Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP)

Linda Prescott-Clements,[1] Cees P M van der Vleuten,[2] Lambert W T Schuwirth,[2] Yvonne Hurst[1] & James S Rennie[1]

OBJECTIVE The drive towards valid and reliable assessment methods for health professions' training is becoming increasingly focused towards authentic models of workplace performance assessment. This study investigates the validity of such a method, longitudinal evaluation of performance (LEP), which has been implemented in the assessment of postgraduate dental trainees in Scotland. Although it is similar in format to the mini-CEX (mini clinical evaluation exercise) and other tools that use global ratings for assessing performance in the workplace, a number of differences exist in the way in which the LEP has been implemented. These include the use of a reference point for evaluators' judgement that represents the standard expected upon *completion* of the training, flexibility, a greater range of cases assessed and the use of frequency scores within feedback to identify trainees' progress over time.

METHODS A range of qualitative and quantitative data were collected and analysed from 2 consecutive cohorts of trainees in Scotland (2002–03 and 2003–04).

RESULTS There is rich evidence supporting the validity, educational impact and feasibility of the LEP. In particular, a great deal of support was given by trainers for the use of a fixed reference point for judgements, despite initial concerns that this might be demotivating to trainees. Trainers were highly positive about this approach and considered it useful in identifying trainees' progress and helping to drive learning.

CONCLUSIONS The LEP has been successful in combining a strong formative approach to continuous assessment with the collection of evidence on performance within the workplace that (alongside other tools within an assessment system) can contribute towards a summative decision regarding competence.

KEYWORDS validation studies [publication type]; education, dental, graduate/*methods; clinical competence/*standards; Scotland; feasibility studies; feedback.

*Medical Education 2008: **42**: 488–495*
doi:10.1111/j.1365-2923.2007.02965.x

INTRODUCTION

The development of assessment tools that are robust, feasible and of educational value remains among the biggest challenges faced by educators within the health professions. The validity of assessment is vital and, as a consequence, there is an increasing drive towards the development of instruments that can be used in the practice setting (i.e. of workplace assessment) to sample across a wide range of contexts and judge.

Much progress has been made with the development of workplace assessment tools recently, with a number of different approaches suggested in the literature. Methods using global ratings for observed performance (especially the mini-CEX [mini clinical evaluation exercise]) seem particularly successful.[1–4] There are, however, minor problems in some instances such as gauging the reference point against which evaluators' judgements measure standards, and use of a limited range of cases.

Longitudinal Evaluation of Performance (LEP) is a similar method that involves the direct observation of patient encounters on a weekly basis, for which global

[1]National Health Service Education for Scotland, Edinburgh, UK
[2]Department of Educational Development and Research, University of Maastricht, Maastricht, The Netherlands

Correspondence: Dr L Prescott-Clements, Educational Projects Manager, NHS Education for Scotland, Floor 5, Thistle House, 91 Haymarket Terrace, Edinburgh EH12 5HE, UK.
Tel: 00 44 131 313 8081; Fax: 00 44 131 313 8001;
E-mail: Linda.Prescott-Clements@nes.scot.nhs.uk

Overview

What is already known on this subject

Workplace assessment is increasingly important in the assessment of competence. The validity of assessment is vital in such circumstances. When used in isolation, psychometric analyses of the construct (e.g. Cronbach's α) may not provide the evidence required to address the broader concepts of validity and educational impact that are so important.

What this study adds

We describe a method of workplace assessment, the longitudinal evaluation of performance (LEP) and the evaluation of its validity. Results indicate that this tool is valid and has a positive educational impact.

Suggestions for further research

Research into the validity of the LEP will continue. An evaluation of the reliability of this tool is currently underway.

ratings are given for broad skill areas with respect to the 'whole task' performed.[5,6] This method, however, adopts a different approach. Ratings are given on a 9-point scale, using a *fixed reference point* for judgements that is the standard expected by the evaluator upon *completion* of the training.[5] Ratings 1–3 represent the category 'Need improvement', 4–6 represent 'Satisfactory' and 7–9 'Superior' performance. Initial concerns that the use of a fixed reference point in this way might demotivate trainees at early stages in training were addressed using a no-penalty approach and extensive training for trainers and trainees (on areas including rating scales, avoiding bias and giving feedback). Being primarily formative, if the LEPs indicate a need for improvement in a certain area there are no lasting consequences for the trainee, other than a requirement for progress to be subsequently demonstrated to an appropriate standard in order to achieve satisfactory completion of training. This approach enables the LEP to be used essentially as a screening assessment that can identify personal strengths and weaknesses, maximising the educational impact of the process and enhancing feedback. It also recognises that different trainees bring different levels of prior knowledge and experience to their training, and consequently may reach the standards required at different rates. Designed to be flexible, the LEP is not used for a prescribed list of cases as this could limit validity in terms of the range of assessment completed. Instead, trainers identify which cases are assessed each week using a comprehensive list of competencies as a guide.[7] Using this approach, appropriate coverage of the curriculum is ensured as the criteria for satisfactory completion of the training must be met, including a requirement that all areas of clinical focus are assessed. Additional sampling restrictions imposed on LEPs ensure the range of LEPs completed during the training include the treatment of different patient age groups and case complexities, and that they use different evaluators. Trainers are able to monitor assessment coverage and sampling using the quantitative and qualitative information provided to them on regular LEP feedback reports.

The LEP has been implemented in Dental Vocational Training (DVT) in Scotland, and the question now arises as to whether this approach is valid within the context of this workplace. Validity is not an easy concept and in an effort to understand it further, research has focused increasingly upon its dissection into constituent spheres, with numerous 'types' of validity being described including content, construct, criterion, concurrent and predictive validity (to name a few). This detailed view of validity being made up of constituent parts has sometimes unfortunately clouded perception and resulted in reports of validity being generalised from the results of a few analyses, such as correlation coefficients with a 'gold standard'.[8] It has been argued recently that we cannot infer validity from a single analysis and that the evaluation of assessment programmes should focus on the larger picture.[9] For example, in isolation the results from correlating assessment data with those from another test or 'gold standard' demonstrate only the degree of similarity (or otherwise) between these measures. In many instances, new assessments are developed to improve on existing methods and in these cases the results from such comparisons may be misleading. If the original method lacked validity, a high degree of correlation may only indicate that the new assessment performs to an equally poor standard. Although a lower correlation may indicate differences between the methods, the direction (i.e. better or worse) can only be determined with supplementary information. In other cases (including this study) an alternative method for comparison may not be available and other analyses will be required to build evidence for validity.

Newer insights[10] try to look at measurement information as the combination of construct-relevant variance and construct-irrelevant variance. Others go even further[11] in that they define validity as fitness for purpose. Traditional construct validity is then only a single small part of the puzzle but, in this view, the way the instrument is used and when, by whom and for which purpose are other equally important areas. For an irrefutable case for validity to be established, evidence originating from many sources should be considered in combination.

The aim of this study was to explore the validity of the LEP from different angles, including the quantitative (construct) and qualitative approaches.

METHODS

A detailed description of the LEP has been published previously.[5] To determine the validity of the LEP, the multi-faceted purpose for which it was designed was considered. In addition to the identification of appropriate (or poor) performance upon completion of training, the LEP was designed to address other important educational requirements.[5,6] The operational definitions of our research question reflected these aims.

- Is the range of assessment covered using the LEP adequate? Does it address all-round competence?
- Does the LEP identify trainees' strengths and weaknesses, and their progress towards the standard of performance required upon completion of this training?
- Does the LEP identify poor performance?
- Does the LEP motivate trainees?
- Is the LEP feasible in practice?
- Is the feedback from the LEP useful?

Two sources of information were used for validity investigations. The first comprised LEP results from 2 consecutive years of DVT trainees (separate cohorts). The second consisted of results from the annual process evaluation questionnaire sent to trainers.

LEP data

Trainees must complete at least 42 LEPs throughout the year to fulfil the requirements for satisfactory completion of DVT. The 2002–03 cohort had 100 trainees and the 2003–04 cohort had 101, resulting in data from almost 10 000 LEPs submitted over the 2 years. Data were entered into a specially designed database (ACCESS 2000; Microsoft Corp., Redmond, WA, USA), and analysed using SPSS (SPSS Inc., Chicago, IL, USA) or EXCEL (Microsoft Corp.).

Evaluation questionnaire for trainers

A comprehensive evaluation questionnaire was sent to trainers towards the end of each DVT year. Trainers were asked to identify their level of agreement with statements using a 5-point scale ('Strongly agree', 'Agree', 'Neither agree nor disagree', 'Disagree' or 'Strongly disagree'). Most trainers were in post for both years of this study and therefore responded to the questionnaire on 2 occasions.

RESULTS

Is an appropriate range of cases assessed using the LEP?

The number of trainees assessed across the full range of 11 clinical foci increased from 67% (2002–03) to 94% (2003–04). In addition, all trainees in the second cohort were assessed across at least 10 areas of clinical focus. Although all participants (both cohorts) were trained in the assessment process and were aware of the criteria for satisfactory completion, the wider range of assessment coverage by the second cohort is most likely a result of trainers becoming familiar with the new assessment programme and being able to manage the assessment process more effectively. Indeed, this high level of curriculum coverage has since been confirmed: in 2004–05, 99% of trainees were assessed across all areas of clinical focus. Results from the trainer evaluation questionnaire support this hypothesis: in 2003–04, 73% of trainers indicated that the types of procedures assessed were carefully monitored to ensure an adequate range of cases, compared with only 53% the previous year (Table 1).

Both cohorts managed to assess a full range of cases of different complexity using the LEP, with 99% of trainees being assessed across encounters considered to be of low, moderate and high complexity.

To ensure an appropriate range of assessment, trainees are required to complete LEPs for the treatment of different patient age groups. A high proportion of trainees were assessed across all patient age groups (children, adults, elderly people) during the training; 93% achieved this goal in 2002–03 and 98% achieved it in 2003–04.

Table 1 Results from the trainers' evaluation questionnaire (both pilot studies)

Question	Response (*cohort 1/cohort 2**)					
	No response	**Strongly disagree**	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
I found the LEP useful	1%/1%	0%/0%	4%/4%	8%/4%	63%/64%	24%/27%
I was uncomfortable observing my trainee's performance	0%/3%	24%/28%	57%/53%	12%/8%	7%/7%	0%/1%
Observation is important in the assessment of vocational skills	0%/0%	0%/0%	0%/1%	1%/0%	58%/50%	41%/49%
Having an external evaluator carry out LEPs in each block was useful	2%/0%	0%/0%	4%/1%	11%/14%	59%/55%	24%/30%
The LEP forms are clear and easy to understand	1%/0%	0%/0%	13%/4%	20%/11%	58%/74%	8%/11%
There is adequate space on LEP forms for feedback and comments	0%/1%	3%/5%	29%/34%	11%/5%	52%/51%	5%/4%
Having a fixed reference point for judgement helps identify progress over time	0%/2%	0%/0%	4%/1%	5%/5%	67%/78%	24%/14%
Having a fixed reference point for judgement helps drive the training	0%/2%	0%/0%	3%/4%	24%/23%	58%/64%	15%/7%
I am comfortable using my judgement against a 9-point scale	0%/2%	1%/0%	4%/2%	7%/11%	74%/76%	14%/9%
I am comfortable with the descriptors 'Need improvement', 'Satisfactory' and 'Superior'	0%/2%	3%/2%	5%/5%	12%/8%	64%/71%	16%/12%
My trainee welcomes my feedback	3%/2%	0%/0%	4%/1%	14%/11%	70%/73%	9%/13%
LEPs had a negative effect on the working relationship I had with my trainee	3%/2%	31%/26%	51%/57%	9%/9%	3%/5%	3%/1%
LEPs were useful in identifying topics for discussion in tutorials	3%/2%	1%/0%	5%/1%	14%/14%	63%/63%	14%/20%
LEP feedback reports were useful	0%/2%	1%/4%	11%/20%	27%/19%	52%/49%	9%/6%
LEPs were primarily driven by my trainee	7%/1%	14%/12%	48%/53%	24%/25%	7%/9%	0%/0%
The types of procedures were carefully monitored to ensure an adequate range of cases were covered	9%/1%	1%/2%	11%/5%	26%/19%	50%/65%	3%/8%
I feel reluctant to give ratings in the 'Need improvement' range	3%/1%	14%/14%	52%/45%	11%/26%	19%/13%	1%/1%
My trainee found 'Need improvement' ratings de-motivating	5%/1%	7%/3%	38%/41%	32%/35%	18%/13%	0%/7%

* Cohort 1 = 102 trainers in dental vocational training [DVT] 2002–2003; Cohort 2 = 98 DVT trainers, 2003–2004
LEP = longitudinal evaluation of performance

Does the LEP identify trainees' strengths and weaknesses, and their progress towards the standard of performance required upon completion of this training? Does the LEP identify poor performance? Does the LEP motivate trainees?

When the average LEP scores (across all trainees in 2003–04) for each month of training are plotted on a graph (for all 8 skill categories), a curve with a positive gradient is consistently shown in each case (Fig. 1). The increase in ratings throughout the year represents the progress made during training. Even more interestingly, the results display a typical learning curve, with a steeper gradient (representing a larger increase in ratings) apparent in the first half of training. This difference can be seen more clearly when the mean *difference* in ratings is calculated and compared between the first half of training (training block 2 scores minus block 1 scores) and the second half of training (block 3 scores minus block 2 scores). Approximately twice as much progress is made during the first half of training, which is perhaps consistent with DVT, where individuals are qualified but have yet to develop their skills in a general practice environment. It is also interesting that a relative (albeit small) decrease in mean ratings for most skill categories was observed at the end of the training year. This was because a larger proportion of the total LEPs submitted at this time represented re-assessment of cases that had previously been awarded 'Need improvement' ratings earlier in the year. Satisfactory completion of this training cannot be awarded unless all the areas identified as requiring improvement have been addressed. The submission of such 'repeat-case' LEPs in the final stages of the training, with generally lower (although still satisfactory) mean ratings, also supports the validity of the tool, indicating correct use of the rating scale and reference point for judgement.

Despite initial concerns, a great deal of support was given for the use of a fixed reference point for judgement (Table 1). In 2003–04, 71% of trainers agreed that this helped drive the training and 92% thought that it helped in identifying progress over time. Most trainers were comfortable using the 9-point scale (85%) and the descriptors used (83%). In addition, only 6% of trainers thought that the use of LEPs had a negative effect on their working relationships with their trainees. Indeed, 26% of trainers strongly disagreed with this statement. Only 14% of trainers were reluctant to give ratings in the 'Need improvement' range, but 20% thought that their trainees found lower ratings demotivating.

In 2003–04, 74% of trainees were given ratings in the 'Need improvement' range, the majority of which were within the first training block. The high proportion of 'Satisfactory' LEPs during this initial phase also suggests that trainers were using LEPs (and the rating scale) correctly and were crediting good performance in addition to highlighting areas in which further progress was required. Of the trainees who received ratings indicating a need for improvement, the vast majority (90%) were successful in repeating LEPs for relevant clinical encounter(s) that demonstrated progress (i.e. ratings ≥ 4) in these areas. Trainees with performance issues outstanding

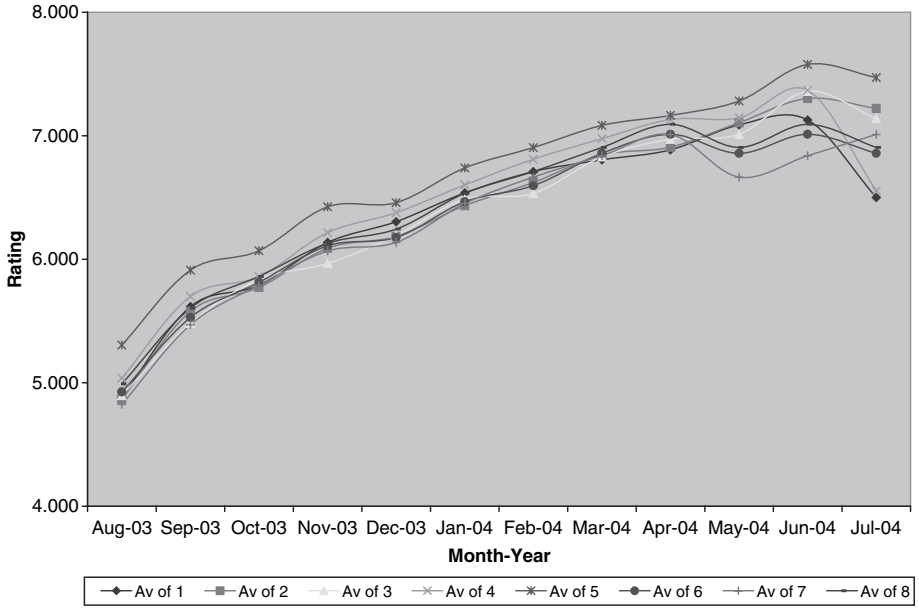


Figure 1 Mean longitudinal evaluation of performance ratings given for each skill category by month during dental vocational training in Scotland, August 2003 to July 2004. Av = average

had their certificates of satisfactory completion withheld pending further assessment evidence.

Is the assessment feasible in practise?

On average, 22 minutes were spent on observing performance and 8 minutes on giving feedback to the trainee. The dentists training grant has been increased to support the extra time spent on assessment. Further evidence supporting the feasibility of the LEP (Table 1) includes the finding that 65% of trainers found the assessment system easy to implement (only 11% of trainers disagreed) and 63% thought the assessment had sufficient flexibility.

Is the feedback from the assessment useful?

Most LEP forms include written feedback from the evaluator. The evaluation questionnaire revealed that 32% of trainers in 2002–03 considered there to be too little space on the LEP form for written feedback. Consequently, this space on the LEP form was almost doubled in size for the following cohort. However, results from this evaluation exercise (2003–04) revealed that a similar proportion (39%) of trainers wanted yet more space for feedback (Table 1, and see Fig. 2).

In general, trainees welcomed feedback from their evaluators during the LEP process. A total of 79% of trainers considered there had been a positive response from their trainees in 2002–03, increasing to 86% in 2003–04. Only 1 trainer in the latter cohort thought that his or her trainee did not welcome feedback. A similar response was given by trainers when asked if they considered LEPs useful in identifying topics for discussion in tutorials, with the majority of trainers in agreement (77% in the first cohort and 83% in the second).

Formal written feedback is provided to individual trainers and trainees at the end of each training block, summarising their progress made to date and highlighting any particular strengths or weaknesses. Although the majority of trainers agreed that these reports were useful (61% in 2002–03, falling to 55% in 2003–04), this was less prominent than for immediate feedback, as indicated above. Many trainers were indifferent about the usefulness of these reports and this may be explained in part by the fact that they represent a summary of information that is already familiar to both trainer and trainee (copies of all LEPs carried out are kept in the training practice for reference purposes). Of particular concern are the 24% of trainers in 2003–04 who disagreed with the statement: 'LEP feedback reports were useful.' However, the reasons indicated were mostly associated with the format of the reports rather than their content, and ways of improving these reports will be investigated as a result of this information.

DISCUSSION

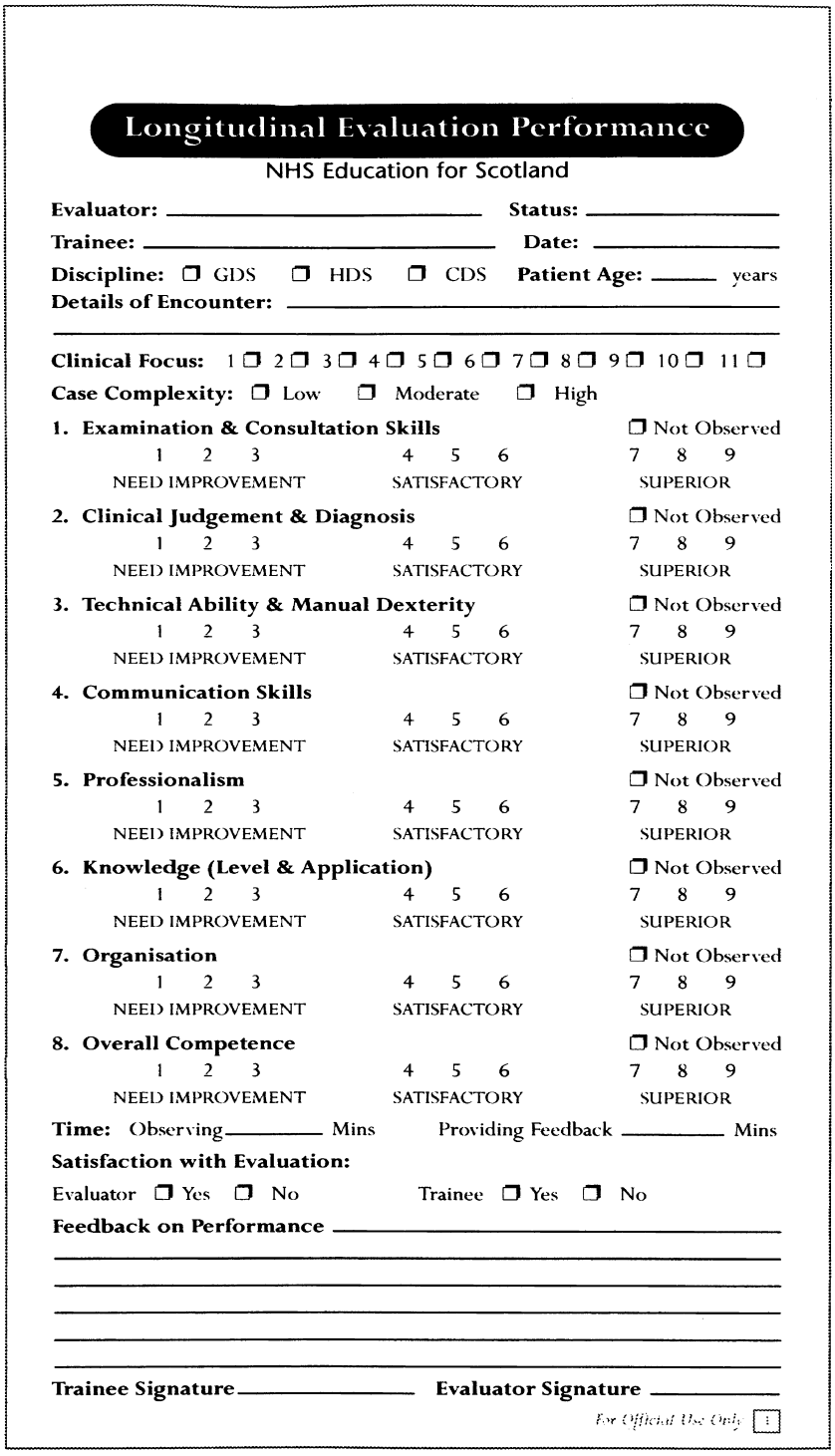
This study involves an investigation into the validity of the LEP, a workplace assessment implemented in DVT in Scotland.

Similar methods of workplace assessment, such as the mini-CEX, have been shown to be highly authentic, valid and reliable for the assessment of doctors in hospital training.[1–4] Although both the mini-CEX and LEP use direct observation of 'real-life' performance and give global ratings across broad areas of competence against a 9-point scale, the differences in implementation highlighted within this paper and in the clinical setting mean that we cannot automatically assume that the 2 instruments have similar psychometric properties.

In acknowledgement of the complexity of the concept of validity, we took a holistic approach to this study and used a wide range of quantitative and qualitative data to build a case. A further challenge was that this assessment is the first to be used in DVT and therefore no 'gold standard' exists with which the validity of the LEP (context-specific) can be compared.

A wealth of evidence has been gathered which supports the validity of the LEP for the assessment of DVT in Scotland. This method has been used effectively in assessing a broad range of competencies relevant to this training, while not being over-prescriptive. The flexibility of this method is generally appreciated by trainers who, as a result, are able to focus assessment specifically to the individual needs of their trainee. The wide range of skills assessed using the LEP ensures that assessment evidence is reflective of the all-round competence of the practitioner.

Using a fixed-reference point for evaluators' judgements, the LEP successfully demonstrates progress over time, indicating for the first time a true 'learning curve' that is pertinent to this training. Not unexpectedly, the differences between ratings awarded to individuals at the beginning, middle and end-points of training (Fig. 1) are much greater than the observed growth demonstrated with the mini-CEX,



Longitudinal Evaluation Performance

NHS Education for Scotland

Evaluator: _____ **Status:** _____

Trainee: _____ **Date:** _____

Discipline: ❒ GDS ❒ HDS ❒ CDS **Patient Age:** _____ years

Details of Encounter: _____

Clinical Focus: 1❒ 2❒ 3❒ 4❒ 5❒ 6❒ 7❒ 8❒ 9❒ 10❒ 11❒

Case Complexity: ❒ Low ❒ Moderate ❒ High

1. Examination & Consultation Skills ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

2. Clinical Judgement & Diagnosis ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

3. Technical Ability & Manual Dexterity ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

4. Communication Skills ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

5. Professionalism ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

6. Knowledge (Level & Application) ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

7. Organisation ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

8. Overall Competence ❒ Not Observed

1 2 3 4 5 6 7 8 9

NEED IMPROVEMENT SATISFACTORY SUPERIOR

Time: Observing_____ Mins Providing Feedback _____ Mins

Satisfaction with Evaluation:

Evaluator ❒ Yes ❒ No Trainee ❒ Yes ❒ No

Feedback on Performance _____

Trainee Signature_____ **Evaluator Signature** _____

For Official Use Only [ ]

Figure 2 Longitudinal evaluation of performance (LEP) form

which does not use the end of training as a reference.[2] As progress can be visualised as ratings increase with time, it can be argued that the educational impact of the LEP on both trainee and trainer is increased and this may also have a positive effect on the motivation of the individual towards improving his or her own performance. Progressive data such as that observed when using a fixed reference point may be considered to highlight the importance of formative assessment within the training cycle (i.e. between assessment, feedback and subsequent improved performance).

A potential limitation of continuous assessments that use rating scales, and particularly those with a fixed reference point for judgement, is that raters will

anticipate that trainees' scores will increase with time and this will positively bias ratings awarded later in the training. Although detailed studies into sources of bias (including the effects of case complexity) would be required to rule this out, it is encouraging that mean scores are shown to decrease in the final stages of this training when LEPs have been submitted for cases where poor performance had previously been identified. Although the scores for such cases demonstrate satisfactory performance, these ratings are clearly lower than others submitted at this stage of training as a negative effect on the mean category scores is observed (Fig. 1).

In addition, the evidence from this study suggests that the focus on a formative approach through maximum feedback and no penalty for lower LEP ratings has enabled trainers to feel more comfortable about awarding ratings which are appropriate to performance, where lower scores are given when required and good performance is rewarded as such. By minimising the consequences of assessment outcomes, it is likely that problems such as leniency and the halo effect have also been reduced. In addition, any areas identified as falling into the 'Need improvement' category must be addressed, the relevant skills improved and the trainee re-assessed to a satisfactory standard before he or she can complete this training and be awarded the NHS list number (licence) that will allow him or her to treat patients independently. As the frequency of each rating is used rather than an average score, poor performance in one area cannot be compensated for in another area, representing a truly competent practitioner.

Finally, we have good evidence that the LEP is feasible within this setting as it takes an average of 30 minutes per encounter (including feedback) to complete.

Contributors: all authors contributed to the discussions leading to this paper. LP-C, LWTS and CPMvdV undertook the principle responsibility for preparing the draft, co-ordinating input from other authors and writing the final version of the paper.
Acknowledgements: none.
Funding: this study was funded by NHS Education for Scotland.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;**123**:795–9.
2 Norcini JJ, Blank LL, Duffy D, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003;**138**:476–81.
3 Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medical residency training. *Acad Med* 2002;**77**:900–4.
4 Holmboe ES, Huot S, Chung J, Norcini JJ, Hawkins RE. Construct validity of the mini clinical evaluation exercise (mini-CEX). *Acad Med* 2003;**78**:826–30.
5 Prescott LE, Norcini JJ, McKinlay P, Rennie JS. Facing the challenges of competency-based assessment of postgraduate dental training: longitudinal evaluation of performance (LEP). *Med Educ* 2002;**36** (1):92–7.
6 Prescott-Clements LE, Hurst Y, Rennie JS. Satisfactory completion of dental vocational training in Scotland: a system of assessment. *Br Dent J* 2003;**Sept** (Suppl):17–21.
7 Prescott LE, McKinlay P, Rennie JS. Comprehensive validation of competencies for dental vocational training and general professional training. *Eur J Dent Educ* 2003;**7**:154–9.
8 Norman G, Swanson D, Case S. Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teach Learn Med* 1996;**8** (4):208–16.
9 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006;**40**:296–300.
10 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educ Res* 1995;**23**:13–23.
11 Guba EG, Lincoln YS. (2001). Guidelines and checklist for constructivist (a.k.a. fourth generation) evaluation. http://www.wmich.edu/evalctr/checklists/checklistmenu.htm. [Accessed 17 August 2006.]

Received 29 May 2006; editorial comments to authors 23 January 2007, 25 July 2007; accepted for publication 19 September 2007