BJPsych The British Journal of Psychiatry (2008) 193, 60–64. doi: 10.1192/bjp.bp.107.038588

Assessing competence in cognitive–behavioural therapy

Andrew J. A. Keen and Mark H. Freeston

Background
Postgraduate courses on cognitive–behavioural therapy (CBT) assess various competencies using essays, case studies and audiotapes or videotapes of clinical work.

Aims
To evaluate how reliably a well-established postgraduate course assesses CBT competencies.

Method
Data were collected on two cohorts of trainees (n=52). Two examiners marked trainees on: (a) two videotapes of clinical practice; (b) two case studies; and (c) three essays.

Results
Essay examinations were more reliable than case studies, which in turn were more reliable than videotaped assessments. The reliability of the latter two assessments was considerably lower than that commonly expected of high-stakes examinations. To assess reliably standard CBT competencies, postgraduate courses would need to examine about 5 essays, 12 case studies and 19 videotapes.

Conclusions
Reliable assessment of standard competencies is complex and resource intensive. There would need to be a marked increase in the number of samples of clinical work assessed to be able to make reliable judgements about proficiency.

Declaration of interest
M.F. is Director of the Diploma in Cognitive Therapy, whose examinations are detailed herein.

The demand for psychological therapies, particularly cognitive–behavioural therapy (CBT), has increased rapidly in recent years under the impetus of the National Service Framework for Mental Health[1] and the emergence of clinical guidelines from the National Institute for Health and Clinical Excellence (NICE) such as those for schizophrenia and depression,[2] and there are indications that further expansion is required.[3] Although the government has laid out guidelines on the organisation and delivery of psychological therapies,[4] the issue of competency remains largely in the hands of core professional bodies (such as the British Psychological Society and the Royal College of Psychiatrists) and organisations specific therapeutic approaches. The main organisation in the UK for CBT is the British Association of Behavioural and Cognitive Psychotherapy, which specifies minimum training standards for those wishing to become accredited cognitive–behavioural psychotherapists.[5] Typically, this includes successful completion of one of approximately 20 postgraduate courses followed by a further period of supervised clinical practice. Thus, postgraduate courses are important gatekeepers in the accreditation process. As far as we are aware, no systematic research has been conducted on attempts by educational bodies to measure competencies in CBT. The aim of this article is to examine the ability of a relatively typical, well-established postgraduate CBT course to evaluate reliably trainees' standard CBT competencies.

Method

Background

The Postgraduate Diploma in Cognitive Therapy is delivered by the Newcastle Cognitive and Behavioural Therapies Centre, and validated by the University of Durham. It is one of the longest-established courses in the UK and has been running since 1991. The course consists of a 5-day induction and 35 days of training on a day-release basis over a 10-month period. The training consists of 2 h of clinical supervision (in pairs) and 3.5 h of lectures/workshops per week. During the course, trainees are expected to see patients with a range of mental health problems.

Participants

Participants were trainees on the Newcastle Postgraduate Diploma in Cognitive Therapy course during the academic years 2002/2003 and 2003/2004. All 52 trainees admitted to the course during this time were included.

Assessment procedure

Examinations

Trainees are required to submit seven separate pieces of work. These consist of two 60 min videotapes of CBT sessions with real patients, two case studies (maximum of 4000 words), and three essays (two: maximum of 4000 words; one: maximum of 8000 words). Written work is anonymised; videotapes are not.

The three different methods of assessment are hypothesised to evaluate important though somewhat different competencies, namely trainees' ability to: (a) conduct cognitive therapy with patients; (b) select and present relevant patient material, conceptualise and write a case formulation, and report and reflect on the course of therapy; and (c) review CBT literature, critically evaluate it and link it to their clinical experience. The course design means that multiple attempts are made to sample these underlying competencies.

Trainees must score 50% on all assessments to complete the course successfully. It is therefore at this level that course examiners make pass/fail decisions. Any failed assignment can be resubmitted once.

Examiners

Two examiners mark all pieces of work independently. The assessed work assigned to examiners to ensure that pairs of markers repeat at a low frequency and that a range of examiners will mark the trainees work over the year. Supervisors do not mark their own supervisees work. There is a pool of about 15 examiners associated with the Newcastle cognitive therapy course. These examiners are all National Health Service employees, are

experienced cognitive therapists, and most have significant experience in core professional and/or postgraduate training and assessment.

When scores are close, an agreed band and mark is awarded, normally the mean. On those occasions where two examiners disagree to a significant degree (e.g. pass *v*. fail, or other banding discrepancy which cannot be resolved by discussion), a third examiner will conduct an assessment.

Quality assurance

Those marking videotapes participate in an annual half-day training exercise for benchmarking. Assessors meet three times a year at the internal board of examiners meeting. The University of Durham oversees annual quality assurance procedures and conducts periodic reviews.

Measures

Videotapes

Examiners evaluate trainees' performances on the two videotapes using the Revised Cognitive Therapy Scale (CTS–R).[6] This scale consists of 12 items all of which reflect one important competence in CBT. The items are: agenda setting and adherence; feedback; collaboration; pacing and efficient use of time; interpersonal effectiveness; eliciting of appropriate emotional expression; eliciting key cognitions; eliciting behaviours; guided discovery; conceptual integration; application of change methods; and homework setting. The CTS–R is widely used by postgraduate training courses in CBT.

Examiners rate all items using a 7-point continuous adjectival scale (0=incompetent, i.e. non-compliance with that aspect of therapy, through to 6=expert, i.e. compliance and very high skill in the face of difficulties). A total mark is calculated and, thus, trainees receive a score out of 72 for both tapes. There is a 35-page manual to assist examiners using the CTS–R (available from the authors on request). Traditionally, both tapes required a score of 36. However, owing to the scale used in the CTS–R (incompetent to expert) and in consultation with the university and external members, from 2003 to 2004, the pass mark appropriate for the range of expertise expected on the course was set at 33 for the first tape and at 36 for the second.

Written assignments

The essays and case studies are marked against separate grids that each have four content dimensions. The four content dimensions for essay assessments are: (a) accessing literature; (b) interpreting literature; (c) integration with clinical practice; and (d) discussion and original thought. For case study examinations, the content dimensions are: (a) identifying information and presenting problem(s); (b) case formulation; (c) course of therapy; and (d) outcome discussion/review. Trainees receive a final percentage mark for submitted essays and case studies.

Statistical analysis

In order to reflect the efforts of the Newcastle course to maximise reliability, on those occasions that a third examiner was asked to provide a mark for a piece of work, we used the two scores that were closest. Descriptive statistics for all seven examinations were calculated. Matched-pairs *t*-tests were used to investigate possible differences between trainees' performances on assessments within examination type (i.e. between tapes 1 and 2; between case studies 1 and 2; between essays 1 and 2; between essays 1 and 3, and between essays 2 and 3). Cohen's *d* effect sizes were calculated when differences were significant.

Generalisability theory

Generalisability theory was used to determine the magnitude of the sources of variation in, and the reliability of, all examinations (this is termed a G-study). Generalisability theory was also used to calculate the most effective ways of reducing measurement error and thereby achieving acceptable levels of generalisability (this is called a D-study). Generalisability theory (rather than classic test theory) was used because it is eminently suitable in those situations wherein the goal is to identify and quantify errors of measurement and in doing so find effective methods for reducing their influence.[7,8] A generalisability coefficient of 0.8 or more is generally accepted as indicating adequate reliability,[9] although some have suggested a coefficient of at least 0.9 in very high-stakes examinations, such as those relating to professional licensure.[10] Generalisability analyses were conducted on trainees' total scores on the CTS–R, case studies and essays because it was at this level that examiners made pass/fail decisions.

Separate analyses were conducted on all seven discrete examinations to ascertain the reliability of individual examinations. Also, separate analyses were conducted at the level of assumed competencies. Thus, the reliability of the three assessment types (videotapes, case studies and essays) was calculated by treating the two videotapes, two case studies and three essay assessments as repeated attempts to measure the same three attributes. There were, therefore, three objects of measurement. The object of measurement on videotape examinations was the overall ability of trainees to conduct CBT (as defined by the CTS–R). The object of measurement on the case study assessments was trainees' ability to conceptualise and report on cognitive therapy casework (as defined by four dimensions on the case study marking grid), and the object of measurement in essay evaluations was the ability of trainees to critically review the CBT literature and link it to practice. All other factors described were considered sources of measurement error (Fig. 1 and 2).

Figure 1 illustrates the study design when generalisability analysis was conducted on all examinations, individually. In this instance, the Newcastle examination procedure meant that the design of the study was equivalent to a two-factor nested ANOVA: examiners were nested within trainees. Thus, we were able to calculate the magnitude of two sources of variation. These were (a) the object of measurement and (b) residual error. Residual error includes all factors not identified within the study design and random error. Figure 2 indicates the study design when generalisability analysis was conducted at the level of assumed competencies. On this occasion, the design of the study was equivalent to a three-factor nested ANOVA. Trainees were crossed with examinations (e.g. all trainees were examined on all occasions), and examiners were nested within both trainees and examinations (e.g. different examiner pairs marked trainees' two assessment videotapes). Consequently, we were able to calculate the magnitude of several different sources of error, as well as the three objects of measurement. First, we were able to assess measurement error due to differences in trainees' performances from one occasion to another within assessment type (i.e. between scores on videotape 1 and videotape 2; between scores on case study 1 and case study 2, and between scores on essay 1, essay 2 and essay 3). This reflects the effect of any instability in trainees' performances because of a general trend towards higher or lower scores across examinations, possibly because of practice, training or some other factor. In addition, we could calculate errors of measurement that reflected the fact that some trainees scored relatively high on one examination and relatively low on another

https://doi.org/10.1192/bjp.bp.107.038588 Published online by Cambridge University Press

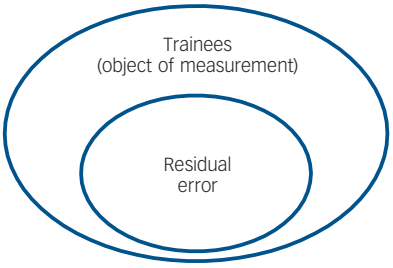

Fig. 1 Sources of variation in the seven individual examinations.

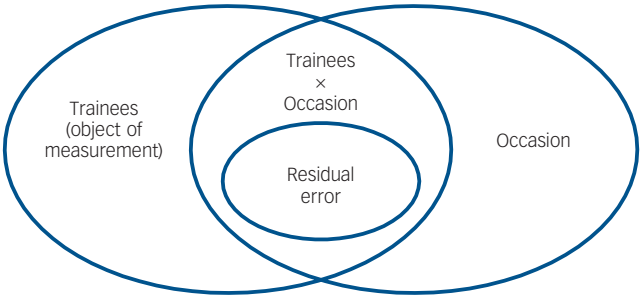

Fig. 2 Sources of variation when combining examinations within assessment type.

(within examination type), whereas the reverse was true for other trainees (e.g. some trainees may have scored relatively high on videotape 1 yet relatively low on videotape 2, whereas others scored relatively low on videotape 1 and relatively high on videotape 2). Finally, we were able to calculate the residual error, which again consisted of factors not included in the design and random error.

On all occasions, absolute rather than relative generalisability coefficients were calculated. That is, our coefficients reflect the reliability of decisions applied to the absolute level of trainees' scores rather than the relative standing of trainees among their peers.

Standard errors of measurement (s.e.m.) at 95% confidence intervals were calculated using generalisability coefficients. The impact of s.e.m. values on examiners' marks was also ascertained.

Results

All 52 trainees across the two cohorts were examined on both videotape 1 and videotape 2. Of these trainees, 51 (98%) submitted case study 1, 50 (96%) submitted case study 2, all submitted essay 1, 49 (94%) submitted essay 2, and 48 (92%) submitted essay 3. Overall, 15 out of 104 (14%) videotapes received a mark from a third examiner. On all occasions, this replaced one of the scores assigned by the original pair of examiners. In addition, 15 out of 101 (15%) case studies received a third mark and similarly on all occasions an original examiner's score was substituted by the third mark. Of the 149 essays submitted by trainees, a third examiner assessed 25 (17%); on 24 out of 25 (96%) occasions the third examiner's marks replaced an original mark.

Scores obtained and changes over time

The means, standard deviations, minimum and maximum scores for all examinations are illustrated in Table 1. Trainees scored significantly higher on videotape 2 than on videotape 1 ($t(51)=2.964$, $P=0.005$); the effect size was small ($d=0.412$). There were no significant differences between case study scores. Trainees scored higher on essay 3 than on essay 2 ($t(46)=2.633$, $P=0.011$) and significantly higher on essay 2 than on essay 1 ($t(48)=2.213$, $P=0.032$); both effect sizes were small ($d=0.384$ and $d=0.316$ respectively). The significant difference between trainees' scores on essay 3 and essay 1 ($t(47)=6.368$, $P<0.001$) reflected a large effect size ($d=0.919$). Thus, there is evidence that trainees improved from assessment early in term II to mid-term III on videotapes but not case studies, and there was progression on essays from term I to term III.

Table 1 Trainees' total scores on all assessments

Assessment	*n*	Minimum	Maximum	Mean (s.d.)
Videotape				
1	52	26.0	48.0	37.3 (4.96)
2	52	28.0	54.0	40.1 (5.65)
Case study				
1	51	47.5	76.0	62.9 (6.08)
2	50	47.5	80.0	63.1 (6.62)
Essay				
1	52	51.0	78.5	61.9 (6.72)
2	49	47.5	81.5	64.7 (8.94)
3	48	47.0	95.0	68.5 (10.18)

The seven examinations as discrete assessments

Table 2 illustrates results of the generalisability analysis on all seven individual examinations. In all instances, systematic differences in trainees' ability (to practise CBT, to conceptualise and report on a case study, and to write an essay) were the main source of variation. However, measurement error was present to a lesser extent in essay assessments than in case study examinations, and present to the largest degree in the two videotape evaluations.

The generalisability coefficients for the seven discrete examinations reflect their ability to minimise error and measure differences among trainees. The coefficients for the design of the Newcastle examination procedure for each assessment (two examiners, one test) reveal that two of the three essays (essays 2 and 3), one case study (case study 2) but neither of the videotape assessments possessed adequate reliability. The fact that we were unable to obtain higher levels of reliability in the single assessments of trainees' practical skills is disappointing because this indicates that we had difficulty generalising across examiners. That is, examiners do not seem to agree largely about the abilities of trainees to deliver CBT effectively.

Combining examinations to evaluate competencies

Treating assessments within examination type as multiple efforts to ascertain trainees' competencies (for example, assuming that trainees' performances on videotapes 1 and 2 indicated their ability to practise cognitive therapy with patients) produced mixed results (Table 3). Systematic differences among trainees accounted for only a minority of total score variance on all three types of examination. This means that efforts to evaluate trainees' competencies resulted in score variability that was much more reflective of measurement error than underlying, relatively stable, knowledge and skills. Again, attempts to measure trainees' skills in CBT proved especially problematic with differences among trainees' ability to conduct CBT explaining only about a seventh of the total score variance.

The most easily identifiable problematic source of measurement error is the instability of trainees' scores across occasions

https://doi.org/10.1192/bjp.bp.107.038588 Published online by Cambridge University Press

Table 2 Magnitudes of sources of variation for videotape, case study and essay examinations as individual assessments with absolute and study absolute generalisability coefficient values

	Percentage of total variance						
	Videotape		Case study		Essay		
Source of variation	1	2	1	2	1	2	3
Trainees	51.08	58.32	62.48	67.47	63.50	80.14	67.01
Residual error	48.92	41.68	37.52	32.53	36.50	19.86	32.99
Absolute G[a]	0.511	0.583	0.625	0.675	0.635	0.801	0.670
Study absolute G[b]	0.676	0.737	0.769	0.806	0.777	0.890	0.803

a. One examiner and one test.
b. Two examiners and one test.

(within assessment type). So, for example, a proportion of trainees scored relatively low on videotape 1 and relatively high on videotape 2, whereas the reverse was true for other trainees. The significant trend towards higher scores on those assessment tapes and essays submitted towards the end of the course introduced relatively little measurement error. There is considerable residual error across all three assessment types, which we were unable to explain because of the naturalistic design of the examination procedure.

In view of the fact that there were particularly large amounts of measurement error in the videotape and case study assessments, it is unsurprising that the generalisability coefficients that reflect the design of the Newcastle assessment procedure (study absolute generalisability values) are not as high as we would have hoped.

If all videotapes, case studies and essays were considered to be seven components of one overall examination, then absolute generalisability = 0.075 (one test, one examiner) and study absolute generalisability = 0.377 (seven tests, two examiners).

D-study analyses

D-study analyses indicated that in order to obtain generalisability coefficients of 0.8 for each assumed competence, the course would need to use two examiners to evaluate 19 videotapes (G=0.808), two examiners to mark 12 case studies (G=0.800), and two examiners to score 5 essays (G=0.803). Increasing the number of examiners rather than the number of tests would yield less increase in generalisability and, therefore, would be less cost-efficient. For example, using 19 examiners to assess two tapes results in G=0.381; 12 examiners to evaluate two case studies produces G=0.461; and 5 examiners scoring two essays results in G=0.680.

Standard error of measurement

Table 4 illustrates the s.e.m. values for the various assessments. It also details the number of trainees who scored below the pass mark on the examinations and the number who fell below the pass mark plus one s.e.m. – this is the score that excludes a failing mark. It is clear that, when due consideration is given to confidence intervals, many trainees' true score on the videotape examinations could fall below the failure threshold. The most significant illustration of this fact is that using examiners' overall mean scores, the confidence intervals of more than nine out of ten (92.3%) trainees' scores overlap the failure point.

Discussion

Findings of the present study

As far as we are aware, this is the first study to use generalisability theory to ascertain the ability of a postgraduate course on CBT to reliably assess trainees' competencies. The results that we have reported here indicate that there are substantial difficulties in examining the knowledge and skills typically thought important in CBT. The more examinations reflected practical skills, the less reliable they became. Thus, essay examinations were more reliable than case study examinations, which were in turn more reliable than efforts to assess practical therapeutic skills. This is especially the case when we consider assessments within examination type as samples of an underlying ability. Again, examinations of practical CBT skills were the most unreliable and generalising from one or two assessments to routine practice is extremely problematic (calculations indicate that about 19 60-min assessments are needed).

Over the past 30 years, research has consistently indicated similar problems in measuring reliably competencies in other areas such as medicine and nursing.[11] Invariably, the major cause of this difficulty was a phenomenon called content specificity. This term describes the fact that performances differ considerably across problems or situations, with correlations of scores typically about 0.1–0.3.[12] For example, doctors' ability to diagnose accurately and make appropriate case management decisions on one case is weakly related to their ability to do so on another, even if both cases are drawn from the same clinical domain.[8,13] So robust is this finding that almost all of those involved in medical education have now abandoned the quest for general skills such as problem-solving and clinical reasoning.[14]

Implications

Ideally, training bodies and clinical supervisors would focus their efforts largely on developing those competencies that lead to improved patient outcome. The first step in this process is to be able to obtain reliable and valid estimates of competence. That is, educational tests and processes are required that can differentiate ability in CBT, and these estimates should relate to patient outcomes. Our results indicate that standard examination

Table 3 Magnitudes of sources of variation for combined videotape, case study and essay examinations, with absolute and study absolute generalisability coefficient values

	Percentage of total variance		
Source of variation	Videotapes	Case study	Essay
Trainees	14.39	20.79	36.98
Occasion	8.47	0.00	6.40
Trainee × occasion	36.03	47.05	21.25
Residual error	41.11	33.41	35.37
Absolute G[a]	0.144	0.208	0.370
Study absolute G	0.307[b]	0.399[b]	0.710[c]

a. One examiner and one test.
b. Two examiners and two tests.
c. Two examiners and three tests.

https://doi.org/10.1192/bjp.bp.107.038588 Published online by Cambridge University Press

Table 4 The number of trainees failing assessments using standard criteria and inclusion of s.e.m. values

Assessment	s.e.m. (+/–)	Examiners' mean *n* (%)	Examiners' mean + s.e.m. *n* (%)
Videotape			
1	5.53	14 (26.9)	35 (67.3)
2	5.68	10 (19.2)	30 (57.7)
1+2	8.93	10 (19.2)	48 (92.3)
Case study			
1	5.73	2 (4.0)	6 (11.8)
2	5.72	0 (0.0)	8 (14.0)
1+2	9.60	0 (0.0)	8 (16.0)
Essay			
1	6.22	0 (0.0)	11 (21.2)
2	5.81	1 (2.0)	9 (18.4)
3	8.86	2 (4.2)	8 (16.7)
1+2+3	9.55	0 (0.0)	11 (23.4)

procedures of practical skills typically employed by university postgraduate courses have relatively low reliability and this in turn imposes a relatively low upper limit on validity.[8]

Educational bodies have endeavoured to address difficulties of reliability by various methods. For example, medical schools have made extensive efforts to reduce the effects of common sources of measurement error by increasing standardisation of cases and increasing sampling of performances across different types of cases using the observed structured clinical examination.[9] However, there has been increasing acknowledgement that the ultimate goal is to assess what health professionals actually do in their daily clinical practice, which presents a considerable challenge.[15–17] The system of formative assessment employed by postgraduate courses on CBT reflects very well this aspiration. Typically, the regular supervision provided is based on both supervisors and supervisees watching and/or listening to many videotapes or audiotapes of therapy with real patients. Theoretically, this means that trainees' clinical work is repeatedly sampled both within and across patients (and therefore most probably across presenting problems). It may be the case that postgraduate courses can use this routinely generated material to increase sampling of trainees' clinical work, although it is a continual challenge for educational institutions to both manage resources and deliver reliable and valid examinations.

Limitations

Our results reflect data collected about only two cohorts of trainees on just one postgraduate course. These trainees may not be a representative sample of all trainees on all postgraduate courses throughout the UK or elsewhere. A larger sample size would also have been beneficial. However, the results detailed herein are not at odds with the general picture that has emerged from myriad studies of the health professions over the past few decades. The naturalistic design of the examination procedure meant we could not dissect more fully the large amounts of residual error. This is especially important because we simply do not know what contextual factors influence significantly trainees' (or experts') clinical performances. For example, perhaps important factors include overall experience, experience of the specific presenting problem, characteristics of patients, or aspects of the match between clinicians and patients. It is also possible that our results reflect the fact that trainees are relative novices and therefore apply their knowledge and skills inconsistently in a similar way to learner drivers. Finally, it is worth noting that when a substantial element of any course is supervision of clinical practice by multiple supervisors, the probable nature and quality of trainees' educational experience will be diverse.[18]

Andrew J. A. Keen, PhD, Child and Family Mental Health Service, Royal Aberdeen Children's Hospital, Aberdeen; **Mark H. Freeston**, PhD, Newcastle Cognitive and Behavioural Therapies Centre, University of Newcastle & University of Durham, UK

Correspondence: Dr Andrew Keen, Child and Family Mental Health Service, Royal Aberdeen Children's Hospital, Westburn Road, Aberdeen AB25 2ZD. Email: Andrew.Keen@ARH.grampian.scot.nhs.uk

First received 2 Apr 2007, final revision 15 Feb 2008, accepted 11 Mar 2008

Acknowledgements

The authors thank the staff at the Newcastle Cognitive and Behavioural Therapies Centre who assisted in data collection.

References

1 Department of Health. *National Service Framework for Mental Health: Modern Standards and Service Models*. Department of Health, 1999.

2 National Institute of Clinical Excellence. *Our Guidance*. NICE, 2006.

3 Layard R. The case for psychological treatment centres. *BMJ* 2006; **332**: 1030–2.

4 Department of Health. *Organising and Delivering Psychological Therapies*. Department of Health, 2006.

5 British Association for Behavioural Cognitive Psychotherapies. *Minimum Training Standards for the Practice of CBT*. BABCP, 2006.

6 Blackburn IM, James IA, Milne DL, Baker C, Standart S, Garland A, Reichelt FK. The revised cognitive therapy scale (CTS–R): psychometric properties. *Behav Cogn Psychother* 2001; **29**: 431–46.

7 Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *Am Psychol* 1989; **44**: 922–32.

8 Streiner DL, Norman GR. *Health Measurement Scales*. Oxford University Press, 2004.

9 Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001; **357**: 945–9.

10 Downing S. Reliability: on the reproducibility of assessment data. *Med Educ* 2004; **38**: 1006–12.

11 Eva, KW. On the generality of specificity. *Med Educ* 2003; **37**: 587–8.

12 van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med* 1990; **2**: 58–76.

13 Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educ Res* 1994; **23**: 23–30.

14 Norman GR. Research in medical education: three decades of progress. *BMJ* 2002; **324**: 1560–2.

15 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990, **65** (suppl 9): S63–7.

16 Brown N, Doshi M. Assessing professional and clinical competence: the way forward. *Adv Psychiatr Treat* 2006; **12**: 81–9.

17 Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ* 2007; **12**: 239–60.

18 Keen A, Klein S, Alexander DA. Assessing the communication skills of doctors in training: reliability and sources of error. *Adv Health Sci Educ* 2003; **8**: 5–16.

https://doi.org/10.1192/bjp.bp.107.038588 Published online by Cambridge University Press