

## Performance-based assessment in continuing medical education for general practitioners: construct validity

J J M Jansen,<sup>1</sup> A J J A Scherpier,<sup>2</sup> J C M Metz,<sup>3</sup> R P T M Grol,<sup>1</sup> C P M van der Vleuten<sup>4</sup>  
& J J Rethans<sup>1</sup>

1 Centre for Research on Quality in Health Care, Universities of Nijmegen and Limburg; 2 Skillslab, University of Limburg, Maastricht; 3 Clinical Training Centre, University of Nijmegen, Nijmegen; and  
4 Department of Educational Development and Research, University of Limburg, Maastricht, the Netherlands

### SUMMARY

The use of performance-based assessment has been extended to postgraduate education and practising doctors, despite criticism of validity. While differences in expertise at this level are easily reflected in scores on a written test, these differences are relatively small on performance-based tests. However, scores on written tests and performance-based tests of clinical competence generally show moderate correlations. A study was designed to evaluate construct validity of a performance-based test for technical clinical skills in continuing medical education for general practitioners, and to explore the correlation between performance and knowledge of specific skills. A 1-day skills training was given to 71 general practitioners, covering four different technical clinical skills. The effect of the training on performance was measured with a performance-based test using a randomized controlled trial design, while the effect on knowledge was measured with a written test administered 1 month before and directly after the training. A training effect could be shown by the performance-based test for all four clinical skills. The written test also demonstrated a training effect for all but one skill. However, correlations between scores on the written test and on the performance-based test were low for all skills. It is concluded that construct validity of a performance-based test for technical clinical skills of general practitioners was demonstrated, while the knowledge test score was shown to be a poor predictor of competence for specific technical skills.

### Keywords

Clinical competence; \*education medical continuing; educational measurement/\*methods; family practice/\*education; KAP; Netherlands; randomized controlled trial

### INTRODUCTION

In measurement of clinical competence the use of direct observation of clinical performance under standardized

conditions has become a popular assessment method because it directly assesses behaviour considered relevant to clinical performance. The method has been extensively studied, providing general supportive evidence for validity and acceptable reliability (Van der Vleuten & Swanson 1990; Colliver & Williams 1993; Vu & Barrows 1994).

Performance-based testing has also been extended to postgraduate education (Stillman *et al.* 1986; Cohen *et al.* 1990; Joorabchi 1991; Grand'Maison *et al.* 1992) and assessment of practising doctors (Rethans *et al.* 1991; Norman *et al.* 1993; Jansen *et al.* 1995). However, the use of this method to assess clinical competence at postgraduate level and among practising doctors has been criticised for lack of validity, because of the rigidity (Cox 1990) or trivialization (Norman *et al.* 1991) of the scoring methods used. These critics suggest that the assessment method based on checklists may be appropriate to assess basic history-taking and physical examination skills, but not in discriminating between different levels of expertise at graduate level and beyond. Nevertheless, validation studies have shown (small) differences in mean scores between senior students and residents (Cohen *et al.* 1990; Joorabchi 1991; Brailovsky *et al.* 1995), and between junior and senior levels within residency training (Stillman *et al.* 1986; Petrusa *et al.* 1990). Few studies have included comparison of residents and practising doctors. In two experiments comparing residents in family medicine and practising doctors, no overall differences in score were found, although one study reported differences in subscores (Brailovsky *et al.* 1995; Jansen *et al.* 1995). This finding could be explained by the failure of the instrument to measure relevant differences in clinical competence as well as by the failure of the theory underlying the construct, i.e. practising doctors are more competent in the skills assessed in the test compared to residents (Crocker & Algina 1986).

One way to further evaluate construct validity is to assess the discriminating power of performance-based tests among groups of practising doctors with differences

in competence. Norman *et al.* (1993) compared a criterion group of competent doctors, with self-referred doctors and doctors referred by the licensing body because of deficiencies, using multiple assessment methods, and found significant differences on the standardized patient-based test but not on the objective structured clinical examination.

Written tests can discriminate very well between different levels of competence at postgraduate level compared to performance-based tests (Swanson *et al.* 1987; Quattlebaum *et al.* 1989; Benson 1991; Norman *et al.* 1994), but have been criticised for lack of validity beyond recall of knowledge (Levine *et al.* 1970; Dixon 1978; Neufeld 1985). However, studies correlating results on written and performance-based test formats have found moderate to high true correlations (Van der Vleuten & Swanson 1990), providing supportive evidence for the assumption of a relation between knowledge and performance of clinical skills (Miller 1990). It has been argued that these high correlations are perhaps a result of memorizing the checklists used in the performance-based test (Van Luijk *et al.* 1990; Norman *et al.* 1991), but in a recent study among family doctors not familiar with the content of checklists used, a moderate correlation was also found between scores on a written test and a performance-based test covering a broad domain of technical clinical skills (Jansen *et al.* 1995). In continuing medical education, however, courses focus on specific topics rather than on a broad domain, and it is not clear if the correlation between knowledge and performance is as high for specific skills.

An experimental study was designed to evaluate construct validity of a performance-based test for technical clinical skills in continuing medical education of general practitioners, and compare results for the specific skills on the performance-based test with scores on a written test of skills. Our research questions were:

1. Can the performance-based test discriminate between groups of practising doctors with different competence for specific technical clinical skills?
2. How accurately can the results for specific skills on the performance-based test be predicted by the scores on corresponding parts of the written test?

## METHODS

A 1-day training course in technical clinical skills was developed. The training focused on four topics: physical examination of the shoulder, injection techniques of the shoulder, cardiopulmonary resuscitation and intravenous cannulation. These topics were selected as having priority based on a survey among 20 general

practitioners actively involved in CME throughout the country. Training was based on national clinical guidelines developed by professional bodies (Grol 1990). The training time was 1 hour for each topic, and each training was given in small groups (8-12 participants) by two experienced trainers with special interest in the subject concerned. It was assumed that such a training would result in a considerable improvement in competence.

## Instruments/materials

The effect of the skills training on performance was assessed by a performance-based test consisting of four OSCE stations, covering the four topics addressed in the course. Checklists were used for scoring performance and for providing feedback, with criteria based on the national guidelines for general practice. The checklist for examination of the shoulder contained 36 items, for injection of the shoulder 20 items, for resuscitation 16 items, and 25 items for intravenous cannulation. Checklists were developed by a committee of general practitioners, reviewed by at least three faculty members and pilot-tested before the course. In addition to the checklist a 10-point global rating scale was used as a general measure of performance.

In one station (shoulder examination) students with experience as standardized patients were used. They were trained for their role by a general practitioner experienced in the training of standardized patients in a 2-hour training session. In the other stations manikins (Resusci-Anni® CPR model; Limbs&Things® shoulder injection model; Syma® arm model) were used.

A total of 36 general practitioners (staff members from two departments of general practice) were involved as raters. One-third of the encounters were double-rated to determine inter-rater reliability. Two weeks before the course the raters received a 1-hour training. To improve consensus, scoring was practised in the training session and inter-rater differences were discussed.

The effect of the skills training on knowledge of the participants was assessed by a written test which covered the content of the course. The 49 items consisted of statements with three answering options: true, false or question mark. The statements covered knowledge about the four technical clinical skills. The number of items for each topic was based on the number of relevant statements that could be constructed, resulting in 20 items about shoulder examination, 10 items about shoulder injection, and 13 items about resuscitation. Only six items about intravenous cannulation were included because it proved difficult to construct more meaningful questions about this technical skill.

## Procedure

The course was announced in a mailing to general practitioners in the region. Participants ( $n=71$ ) were divided at random into two groups. At the course one group (A,  $n=32$ ) started with the training of shoulder examination and injection techniques, followed by the training on resuscitation and intravenous cannulation, while the other group (B,  $n=39$ ) received the training in the opposite order (Fig. 1). The performance-based test was administered between the two training sessions.

Because of the randomized assignment of the participants the two groups could serve as each others' controls for the different topics. As group A received the training on examination and injection of the shoulder before entering the performance-based test, while group B received this training after the test, the effect of this training could be evaluated by comparing the scores of both groups on the stations assessing examination and injection of the shoulder. The same comparison could be made for resuscitation and intravenous cannulation, where group A served as a control for group B. The participants received immediate feedback at each station on their performance by the rater using the checklist.

The knowledge test was mailed to all participants 1 month before the course and administered again directly after the training (pretest-post-test design). Participants only received feedback on their scores after the post-test.

## Statistical analysis

The complete results on all four performance stations were available for 71 participants. As 10 participants failed to return the written pretest, complete data on the written tests were available for 62 participants. Raw

scores on the performance-based test and on the written test (number of correct items) were converted into a percentage score, and *T*-test was used to compare mean scores. Reliability of the knowledge test score was determined by calculating a Cronbach's alpha reliability coefficient (Cronbach *et al.* 1972) and for the performance-based test inter-rater reliability was assessed using intra-class correlation coefficients (Kramer & Feinstein 1981). Correlations between knowledge test score and performance-based test score were determined using Pearson product-moment coefficients (Welkowitz *et al.* 1982).

## RESULTS

### Subjects

The 71 participants had a mean age of 41 years (range 30-55) and 10 years of experience (range 1-24) as family doctors. Most doctors (69%) worked full-time in their practice, with the remaining working 3-4 days (20%) or less (11%) in a practice. Mean practice size was 2500 patients (range 600-3600). Practice localizations were largely (sub)urban (41%) or small town (36%), and 23% were rural. Only 25% worked in a solo practice, 41% in a duo-practice and the remaining 34% worked in a group practice or health centre. Compared to the population of Dutch general practitioners, there were more female doctors and part-timers among the participants, and fewer doctors working in a solo practice, while age distribution, practice size and practice localization of the participants can be considered as representative. Doctors in group A ( $n=32$ ) and group B ( $n=39$ ) did not differ in characteristics, nor on the written test score prior to the course, suggesting that randomization had been successful.

### Reliability

The inter-rater reliability coefficients for the checklist scores on the four stations of the performance-based test were: 0.97 for examination of the shoulder; 0.98 for injection of the shoulder; 0.93 for intravenous cannulation; and 0.79 for resuscitation (the values based on the rating scale were, respectively, 0.88, 0.89, 0.75 and 0.70). These figures indicate that interobserver variability was minimal. The reliability coefficient for the written test was 0.72 for the pretest and 0.64 for the post-test.

### Scores

Table 1 shows the results for the performance-based test for the checklist score and rating scale. Before training,

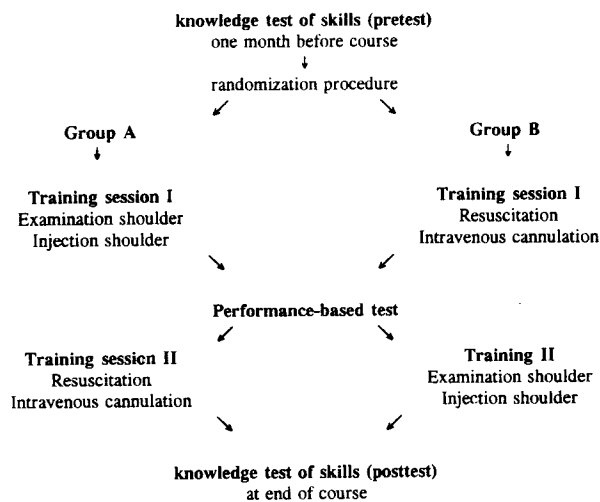


Figure 1 Design for training course and assessment sequence.

	n	Checklist			Rating scale		
		Mean	SD	T-test*	Mean	SD	T-test*
Examination shoulder							
before training	39	51.5	15.8	$P < 0.001$	65.9	13.1	$P < 0.001$
after training	32	73.7	10.3		76.9	11.2	
Injection shoulder							
before training	39	38.7	20.6	$P < 0.001$	50.9	16.3	$P < 0.001$
after training	32	73.8	14.6		72.2	10.4	
Resuscitation							
before training	32	65.8	12.2	$P < 0.001$	60.0	11.8	$P < 0.001$
after training	39	78.0	12.3		75.4	10.9	
Intravenous cannulation							
before training	32	50.3	25.5	$P < 0.001$	49.4	24.4	$P < 0.001$
after training	39	77.9	15.0		76.4	14.4	

All scores expressed as percentage of maximum score. \* T-test for difference before-after training.

the mean scores on all stations revealed considerable deficiencies in performance, especially for the shoulder injection, while performance concerning resuscitation was relatively good. Based on the checklist score a significant improvement was found on all stations after training, with a mean increase in score of 24% (range 12-35%) of the maximum score, and smaller standard deviations in the group who had received training on three of four topics, supportive for a training effect. The increase of the score on the resuscitation station was somewhat lower compared to the other stations. The rating scale scores mirrored closely the checklist scores, with ratings being only somewhat less stringent for pre-training performance on the shoulder stations.

Table 2 provides the scores on the written test 1 month before and directly after the training for the different topics. The scores showed significant improvement on all topics except for intravenous cannulation. The pretest score for intravenous cannulation was high, indicating that questions were probably relatively easy and limiting possibility of improvement.

### Correlation

The scores on the checklists and the general ratings were correlated for all four stations, resulting in a correlation coefficient of 0.80 for examination of the shoulder, 0.87 for injection of the shoulder, 0.60 for resuscitation and 0.80 for intravenous cannulation. The checklist scores on the performance-based test were correlated with the pretest scores and post-test scores on the knowledge test.

**Table 1** Checklist and rating scale scores on the performance-based test

The scores on the performance-based stations for participants before the training were matched with their scores on the corresponding parts of the written pretest, while for the scores on the stations after the training the corresponding parts of the written post-test were used. The results are presented in Table 3. Correlations between scores on the knowledge test and the performance-based test are variable, decreasing from significant to non-significant after training for 'injection of the shoulder', while increasing to significant ( $P < 0.05$ ) for

**Table 2** Scores on the written test before and after training

	Mean	SD	Paired T-test
Examination shoulder (20 items)			
before training	66.3	15.8	$P < 0.001$
after training	79.3	13.0	
Injection shoulder (10 items)			
before training	54.8	20.1	$P < 0.001$
after training	77.7	13.0	
Resuscitation (13 items)			
before training	56.9	13.2	$P < 0.001$
after training	67.6	13.8	
Intravenous cannulation (6 items)			
before training	75.5	22.3	$P = 0.264$
after training	78.8	14.2	

All scores expressed as percentage of maximum score.

**Table 3** Correlations of the performance-based test scores (checklist and rating scale) with the knowledge test scores

	Checklist	Rating scale
Examination shoulder		
pretest score	0.20	0.28
post-test score	0.43*	0.23
Injection shoulder		
pretest score	0.35*	0.30
post-test score	-0.20	0.03
Resuscitation		
pretest score	0.14	-0.20
post-test score	0.35*	0.01
Intravenous cannulation		
pretest score	0.24	0.24
post-test score	-0.05	-0.20

\*  $P < 0.05$ .

'examination of the shoulder' and 'resuscitation'. Correlation of the general rating with the written tests resulted in comparable figures.

## DISCUSSION

A considerable training effect was demonstrated on the performance-based test (both on the checklist and on the general rating scale) for all four clinical skills in a short hands-on skills training in small groups for practising family doctors. These results suggest that a performance-based assessment method can indeed discriminate between different levels of proficiency among practising doctors which provides support for construct validity. Other recent studies have demonstrated similar results for different technical clinical skills (Nyquist *et al.* 1994; Carney *et al.* 1995). Inter-rater reliability was high as has been reported in other studies concerning clinical skills (Wakefield 1985), with rating scales having a somewhat lower reliability (Van Luijk & Van der Vleuten 1992).

The knowledge test score also improved for all but one skill as a result of the training. The knowledge test failed to demonstrate a training effect for intravenous cannulation, while performance did improve by more than 25%. A likely explanation is that questions in the knowledge test were too easy, so discriminating power was lost.

Correlations between checklist scores and general ratings were high for all stations, except resuscitation, which showed a moderate correlation. This could indicate that some relevant performance aspects were not well covered by the checklist. For the other three stations the high correlations with the general ratings are sup-

portive for content validity of the checklist since the raters were experienced general practitioners, and therefore were considered experts in the evaluation of performance of their peers. These results indicate that both rating scales and checklists seem appropriate measurement tools in assessment of performance of technical clinical skills of general practitioners.

Correlations between scores on the written test and the performance-based test were variable but low. Even when leaving intravenous cannulation out of consideration because of the above-mentioned problems, knowledge of a skill was not a reliable predictor of proficiency for that specific technical clinical skill as knowledge predicted only a very small part of the variance on the performance-based test for the different skills. The low reliability of the written test used may have had a negative influence on the correlations. However, the content of each specific skill puts a limit to the number of meaningful items from which a written test can be sampled, contrary to assessment of clinical competence as a general construct where the domain from which items for test construction can be sampled is very large. Correction for unreliability was therefore not considered appropriate. The results are consistent with an earlier study (Vu & Barrows 1990). Although scores on knowledge tests and performance-based tests can have a high correlation when generalized over a broad domain (Newble & Swanson 1988; Van der Vleuten, Van Luijk & Beckers 1988; Jansen *et al.* 1995), this relation is not necessarily replicated for specific skills.

In conclusion, while both the performance-based test and written test were able to demonstrate a training effect, they apparently measured different things: performance ('shows how') and knowledge ('knows'), applying the terminology of Miller (1990). Knowledge, perhaps useful as a predictor of performance when generalized over a broad domain, resulted in being a poor predictor of performance for specific technical skills. For assessment of mastery of specific technical clinical skills a performance-based test is preferably used, and both checklists and rating scales seem suitable.

## REFERENCES

- Benson J A (1991) Certification and recertification: one approach to professional accountability. *Annals of Internal Medicine* 114, 238-42.
- Brailovsky CA, Grand'Maison P & Lescop J (1995) Construct validity of the objective structured clinical examination used in the Quebec licensing examination. In: *Proceedings of the Sixth Ottawa Conference on Medical Education*. (ed. by AI Rothman & R Cohen), (pp. 373-4). University of Toronto Bookstore, Toronto.

- Carney P A, Dietrich A J, Freeman D H & Mott L A (1995) A standardized-patient assessment of a continuing medical education program to improve physicians' cancer-control clinical skills. *Academic Medicine* **70**, 52-8.
- Cohen R, Reznick R K, Taylor B R, Provan J & Rothman A (1990) Reliability and validity of the objective structured clinical examination assessing surgical residents. *The American Journal of Surgery* **160**, 302-5.
- Colliver J A & Williams RG (1993) Technical issues: test application. *Academic Medicine* **68**, 454-60.
- Cox K (1990) No Oscar for OSCA. *Medical Education* **24**, 540-5.
- Crocker L & Algina J (1986) *Introduction to Classical and Modern Test Theory*. Harcourt Brace Jovanovich, Orlando, Florida.
- Cronbach L J, Gleser G C, Nanda H & Rajaratnam N (1972) *The Dependability of Behavioral Measurements*. John Wiley, New York.
- Dixon J (1978) Evaluation criteria in studies of continuing education in the health professions: a critical review and a suggested strategy. *Evaluation and the Health Professions* **1**, 47-65.
- Grand'Maison P, Lescop J, Rainsberry P & Brailovsky C A (1992) Large-scale use of an objective structured clinical examination for licensing family physicians. *Canadian Medical Association Journal* **146**, 1735-40.
- Grol R P T M (1990) National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. *British Journal of General Practice* **40**, 361-4.
- Jansen J J M, Tan L H C, Van der Vleuten C P M *et al.* (1995) Assessment of competence in technical clinical skills of general practitioners. *Medical Education* **29**, 247-53.
- Joorabchi B (1991) Objective structured clinical examination in a pediatric residency program. *American Journal of Diseases of Children* **145**, 757-62.
- Kramer M S & Feinstein A R (1981) Clinical biostatistics. LIV. The biostatistics of concordance. *Clinical Pharmacology and Therapeutics* **20**, 111-23.
- Levine H G, McGuire C H & Nattress L W (1970) The validity of multiple choice achievement tests as measures of competence in medicine. *American Education Research Journal* **1**, 69-83.
- Miller G E (1990) The assessment of clinical skills/competence/performance. *Academic Medicine* **65**, S63-7.
- Neufeld V R (1985) Written examinations. In (ed. by VR Neufeld & GR Norman. *Assessing Clinical Competence* pp. 94-118. Springer, New York.
- Newble D I & Swanson D B (1988) Psychometric characteristics of the objective structured clinical examination. *Medical Education* **23**, 325-34.
- Norman G R, Van der Vleuten C P M & De Graaff E (1991) Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education* **25**, 119-26.
- Norman G R, Davis D A, Lamb S *et al.* (1993) Competency assessment of primary care physicians as part of a peer review program. *Journal of the American Medical Association* **270**, 1046-51.
- Norman G R, Trott A D, Brooks L R & Smith E K M (1994) Cognitive differences in clinical reasoning related to postgraduate training. *Teaching and Learning in Medicine* **6**, 114-20.
- Nyquist J G, Naylor A J, Woodward-Lopez G & Dixon S (1994) Use of performance-based assessment to evaluate the impact of a skill-oriented continuing education program. *Academic Medicine* **69**, S51-3.
- Quattlebaum T G, Darden P M & Sperry J B (1989) In-training examinations as predictors of resident clinical performance. *Pediatrics* **84**, 165-72.
- Petrusa E, Blackwell T & Ainsworth M (1990) Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Archives of Internal Medicine* **150**, 573-7.
- Rethans J J, Sturmans F, Drop R, Van der Vleuten C & Hobus P (1991) Does competence of general practitioners predict their performance. *British Medical Journal* **303**, 1377-85.
- Stillman P L, Swanson D B, Smee S *et al.* (1986) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine* **105**, 762-71.
- Swanson D B, Norcini J & Grosso L J (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* **12**, 220-46.
- Van der Vleuten C P M, Van Luijk S J & Beckers H J M (1988) A written test as an alternative to performance testing. *Medical Education* **23**, 97-107.
- Van der Vleuten C P M & Swanson D B (1990) Assessment of clinical skills with standardized patients; state of the art. *Teaching and Learning in Medicine* **2**, 58-76.
- Van Luijk S J, Van der Vleuten C P M & Van Schelven S M (1990) The relationship between content and psychometric characteristics in performance-based tests. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, R J Hiemstra, A J JA Scherpbier & RP Zwierstra) pp. 202-7. Boekwerk, Groningen.
- Van Luijk S J & Van der Vleuten C P M (1992) A comparison of checklists and rating scales in performance-based testing. In: (ed. by I R Hart & R M Harden) *Current Development in Assessing Clinical Competence* (pp. 357-362). CanHeal, Montreal.
- Vu N V & Barrows H S (1990) Validity and accuracy of performance and written evaluations in assessing history and physical examination skills. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, RJ Hiemstra, A J J A Scherpbier & R P Zwierstra), (pp. 283-7). Boekwerk, Groningen.
- Vu N V & Barrows H S (1994) Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educational Researcher* **23**, 23-30.
- Wakefield J. Direct Observation (1985) In: *Assessing Clinical Competence* (ed. by V R Neufeld & G R Norman) pp. 51-70. Springer, New York.
- Welkowitz J, Ewen R B & Cohen J (1982) *Introductory Statistics for the Behavioral Sciences*. Harcourt Brace Jovanovich, Orlando, Florida.

Received 4 October 1995; editorial comments to authors 15 December 1995; accepted for publication 14 February 1996