

Assessment of competence in technical clinical skills of general practitioners

J J M Jansen[1], *L H C Tan*[2], *C P M van der Vleuten*[3], *S J van Luijk*[3], *J J Rethans*[1] & *R P T M Grol*[1]

[1]Centre for Research in Quality Assurance in General Practice, Universities of Nijmegen and Limburg, The Netherlands, [2]National Centre for Evaluation of Vocational Training for General Practitioners, University of Utrecht, The Netherlands and [3]Department of Educational Development and Research, University of Limburg, Maastricht, The Netherlands

SUMMARY

Technical clinical procedures constitute an important part of the work of general practitioners. Assessment of competence in the relevant skills is important from the perspective of quality assurance. In this study, the psychometric characteristics of three different methods for assessment of competence in technical clinical skills in general practice were evaluated. A performance-based test (8 stations), a written knowledge test of skills (125 items) and a self-assessment questionnaire (41 items) on technical clinical skills were administered to 49 GPs and 47 trainees in general practice. The mean scores on the performance-based test and the written knowledge test of skills showed no substantial differences between GPs and trainees, whereas the GPs scored higher on the self-assessment questionnaire. While the correlation of the score on the knowledge test of skills with the score on the performance-based test was moderately high, the score on the self-assessment questionnaire showed a rather low correlation with the performance-based test. Although performance-based testing is obviously the best method to assess proficiency in hands-on skills, a written test can serve as a reasonable alternative, particularly for screening and research purposes.

Keywords

*clinical competence; *family practice; psychometrics; employee performance appraisal; self-evaluation programs; educational measurement; students, medical

INTRODUCTION

Technical clinical procedures constitute an important part of the daily work of doctors (Lamberts *et al.* 1991), and proficiency in technical clinical skills is considered a relevant aspect of clinical competence (Fabb 1983). From the perspective of quality assurance of medical care it is therefore important to gather reliable data on competence in relevant technical clinical skills, as a basis for planning continuing education programmes (Berg 1979). The aim of the study presented here was to identify and evaluate different methods for assessment of competence in technical clinical skills in general practice.

Direct observation of performance under standardized conditions has been identified as the first choice assessment method. This method, originally described by Harden & Gleeson (1979) as the objective structured clinical examination (OSCE), has been extensively researched, mainly in undergraduate programmes and to a lesser degree in postgraduate education (Hart *et al.* 1986, 1992; Hart & Harden 1987; Bender *et al.* 1990). It has generally been considered a valuable method, because of good validity. However, the OSCE has some disadvantages in terms of organizational complexity and resources needed (Anderson & Kassebaum 1993; Reznick *et al.* 1993). This threatens feasibility for widespread use in postgraduate quality assurance schemes. The use of a written test and self-assessment were therefore considered as potential alternative methods for performance-based testing.

Theoretically, a relationship is assumed between knowledge and competence in skills (Patrick 1992). At graduate level the correlation between scores on performance-based tests and written tests assessing clinical competence seems variable (Van der Vleuten & Swanson 1990). Some of the differences found can perhaps be explained by differences in content of the tests compared. Newble & Swanson (1988) reported a moderately high correlation (0·88) between an objective structured clinical examination (patient stations) and a short-answer test in the final-year examination, using the same blueprint for both tests. Van der Vleuten *et al.* (1988) also found a high correlation (0·89) between a written test and a performance-based test constructed according to the same blueprint among senior medical students. These studies showed that a written test score has potential predictive value for a performance-based test score in a population of graduating students. However, this could be

Correspondence: J J M Jansen, Department of General Practice, University of Limburg, PO Box 616, 6200 MD Maastricht, The Netherlands

quite different among practising doctors working in variable practice conditions and having variable continuing medical education experience.

The consideration of self-assessment as another alternative method originated from the literature on adult learning (Fuhrmann & Weissburg 1978), which views self-assessment as an important requisite for effective learning. Yet little research has been published concerning the validity of self-assessment. Results show low to moderate correlations between self-assessment and objective methods, with higher correlations between self-assessment and performance compared to self-assessment and knowledge (Gordon 1991, 1992). Most research is based on undergraduate student populations. As self-assessment is considered a skill which has to be acquired, experienced professionals might be more accurate than undergraduate students in self-assessment of their performance of technical clinical skills (Wooliscroft *et al.* 1993).

The specific research-questions of the study presented here were:

1 Do the identified methods to assess competence in technical clinical skills discriminate between different levels of experience among GPs?
2 What is the reliability of the three different test methods?
3 What is the relationship of the scores of the written test and the self-assessment questionnaire with the performance-based score?

METHODS

Subjects

In March 1992, a test was administered to 49 GPs (all involved as teachers in vocational training in general practice) and 47 trainees in general practice, recruited from two university training institutions. The GPs' experience in practice ranged from 5 to 25 years (mean 13 years): 11 had less than 10 years' experience, 23 had 10–15 years' experience and the remaining 15 had more than 15 years of working experience as a GP. The trainees were at different stages of vocational training: 12 at 3 months (beginners), 16 at 7 months (intermediate), and 19 between 19 and 23 months (advanced).

Instruments

The test consisted of three different parts: a performance-based test, a written knowledge of skills test and a self-assessment questionnaire.

Performance-based test. Stations for the performance-based test (PBT) were developed by a national committee of six practising GPs and two test experts, and based on nationally accepted reference literature for GPs, including national guidelines for GPs as developed by the Dutch College of General Practitioners (Grol 1990).

Check-lists for scoring contained items considered as crucial for adequate performance, as agreed upon by consensus by the committee. Each item was defined in one or more subitems. An illustration is provided in Fig. 1. Performance on all subitems had to be adequate to obtain a favourable marking of the item. Each correct item was given one credit point. Incorrectly or not performed items received no points.

The testing time per station varied between 10 and 20 minutes, adding to a total testing time of 2 hours for the eight stations. Four stations included the management of clinical problems (chest pain, urinary tract infection, impaired hearing, ankle sprain), and standardized patients were used. The remaining four stations covered the performance of isolated technical skills (ophthalmoscopy, urinary catheterization, resuscitation, insertion of an intrauterine device). Mannequins were used in these stations. There were no written (follow-up) stations.

Knowledge test of skills. The written knowledge test of skills (KTS) contained 125 items concerning different technical clinical skills relevant to general practice. The items had the form of statements requiring judgement as true or false (see Fig. 1). If in doubt about the correct answer, a question mark could be used. Of the 125 items, 75 were constructed with a content corresponding to the performance-based test. The remaining 50 questions focused on relevant technical skills not covered by the PBT, thus allowing comparison of prediction of the performance-based scores by the different subsets of items.

Self-assessment questionnaire. The self-assessment questionnaire (SAQ) consisted of 41 items, with 20 items corresponding to the content of the PBT. The remaining items corresponded to skills only covered by the written test. For each item the candidates were prompted to indicate the level of their proficiency in the particular skill using a 5-point Likert-scale (very poor–poor–regular–good–very good) (see Fig. 1).

Procedure

The candidates were tested on four different days, and on two different sites. After the SAQ was completed, the candidates passed through the first part of the PBT, subsequently the KTS, and finally the second part of the PBT. This test sequence was used for logistical reasons.

A group feedback session was held at the end. Candidates and raters were prompted to comment on the content of the test and the testing procedure was

Scoring grid Resusciation

Testing date:
Ratercode:
Candidate:

	Not performed	incorrect	correct
Initial pocedures			
1. Checks consciousness – tries to wake patient with loud voice – gives adequate painstimulus	☐	☐	☐
2. Checks circulation – onesided feeling for carotid-pulsations	☐	☐	☐
3. Checks if airway is free	☐	☐	☐
4. The first 3 items are performed within 30 seconds	☐	☐	☐
5. The first 3 items are performed in the above mentioned order	☐	☐	☐
Resuscitation			
6. Starts directly with resuscitation	☐	☐	☐
7. Cardiac massage is performed correctly – shoulders of resucitator are above sternum patient – hands are crossed on the sternum two fingers above xyphoid – rythm: 15 compressions in 10 seconds	☐	☐	☐
8. Performs two insufflations after each 15 compressions	☐	☐	☐
9. Performs insufflations correctly – brings head of patient in hyperextension – fully covers mouth of patient during insufflation – doesn't allow air to escape from nose of patient – watches whether chest rises during insufflation – chest rises during insufflation	☐	☐	☐

Questions on resuscitation

The GP decides to resuscitate an infant (less than 1 year old), who has no signs of spontaneous breathing nor arterial pulsations. The head of the infant is hyperextended.

1. The correct extent of hyperextension of the head is LESS with an infant compared to an adult.

The GP places his mouth over the nose and mouth of the infant.

2. This is a correct procedure of insufflation of an infant.

During the resuscitation the GP gives thorax compression at a rate of about 90 per minute

3. This is a correct rate for intants.

4. During resusictation of an adult the adequate rate of compressions is closer to 80 per minute than to 60 per minute.

5. During resuscitation with one resuscitator the recommended schedule is: 15 compressions followed by 1 insufflation.

Self-assessment questionnaire

	how often performed						proficiency				
	0–1x per year	2–6x per year	7–15x per year	16–50x per year	>50x per year	not relevant for gp	very bad	bad	regular	good	very good
14. Resuscitation adult	1	2	3	4	5	6	1	2	3	4	5
15. Resuscitation child	1	2	3	4	5	6	1	2	3	4	5

Figure 1 Scoring grid, questions and self-assessment items on resuscitation

evaluated. As a result of comments by the candidates and raters 6 out of 78 items were removed from the check-lists of the PBT before final analysis, and 10 items were removed from the KTS, leaving 115 items for analysis. As a consequence of the link between the items on the KTS and the SAQ, six items were removed from the questionnaire, having 35 items remaining on the SAQ.

The standardized patients were recruited from a group of experienced standardized patients from one of the participating universities. They were trained by a GP experienced in the training of standardized patients.

A total of 36 GPs (staff members of departments of general practice) were involved as raters. One-third of the encounters were double-rated. One week before the test the raters received a 2-hour training. During the training session scoring was practised and results were compared and discussed. The interrater reliability was 0·82 for the total check-list scores (intraclass correlation, including absolute and relative differences between raters in the error term).

Analysis

For the PBT, the individual test score was calculated as the mean of the scores on the different stations. The KTS score was based on the sum of correct answers, and the score on the SAQ was calculated as the sum of scores on the Likert 5-point scale. All scores were expressed as percentages of the maximum score.

The statistical analysis performed included a one-way analysis of variance using a multiple comparisons test (Student–Neuman–Keuls) for differences between groups. Generalizability theory (Cronbach *et al.* 1972) was used to calculate the reliability coefficients for relative and absolute decisions, and interrater reliability. Correlations were calculated as Pearson product–moment coefficients.

Generalizability theory may be considered as an extension of classical test theory. In classical test theory, the observed variance is seen as composed of two sources: true score variance and error variance. Reliability is defined as the ratio between the true score variance and error score variance. Generalizability analysis allows for

Table 1 Scores on the performance-based test (PBT), the knowledge test of skills (KTS) and the self-assessment questionnaire (SAQ)

	PBT total score (8 items)			KTS total score (115 items)			KTS subscore (72 items)			SAQ total score (35 items)			SAQ subscore (21 items)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Trainees ($n = 47$)	56	9	35–76	61	7	44–79	58	6	44–74	64	7	50–82	61	7	45–84
GPs ($n = 48$)	56	9	34–75	65[a]	7	51–79	61	7	46–78	70[c]	9	53–88	69[c]	9	49–89
Trainees															
Beginners ($n = 12$)	53	10	35–76	56[b]	7	44–68	56	6	44–63	63	7	50–75	60	7	45–72
Intermediate ($n = 16$)	57	8	41–73	61	5	52–71	58	5	50–68	64	5	54–75	61	7	51–74
Advanced ($n = 19$)	58	8	39–72	63	7	52–78	60	6	49–74	65	7	55–82	64	8	54–84
GPs															
<10 years ($n = 11$)	56	7	44–67	65	8	54–77	61	8	49–74	68	7	60–84	65	7	45–72
10–15 years ($n = 23$)	55	9	34–72	64	6	51–74	60	7	46–72	71	9	53–86	69	9	51–74
15 years ($n = 14$)	57	9	51–75	67	7	54–79	61	8	50–78	72	10	58–88	72	10	54–84

Note: all entries are expressed as percentage scores.
[a]GPs > trainees $P < 0{\cdot}001$.
[b]Beginners < intermediate = advanced $P < 0{\cdot}05$.
[c]GPs > trainees $P < 0{\cdot}001$.

Table 2 Reliability indicators of the performance-based test (PBT), knowledge test of skills (KTS) and self-assessment questionnaire (SAQ) for total scores and subscores

	Norm-referenced reliability	Testing time to reach 0·80 (hours)	Domain-referenced reliability	Testing time to reach 0·80 (hours)	SEM* (%)
PBT	0·43	10·0	0·35	14·5	7·7
KTS	0·68	1·7	0·64	2·1	4·5
KTS subscore	0·43		0·37		5·8
SAQ	0·92	0·1	0·90	0·1	2·8
SAQ subscore	0·87		0·83		3·8

*Standard error of measurement (SEM) expressed as percentage of maximum score.

partition of the error variance into multiple sources. Depending on the perspective (relative or absolute interpretations), multiple error variances can be estimated, resulting in multiple reliability coefficients. The norm-referenced reliability coefficient is appropriate when test scores are used for the rank ordering of the candidates (e.g. candidate A is better than candidate B). The domain-referenced reliability coefficient is appropriate for absolute score interpretations (e.g. candidate A masters 70% and candidate B 60%).

RESULTS

Scores

Complete data were available from all 96 candidates on the PBT and on the KTS. One candidate failed to complete the SAQ. There were no statistically significant differences between sites and days of administration.

Table 1 shows the scores of the candidates on the PBT, the KTS and the SAQ. Results of experienced GPs were compared with trainees. Within both groups the results were broken down for differences in experience.

On the PBT there was no difference in mean scores between GPs and trainees. There was also no difference in score among the GPs with different years of practice experience. Within the group of trainees there was a trend of slight improvement in scores in relation to stage of vocational training.

The results on the KTS showed a statistically significant difference between the mean scores of GPs and trainees ($P < 0{\cdot}001$). There was no difference in score between GPs with varying years of experience in practice. The mean scores of the trainees increased with experience level, reaching a statistically significant difference only for the scores of the beginners group versus the scores of the other groups ($P < 0{\cdot}05$). The subscores on the KTS, based on the answers to the 75 questions linked with the PBT, however, showed no statistically significant differences.

On the SAQ there was a significant difference between GPs and trainees for the total score as well as for the subscore ($P < 0{\cdot}001$). Within the trainee group as well as within the group of GPs there was a small increase in score associated with level of training respective to years of practice. However, the differences were not statistically significant ($P > 0{\cdot}05$).

Reliability

In Table 2 the results are presented of the generalizability analysis based on the personal scores. The norm-referenced reliability coefficient reflects the reliability of the rank ordering of candidates. A reliability of 0·80 is often considered as a minimum requirement if scores are used as a basis for individual decision-making. The required testing time to reach such a norm-referenced reliability was calculated for the different tests, resulting in considerable time required for the PBT. The domain-referenced reliability coefficient indicates how reliable the absolute scores are. It is naturally more severe since not only the differences in rank ordering but also the absolute differences in scores on the items (item or test difficulty) are taken into account. This explains why the required testing time to reach a reliability coefficient of 0·80 is considerably longer compared to the norm-referenced approach. Table 2 also includes the standard error of measurement (SEM) for the different tests as an alternative reliability index. The SEM can be used to estimate a confidence interval for individual test scores (multiplying the SEM by 1·96, a 95% confidence interval is obtained, e.g. the 95% confidence interval for the KTS score of candidate A with a test score of 70% ranges from 61% to 79%). Large confidence intervals are to be taken into account for the performance based test.

Correlations

The correlations between total test scores on the different assessment methods were calculated. Calculations were repeated using the subscores of 75 items of the KTS and 19 items of the SAQ linked to the content of the PBT. The correlations were recalculated after correction for attenuation caused by the unreliability of the tests, as this tends to obscure existing relations between scores. These disattenuated correlations are indicative of the correlations which would result when the tests used had perfect reliabilities. The results are presented in Table 3.

The observed correlations between PBT and KTS were low, with a slightly stronger correlation of the subscores compared to total scores. The same relation can be seen between PBT and SAQ. The correlations between KTS and SAQ were within the same range. However, correcting the scores for unreliability gave moderate to high disattenuated correlations between PBT and KTS, somewhat lower correlations between the PBT and SAQ, and even lower correlations between KTS and SAQ.

DISCUSSION

Although the results do show some small differences in mean scores between practising GPs and trainees, the overall results on the PBT and KTS indicate that competence in technical clinical skills (as measured by the KTS

or the PBT) shows no substantial differences. Only on the SAQ score do trainees and GPs differ consistently.

The proficiency in technical clinical skills seems to show little general improvement or deterioration during vocational training and thereafter, whereas the higher SAQ score associated with more advanced levels of training or experience most likely reflects a general self-attribution: as a result of experience GPs tend to feel more confident about their competence concerning technical clinical skills, without necessarily being more competent.

It has been difficult to demonstrate changes in scores on PBTs related to training or experience at postgraduate level, whereas these changes can easily be demonstrated on written tests (Norman *et al.* 1994). This questions the validity of the use of PBTs to discriminate between different degrees of expertise among GPs. However, as the scores on the KTS also showed no substantial differences related to experience, we believe the scores on the PBT reflect the absence of substantial differences of competence between groups with different training levels and experience.

The results of the reliability analyses were comparable with results in the literature, taking testing time into account (Van der Vleuten & Swanson 1990). Testing time was for all but one test too short to obtain reproducible scores. The high reliability of the SAQ reflects the strong influence of global self-attributions (Gordon 1991).

There was a positive correlation between knowledge of skills and proficiency on these skills. The existence of this specific relation is supported by the finding of a higher correlation, linking the subscores of the test. These findings indicate that a written knowledge test of skills can predict performance on these skills to some extent, if developed according to the same blueprint. This implies that a written test might be useful in situations where performance-based tests are difficult to apply, e.g. for screening purposes. The PBT could then be reserved for a (smaller) group identified to merit further evaluation, and thus a more efficient use of the PBT is achieved.

The correlation between self-assessment and proficiency in technical skills was moderate. Other studies reported low to absent correlations between self-assessment methods (Gordon 1991; Stillman *et al.* 1986, 1990). However, in contrast to the written test, the subscore of self-assessment showed only a slightly higher correlation with the PBT, suggesting that GPs have a rather general notion about their proficiency in technical clinical skills. It would be interesting to investigate whether a training programme in self-assessment could improve this skill (Gordon 1992).

In conclusion, while performance-based testing is obviously the best method to assess proficiency in hands-on skills, a written test can serve as a reasonable alternative in some situations, as it is relatively easy to administer and not very costly. Self-assessment, although positively correlated with performance, is a less viable alternative as it seems to reflect a general notion of competency.

Table 3 Correlations between (sub)scores of the performance-based test (PBT), knowledge test of skills (KTS) and the self-assessment questionnaire (SAQ)

	PBT	KTS	KTS sub	SAQ	SAQ sub
PBT		0·54	0·77	0·40	0·47
KTS	0·29[b]		1·00	0·37	0·40
KTS sub	0·33[b]	0·87[c]		0·38	0·49
SAQ	0·25[a]	0·29[b]	0·24[a]		1·00
SAQ sub	0·29[b]	0·31[b]	0·30[b]	0·96[c]	

Note: Observed correlations in lower triangle ([a]$P < 0{\cdot}05$, [b]$P < 0{\cdot}01$, [c]$< 0{\cdot}001$) and disattenuated correlations in upper triangle.

REFERENCES

Anderson M B & Kassebaum D G (1993) Proceedings of the AAMC's Consensus Conference on the Use of Standardized Patients in the Teaching and Evaluation of Clinical Skills. *Academic Medicine* **68**, 437–83.

Bender W, Hiemstra R J, Scherpbier A J J A & Zwierstra R P (eds) (1990) *Teaching and Assessing Clinical Competence.* BoekWerk Publications, Groningen.

Berg A O (1979) Does continuing medical education improve the quality of medical care? A look at the evidence. *Journal of Family Practice* **8**, 1171–4.

Cronbach L J, Gleser G C, Nanda H & Rajaratnam N (1972) *The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles.* John Wiley & Sons Inc, New York.

Fabb W E (1983) *The Assessment of Clinical Competence in General Family Practice.* MTP Press, Hingham, USA.

Fuhrmann B S & Weissburg M J (1978) Self-assessment. In: *Evaluating Clinical Competence in the Health Professions* (ed. by K M Morgan & D M Irby), pp. 139–50. CV Mosby, St Louis, MO.

Gordon M J (1991) A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine* **66**, 762–9.

Gordon M J (1992) Self-assessment programs and their implications for health professions training. *Academic Medicine* **67**, 672–9.

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.

Harden R M & Gleeson F A (1979) Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* **13,** 41–54.

Hart I R, Harden R M & Des Marchais J (eds) (1992) *Current Developments in Assessing Clinical Competence.* Canadian Health Publications, Montreal.

Hart I R, Harden R M & Walton H J (eds) (1987) *Further Developments in Assessing Clinical Competence.* Health Publications, Montreal.

Hart I R, Harden R M & Walton H J (eds) (1987) *Further Developments in Assessing Clinical Competence.* Canadian Health Publications, Montreal.

Lamberts H, Bouwer H & Mohrs J (1991) *Reason for Encounter-, Episode- and Process-Oriented Standard Output from the Transition Project.* Department of General Practice, University of Amsterdam, Amsterdam.

Newble D I & Swanson D B (1988) Psychometric characteristics of the objective structured clinical examination. *Medical Education* **23,** 325–34.

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.

Patrick J (1992) *Training: Research and Practice*, pp. 19–71. Academic Press, London.

Reznick R K, Smee S, Baumber J S *et al.* (1993) Guidelines for estimating real cost of an objective structured clinical examination. *Academic Medicine* **68,** 513–17.

Stillman P, Swanson D, Smee S *et al.* (1986) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine* **105,** 762–71.

Stillman P L, Regan M B, Swanson D B *et al.* (1990) An assessment of clinical skills of fourth-year students at four new England medical schools. *Academic Medicine* **65,** 320–326.

Van der Vleuten C P M & Swanson D B (1990) Assessment of skills with standardized patients: state of thc art. *Teaching and Learning in Medicine* **2,** 58–76.

Van der Vleuten C P M, van Luyk S J & Beckers H J M (1988) A written test as an alternative to performance testing. *Medical Education* **23,** 97–107.

Wooliscroft J O, TenHaken J, Smith J & Calhoun J G (1993) Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. *Academic Medicine* **68,** 285–94.

Received 15 *November* 1994; *editorial comments to authors* 30 *January* 1995; *accepted for publication 6 April 1995*