations.

### Practice points

- Our OSCE appeared to measure declarative knowledge, perhaps due to the use standardized patients with no clinical findings.
- Our ITER loaded on just two principal components – information processing and professionalism.
- Overlapping items may partly explain poor discriminant validity of our ITER.
- A halo effect and an error of leniency also contribute to poor discriminant validity.

*Correspondence*: Dr. Kevin McLaughlin, Division of Nephrology, Foothills Hospital, 1403 29th Street NW, Calgary, Alberta T2N 2T9, Canada. Tel: 403 944 2510; Fax: 403 944 3199; Email: kevin.mclaughlin@calgaryhealthregion.ca



ISSN 0142–159X print/ISSN 1466–187X online/09/020036–4 © 2009 Informa UK Ltd.
DOI: 10.1080/01421590802334309

## Method

### Study design and sample

This was a prospective observational study, conducted over 12 months, during which we collected the results of all evaluations for 103 final year medical students during the Internal Medicine clerkship at the University of Calgary. The Internal Medicine clerkship is 12 weeks long and includes a mandatory four weeks rotation on a medical teaching unit (MTU) at one of two university-affiliated hospitals. We collected each student's in-training evaluation report (ITER) for the MTU rotation, in addition to the results from their formative objective structured clinical examination (OSCE) and the summative multiple choice question (MCQ) examination.

We did not seek approval by an ethics board as this study was part of an ongoing quality improvement initiative for the Internal Medicine clerkship program.

### Components of the evaluation tools

The ITER comprised eight individual items in addition to an overall score of student's performance. The eight individual items on the ITER were: data processing skills; clinical skills; knowledge of subject area; relationships with patients and their families; professional relationships; educational attitudes; initiative, interest and team relationships; and attendance and dependability. Each individual item was rated using a four-point scale (1 = unsatisfactory; 2 = below expected level; 3 = at expected level; and 4 = above expected level). The overall student's performance was rated using a five-point scale (1 = unsatisfactory; 2 = below expected level; 3 = at expected level; 4 = above expected level; and 5 = outstanding). All preceptor raters were specialists in General Internal Medicine and the ITER ratings reflected input from all preceptors and residents.

The OSCE included eight stations where students took a history and demonstrated examination routines on a standardized patient. The MCQ examination comprised 60 problem-solving questions for which there was a single best answer.

### Statistical analyses

We assessed reliability of the evaluation tools using Cronbach's $\alpha$ coefficient. To assess discriminant validity we performed exploratory factor analysis on the individual components of the ITER, in addition to the OSCE and MCQ. This technique reduces a set of items to a smaller number of underlying principal components and, in so doing, uncovers the latent structure of the set of items (Kerlinger & Lee 2000). Factor analysis can evaluate discrimination by testing statistically whether two or more items differ. Items are considered to be measuring different constructs if they load most heavily on different principal components (Straub 1989). Items that load most heavily, or converge, on the same principal component are considered to be measuring the same construct.

In our analysis we firstly created a Pearson product moment correlation matrix for the ITER items and then used principal component analysis to extract factors. We used a cut-off threshold for factor extraction of eigenvalue $\geq 1$ (Kaiser rule). We then performed factor loading on extracted factors, followed by factor rotation using the Varimax method with Kaiser normalization (Kerlinger & Lee 2000). We used a cut-off threshold of 0.5 for factor loading. We used SPSS statistical software for all our analyses.

## Results

The alpha coefficient for the ITER was 0.86. For overall performance no student was rated unsatisfactory, 2.9% were below expected level, 44.7% were at expected level, 36.9% were above expected level, and 15.5% were outstanding. The mean overall rating [95% CI] was 3.6 [3.49, 3.81], which was significantly higher than 3.0, the rating corresponding to performance 'at the expected level' ($p < 0.0001$). There was no difference in the mean ITER rating between any of the 12 week rotations. For the OSCE the alpha coefficient was 0.73. The mean ($\pm$SD) OSCE score was 85.0% ($\pm$3.5) with a minimum performance level (MPL) score of 77.8%. For the MCQ the alpha coefficient was 0.75. The mean MCQ score was 69.8% ($\pm$8.9) with a MPL score of 56.2%.

Based on the Kaiser rule, we extracted three principal components (eigenvalues = 5.0, 1.2 and 1.0) accounting for 65% of the total variance. The varimax rotated factor matrix is shown in Table 1 along with the eigenvalues and variance for the individual components. Based upon the pattern of factor loading observed we identified three principal components in student evaluation, which we considered to represent information processing, professionalism, and declarative knowledge, respectively. The nine components of the ITER loaded on two factors – information processing and professionalism – with the overall rating for student performance loading on information processing. Both the OSCE and the MCQ examination loaded on a single factor, declarative knowledge.

## Discussion

### Principal components of evaluation

In this study we found that the eleven components of the evaluation on the Internal Medicine clerkship loaded on three factors, which we labeled as information processing, professionalism, and declarative knowledge, respectively.

The MCQ examination is accepted as being an evaluation of knowledge, so it was not surprising to find this loading heavily on declarative knowledge. Although typically considered an evaluation of higher cognitive function, more in keeping with information processing, the OSCE also loaded on declarative knowledge. This may be due to the fact that in our OSCE we use standardized patients without clinical findings – if there is no information for students to process the evaluation is based upon the demonstration of examination routines.

In contrast to the MCQ and OSCE, the ITER is considered as a tool to evaluate multiple attributes simultaneously. But our nine item ITER loaded on only two factors. This lack of discriminant validity means that we were evaluating far fewer attributes than we had initially intended. So why does our ITER lack discriminant validity?

**Table 1.** Orthogonally rotated principal component matrix to the normalized varimax criterion of ITER, OSCE and MCQ variables.

| | Principal component | | |
|---|---|---|---|
| Variable | Information processing | Professionalism | Declarative knowledge |
| Data processing skills | 0.76 | | |
| Overall ITER rating | 0.73 | | |
| Knowledge of subject area | 0.70 | | |
| Clinical skills | 0.69 | | |
| Educational attitudes | | 0.80 | |
| Attendance and dependability | | 0.69 | |
| Initiative, interest and team relationships | | 0.67 | |
| Professional relationships | | 0.62 | |
| Relationships with patients and their families | | 0.55 | |
| MCQ | | | 0.86 |
| OSCE | | | 0.55 |
| Eigenvalues | 5.0 | 1.2 | 1.0 |
| Percent of variance | 45.7 | 10.6 | 8.9 |

### Variables affecting discriminant validity of evaluation tools

The ability of an evaluation tool to identify separate constructs is influenced by the statistical test used to detect discrimination, the nature of the items in the evaluation tool, and sources of bias in evaluation.

Items can appear to converge on a single construct if the threshold for identifying principal components is set too high. The Kaiser rule of including components with eigenvalues of $\geq 1$ is, however, considered to be a conservative criterion and tends to overestimate, rather than underestimate, the number of component (Lance et al. 2006). It is, therefore, unlikely that our analysis underestimated the number of principal components.

Items converge when they measure overlapping constructs. This may account for some of the poor discriminant validity of our ITER. Five ITER items converged on the principal component professionalism, of which three included the word 'relationships', while the other two were attitudes and dependability – key attributes to forming good relationships. It is not surprising, therefore, that these five items converged onto a single component as they appear to be measuring overlapping constructs. But overlapping constructs is not a good explanation for the convergence of other attributes, such as data processing skills, clinical skills and knowledge of the subject area.

Non-overlapping constructs may converge on factor analysis as a result of a systematic bias in rating of the ITER. On a clinical rotation attributes such as data processing skills, clinical skills and knowledge of the subject area are usually inferred from presentation of cases to the preceptor, rather than being directly and independently evaluated. Inferring three attributes from a single source of data encourages determination bias whereby the perception of a particular trait, or competency, is influenced by the perception of the former competencies in a sequence of interpretations, i.e. a 'halo effect' (Thorndike 1920; Iramaneerat & Yudkowsky 2007). In addition to a halo effect our data suggested another source of constant error in ratings of the overall performance – an error of leniency – given the fact that the mean score for overall performance on the ITER was significantly higher than 'at expected level'.

## Study limitations

This study looked at the evaluation tools for a single clerkship within a single centre, which limits both statistical power and generalizability of our results. But, given the idiosyncratic nature of evaluation tools used for different clerkships and different centres, it would be logistically difficult to perform factor analysis on in-house evaluations from more than one clerkship and/or centre. We plan to extend our analyses to see if our results are at least consistent between different clerkships in our centre.

Another limitation relates to the fact that factor analysis is partly qualitative. Consequently, our principal components were inferred from factor loading – it is not possible to 'prove' that the principal components represent the constructs that we have interpreted them as representing. We plan to perform confirmatory factor analysis which, if consistent, could offer further support to our interpretations.

## Conclusions

Despite using 11 evaluation items on three tools, we identified only three prinicipal components of evaluation on the Internal Medicine clerkship rotation: information processing, professionalism and declarative knowledge. Both our MCQ and OSCE appeared to measure declarative knowledge. Our ITER appeared to measure only two attributes – information processing and professionalism – rather than the intended nine separate attributes, due to a combination of overlapping attributes and constant errors, including a halo effect and error of leniency. This finding has important implications for all levels of training, particularly when regulatory bodies expect

us to evaluate multiple roles of students and physicians – typically using an ITER. Perhaps we need new evaluation tools to really measure what they are asking us to measure.

## Notes on contributors

Drs. McLAUGHLIN and CODERRE are Associate Professors of Medicines at the University of Calgary.

Dr. VITALE is a Clinical Assistant Professor of Medicines at the University of Calgary.

Dr. VIOLATO is a Professor in the Department of Community Health Sciences at the University of Calgary.

Dr. WRIGHT is an Associate Professor of Family Medicine at the University of Calgary.

***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

## References

Accreditation Council for Graduate Medical Education. http://www.acgme.org/acwebsite/home (accessed September 2008).

Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. 2003. Assessing personal qualities in medical school admissions. Acad Med 78:313–321.

CanMEDS. 2005. The 2005 CanMEDS Physician Competency Framework. Better Standards. Better Physicians. Better Care. (The Royal College of Physicians and Surgeons of Canada, Ottawa).

Cohen J. 2001. Facing the Future. 2001. President's Address, 112th Annual Meeting of the Association of American Medical Colleges. http://www.aamc.org/newsroom/pressrel/2001/011104a.htm

Daelmans H, van der Hem-Stokroos H, Hoogenboom R, Scherpbier A, Stehouwer C, van der Vleuten C. 2005. Global clinical performance rating, reliability and validity in an undergraduate clerkship. Neth J Med 63:279–284.

Ferguson E, James D, Madeley L. 2002. Factors associated with success in medical school: systematic review of the literature. BMJ 324:952–957.

Iramaneerat C, Yudkowsky R. 2007. Rater errors in a clinical skills assessment of medical students. Eval Health Prof 30:266–283.

Kerlinger FN, Lee HB. 2000. Foundations of Behavioral Research, 4th ed. (Nelson Thomson Learning, Toronto).

Lance CE, Butts MM, Michels LC. 2006. The sources of four commonly reported cutoff criteria: What did they really say? Org Res Method 9:202–220.

Medical School Objectives Writing Group. 1999. Learning objectives for medical student education – guidelines for medical schools: report I of the Medical School Objectives Project. Acad Med 74:13–18.

Straub DW. 1989. Validating Instruments in MIS Research. MIS Quarterly 13:147–166.

Thorndike EL. 1920. A constant error on psychological rating. J Appl Psychol 4:25–29.