education & training

Psychiatric Bulletin (2008), **32**, 271–273. doi: 10.1192/pb.bp.107.016576

GEOFFREY F. SEARLE

Is CEX good for psychiatry? An evaluation of workplace-based assessment

Workplace-based assessment is a core element of the changes introduced in Modernising Medical Careers (MMC), yet it has been sparsely researched in the UK and there are still psychometric and validity data needed to support its introduction. Supervisors and tutors are wary of it, but it could replace external clinical examinations leading to a Certificate of Completion of Training. The theoretical background of assessment in the workplace is 'Miller's pyramid' of competence in which the trainee progresses from 'knowing' (tested by paper exam), through 'knows how', to 'shows how' (tested by the Observed Standardised Clinical Examination), and finally 'does' (assessed in the workplace) (Miller, 1990).

The Clinical Evaluation Exercise (CEX) was developed in the 1960s in the USA to assess doctors in postgraduate training, and in the early 1970s it replaced the oral examination required for certification. The procedures are quite straightforward – a senior staff member observes the trainee take a history and clinically examine a real patient and assesses his or her performance immediately afterwards. There are many minor variations of the format, but all have numerical (Likert) scales under several headings. In the study discussed below the assessed competencies were: history taking, investigations, record keeping, problem solving/diagnosis, emergency care, attitude to/relationship with patient, team working, learning/teaching, and responsible attitude. The scale scores' descriptors had the following ratings: 1–2 'unsatisfactory', 3–7 'satisfactory', and 8–9 'superior'. The outcome of the assessment is discussed with the trainee immediately, thus providing a formative feedback; the whole process takes about 60 minutes.

This clinical evaluation exercise (modified and renamed the Assessment of Clinical Expertise, ACE) is one of the several tools being used for the assessment of psychiatrists in training as part of MMC. In a competency-based system each judgement must be verified by several senior staff members using more than one assessment tool to securely validate the achievement of the competency. The tools must be easy to use, reliable, robust and valid. A previous study of performance-based tests conducted in the USA (Thompson *et al*, 1990) found an excess of high ratings (95.6% being from 6 to 9), considerable correlation between items (r range from 0.72 to 0.92) and poor interrater reliability (r=0.64, range 0.16–0.88). Kroboth *et al* (1992) concluded that six to ten CEX observations or a whole day of clinical testing would be required to achieve adequate reliability.

Care must be exercised concerning interrater reliability. Oral clinical examinations have poor interrater reliability coefficients (0.45 for 1 hour of testing) (Swanson, 1987). Analysis of the interrater reliability of Part 2 MRCPsych, examiners showed kappa scores of 0.4–0.5 (Oyebode *et al*, 2007). Consequently, when comparing any method against a clinical examination with a viva voce, high correlations are unachievable because of the weakness of the standard. A coefficient above 0.8 is expected for a 'gold standard' assessment.

In spring 2003 the Department of Health introduced a Record of In-Training Assessment (RITA) for senior house officers. This study on the workplace-based assessment was undertaken across the Solent rotation training scheme by adopting the CEX to familiarise trainees and trainers with this type of assessment and provide a record of in-training assessment. This allowed the CEX to be tested and validated against the MRCPsych examination results.

Method

The CEX was modified for psychiatry. No 'pass' or 'fail' scores were set, although they are implied on the instrument. Formal training was not attempted but supervisors/raters were offered tutorials.

Each of the 72 trainees in the Solent rotation arranged quarterly assessments with their educational supervisor in the second and the final month of their attachments. Detailed implementation varied with the administrative back-up available. In one area an assistant sent out forms and issued reminder letters. Elsewhere trainees arranged assessments themselves, the forms being checked, but not collected, by college tutors. Exam results were collated centrally. The second-month assessment occurred in March and September before exam results were known, so the assessor was not influenced by their outcomes (Box 1).

education & training

Box 1. Assessment statistics

Reliability – describes the proportion of the variance between measurements which is caused by a 'true' difference between individuals. Good reliability is essential before examining validity (see below). Intraclass correlation calculates the reliability coefficient of multiple measurements of the same variable – the closer to 1 the better. The Pearson correlation coefficient is between different variables (an interclass correlation) – values around 0.5 imply separate, but related measurements. Cronbach's α describes the homogeneity of the scale items and should be between 0.7 and 0.9.

Validity – describes how accurately a scale measures what it is intended to measure. Its value is always less than the reliability value (never more than the square root of the reliability coefficient).

Streiner & Norman, 2003

Results

Reliability

The scale had a mean item score of 7.08, n=1791 item scores (Mode 8, range 2–9, mean item s.d.=1.15, Cronbach's α=0.92, n=9). The Pearson correlation coefficients between the items varied between 0.44–0.74.

To examine interrater reliability a paired set of ratings were selected, with the second assessment in one attachment compared with the first assessment in the next. The two mean ratings were not highly correlated. There was a highly significant difference in the means with the rating at the end of the previous post higher than the subsequent one (mean difference of mean scores 0.62, s.e=0.095, n=95, $P<0.0001$). The intraclass correlation coefficient was 0.40 (for single measure 95% CI 0.22–0.56, n=95).

On examination of individual subscale scores, 73% of ratings were within one of each other (range 64%–86%, n=95) with all showing the same highly significant difference between the late and early scores as the mean score. The average intraclass correlation was 0.25 (range 0.14–0.36). Each senior house officer was included only once in the interrater and validity analyses to ensure each pair of values were independent and to minimise the risk of bias.

Validity

When the relation of a CEX to exam results is analysed (by selecting assessments made during an exam attempt) there is a statistically significant relationship between mean scale scores and exam success for Part 1, but not Part 2 of the MRCPsych examination (Table 1).

Although the difference in mean scores is very similar (Part 1, 0.62; Part 2, 0.70) the smaller number of cases and greater variance for the Part 2 scores reduces the significance of the results.

Using the methodology of foundation year workplace-based assessments, demanding that all subscale scores are above a set 'pass' score yields similar results. With the pass mark lying between 6 (too many false passes) and 7 (too few true passes) for Part 1 and 6 for Part 2.

Table 1. MRCPsych pass/fail compared with CEX score

Part 1 MRCPsych Mean score (s.e.) t=2.15, d.f.=32, P=0.04	Pass (n=18) 7.29 (0.19)	Fail (n=16) 6.67 (0.22)
Part 2 MRCPsych Mean score (s.e.) t=1.57, d.f.=20, P=0.15	Pass (n=14) 7.65 (0.17)	Fail (n=8) 6.95 (0.42)

Practicality

There were 263 forms returned by 112 senior house officers who were assessed by 78 supervisors, representing an overall return rate of 61% (753 out of 1225, range 22%–94%). Removing the only centre with administrative support (which returned 94%) reduces the rate to 40% (280 out of 695, range 22%–55%).

Discussion

The ranges of intraclass correlations in the study by Kroboth *et al* (1992; range 0.23–0.50) and in this study (0.14–0.37) are similar, with the weaker correlations perhaps being the result of the absence of structured rater training. Thompson *et al* (1990) found a high mean item rating of 7.3–7.6 (7.1 in this study) with correlation between items implying that a single underlying component was being rated which was not supported by this work. This suggests that the assessment is not needlessly complex.

The interrater reliability was weak, especially when individual subscales are considered. Earlier work supports the finding of inadequate interrater reliability. There was no structured and validated education of raters, which would have been ideal. Thompson *et al* (1990) did not mention rater training and Kroboth *et al* (1992) issued a manual and held a single meeting for raters. The assessments were performed by educational supervisors (consultants), whereas MMC will use many other members of the clinical team. The significant difference between assessments early in the post and later implies an important acquaintance effect with trainees who are well known to the rater being scored higher.

This naturalistic study demonstrates that the CEX is a practical method of workplace-based assessment which shows significant statistical association between aggregated scores and examination results.

The rate of return of the scales was highly dependant upon the availability of administrative support. The relatively low rate of return introduces the possibility of bias in those ratings received (trainees handing in only 'good' assessments).

The greatest problem is how many assessments are going to be required before adequate reliability is achieved for a 'high stakes' pass or fail decision. The poor

https://doi.org/10.1192/pb.bp.107.016576 Published online by Cambridge University Press

interrater reliability of the individual items shown in this study implies that far more than the planned three assessments will be required to produce a robust assessment, yet achieving a pass score at these assessments will be critical for the progression of a trainee onto the next stage of training.

The problem of inadequate reliability is made more difficult when integrating competency-based curricula with workplace-based assessments which are not individually tailored to those competencies. The original model from the USA is that the CEX assesses clinical skills inaccessible to written testing using nine elements with their competence descriptors integral to the scale. Similarly in the foundation years' assessments each tool focused on a particular competency. However, the individual competencies of the psychiatry curriculum cannot be explicitly contained within the necessarily limited number of workplace assessment scales and thus their achievement has to be recorded separately. The first year of psychiatry training has approximately 40 individual competencies to be signed off and it is expected that this will be done on at least three separate occasions by three different assessors for each competency. This is potentially an enormous burden upon the trainee, supervisors and administrators. Creating, validating and administering a myriad of different assessments would require enormous resources but without them it may be impossible to clearly and robustly record the achievement of a competence. The College Assessment of Clinical Expertise has been modified to use a six-point scale and examine only the elements of the CEX that can be shown during the assessment – the original implicitly included knowledge of the candidate's performance outside the assessment.

There is an urgent need for the validation of all the tools used for assessment under MMC to demonstrate what is an adequate number of assessments and to show they have adequate psychometric properties. The most important issue of linking scores on workplace-based assessments and the achievement of competencies must be rapidly resolved. As yet there is little information on the effect of using assessors who are not senior medical clinicians to assess experienced postgraduate trainees for a wide variety of competencies and this must be researched. Containing and organising the potential explosion of educational administration will be an enormous challenge. Research into the cost of this exercise is essential as the diversion of resources into a flawed labyrinthine assessment and validation process will add considerably to the pressure upon the already scant resources.

Declaration of interest

None.

Acknowledgements

Thanks to Peter Taylor for helping with data analysis, and Deborah Hutchinson, Ray Vieweg, Jim Ormsby, Bruce Allen, Marianne Gemmeke, Ian Ellison-Wright, Nicola Scammell and Yvonne Remnant for helping with data collection and collation.

References

KROBOTH, F. J., HANUSA, B. H., PARKER, S., *et al* (1992) The inter-rater reliability and internal consistency of a Clinical Evaluation Exercise. *Journal of General Internal Medicine*, **7**, 174–179.

MILLER, G. (1990) The assessment of clinical skills, competence, performance. *Academic Medicine*, **65**, S63–S67.

OYEBODE, F., GEORGE, S., MATH, V., *et al* (2007) Inter-examiner reliability of the clinical parts of MRCPsych part II examinations. *Psychiatric Bulletin*, **31**, 342–344.

STREINER, D. L. & NORMAN, G. R. (2003) *Health Measurement Scales: A Practical Guide to Their Development and Use* (3rd edn). Oxford University Press.

SWANSON, D. (1987) A measurement framework for performance based tests. In: *Further Developments in Assessing Clinical Competence* (ed I. Hart & R. M. Harden), pp. 13–36. Can-Heal Publications.

THOMPSON, W. G., LIPKIN, M. JR, GILBERT, D. A., *et al* (1990) Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluator Form. *Journal of General Internal Medicine*, **5**, 214–217.

Suggested reading

DAVIES, H. (2005) Work based assessment. *BMJ Career Focus*, **331**, 88–89.

Geoffrey F. Searle Consultant Psychiatrist, Crisis Home Treatment Team, Hahnemann House, Hahnemann Road, Bournemouth BH2 5JW, email: geoff.searle@dhft.nhs.uk

https://doi.org/10.1192/pb.bp.107.016576 Published online by Cambridge University Press