

## Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice

Paul Ram,<sup>1</sup> Cees van der Vleuten,<sup>2</sup> Jan-Joost Rethans,<sup>1</sup> Berna Schouten,<sup>1</sup> Sjoerd Hobma<sup>1</sup> & Richard Grol<sup>1</sup>

This study compares the predictive values of written-knowledge tests and a standardized multiple-station examination for the actual medical performance of general practitioners (GPs) in order to select effective assessment methods to be used in quality-improvement activities.

A comprehensive assessment was performed in four phases. First, 100 GPs from the southern part of the Netherlands were assessed by a general medical knowledge test and by a knowledge test on technical skills. Second, in order to check for time-order effects, participants were randomly divided into two groups of 50 each, comparable on scores of both knowledge tests and on professional characteristics. Finally, both groups went through a multiple station examination using standardized patients and a practice video assessment of real surgery, but in opposite orders. Consultations were videotaped and assessed by well-trained peer observers. The drop-out rate was 10%.

In both groups the predictive value of medical knowledge tests, ranging from 0.43 to 0.56 (Pearson correlation disattenuated), proved to be comparable with the predictive value of the multiple-station exam-

ination for actual performance (0.33–0.59). The overall explained variance of scores of the practice video assessment, measured by multiple regression analysis with performance scores as dependent variables and scores on the knowledge tests and the multiple-station examination as independent variables was moderate (19%). A time-order effect showed in only one direction: from practice video assessment to the multiple-station examination. The GP's professional characteristics did not contribute to the explanation of variation in performance. Medical knowledge tests can predict actual clinical performance to the same extent as a multiple-station examination. Compared with a station examination, a knowledge test may be a good alternative method for assessment the procedures of a large number of practising GPs.

*Keywords* Clinical competence; education, measurement, \*methods; education, medical, continuing, \*methods; family practice, \*education.

*Medical Education* 1999;33:197–203

### Introduction

After completing a vocational training scheme, a general practitioner (GP) is supposed to have reached a level of competence which should reflect the demands of diagnostic, patient management and communication skills in daily practice. There is growing consensus that this level of competence does not imply the same level of actual performance in day-to-day practice during the

rest of the GP's professional life<sup>1,2</sup>. Performance may change due to several inter-related factors such as changes in medical knowledge, practice organization or age<sup>3–6</sup>. To sustain an acceptable level of performance GPs should attend postgraduate courses. Although the time investment for these courses may be set, the choice of topics usually depends on the individual GP and his or her subjective needs. Subjective selection of topics in continuing medical education (CME) is not very effective in changing the practice behaviour of doctors<sup>7</sup>. If deficiencies in practice performance could be objectively identified, CME and quality improvement could be better matched to the real needs in order to maximize effectiveness<sup>8</sup>. Therefore, an assessment procedure is needed that is both valid and reliable in revealing deficiencies in the performance of GPs and

<sup>1</sup>Centre for Quality Research, Universities of Maastricht and Nijmegen, The Netherlands,

<sup>2</sup>Department of Educational Development and Research, Maastricht University, The Netherlands

*Correspondence:* Paul Ram, MD, Maastricht University, Department of General Practice, PO Box 616, 6200 MD Maastricht, The Netherlands

appropriate for use on a large scale. Knowledge tests such as multiple choice questions (MCQs) are efficient in handling large numbers of GPs and can easily cover a wide range of subjects<sup>5</sup>. On the other hand, critics raise doubts about the validity and acceptability of MCQs, claiming that selecting options from a list of alternatives is quite different from clinical reality<sup>9,10</sup>. Moreover, these methods mainly assess aspects of competence, i.e. what a doctor is capable of doing, and this may not be the same as what a doctor actually does in his day-to-day practice (performance), the ultimate focus of quality improvement<sup>11</sup>. In this perspective the question arises whether results on written tests do predict actual performance in practice. Different studies have shown a moderate positive relationship between written tests and performance measurements<sup>12-14</sup>. Performance in these studies was assessed either indirectly by chart review and chart-stimulated recall, or directly by peer rating by colleagues who worked with the observed colleague. Assessment by using medical records has its limitations, since clinical data are often incomplete and inconsistent<sup>15</sup>. Observation by co-operating colleagues may be quite subjective and is not useful in single-handed practices. Direct observation using trained (peer) observers, not co-operating with the observed physician, may be more valid and reliable. However, assessment in daily practice is difficult to standardize<sup>16</sup>. Better controlled competence-based tests closely linked to professional reality, called multiple-station examination or objective structured clinical examination (OSCE), could be alternatives<sup>17,18</sup>. In these methods clinical skills must be demonstrated in standardized simulated clinical situations using direct (video) observation to score performance. These tests have more authenticity than knowledge tests and include aspects of problem solving, patient management and attitude. Therefore, these are assumed to have better predictive value for actual performance than written tests. However, implementation on a large scale may be more difficult, because of the high resource requirements and the organizational complexity<sup>19</sup>. Moreover, research on the predictive value of these methods has resulted in conflicting evidence<sup>2,20</sup>. In all, more insight is needed into the value of both knowledge testing and station examinations to predict performance in regular clinical practice.

The study presented here compares the predictive value of written-knowledge tests (less valid, feasible) with the predictive value of direct observation in a multiple-station examination (more valid, less feasible) for medical performance in daily practice.

## Methods

### Subjects

All general practitioners in the south of the Netherlands were invited by letter to participate in this intensive study and 220 (25%) reacted positively. For budgetary reasons 135 were randomly selected, 100 to be assessed, 35 to participate as peer observers. Data on personal and professional characteristics (age, gender, single-handed, being a medical teacher, working full-time or part-time, urbanization area) were collected.

### Instruments and procedure

Knowledge tests and observation instruments were screened on their psychometric qualities. Two instruments with established validity and reliability were included. These were a general medical knowledge test and a scoring list for direct observation of medical performance and communication with patients, the MAAS-Global<sup>20,21</sup>. The general medical knowledge test (Cronbach's alpha = 0.64) consists of 60 short cases with 131 items about complaints and diseases, stratified according to chapters of *International Classification of Primary Care*<sup>22</sup>. Using the blueprint of the general medical knowledge test, a knowledge test on medical technical skills was constructed (124 items, Cronbach's alpha = 0.60) in order to cover the broad cognitive domain of general practice as completely as possible<sup>23</sup>. This knowledge test on skills is focused on knowledge pertaining to why and how a certain skill should be performed, the skills being used in daily practice and selected from nationally accepted job description and guidelines for general practice<sup>24</sup>. Answers on both knowledge tests were given in a true/false/? format. Examples of items of both knowledge tests are shown in Table 1.

Participants were assessed in three sequential steps (see Fig. 1). First, the two knowledge tests, both paper-and-pencil tests, were sent to the participants. After the knowledge tests had been completed, a randomization was carried out in order to check for order effects caused by the design. Two groups of 50 GPs each, comparable on demographic characteristics and on scores of the knowledge tests, were formed. In steps two and three both groups were observed directly. Group 1 went first through a multiple-station examination, a real life-like surgery at a skills laboratory of the medical school, using eight stations with cases presented by standardized patients. All consultations were

**Table 1** Example cases: with subsequent questions from the general medical knowledge test (1) and knowledge test on technical skills (2)

---

1. Mrs. Cleveland, 75-year-old, is known with an advanced arthrosis of her left knee. She enters the room, walking slowly and leaning to the right on her stick. The left knee is swollen and warm.  
There is a hydrops.  
Among the therapeutical measures, adequate at this moment, is/are:

- application of icepacks true/false/?
- massage and exercises true/false/?
- intra-articular injection of corticosteroids true/false/?
- holding the stick in the other hand (left) true/false/?

Literature: Linden AJ van der, Claessens H. Leerboek orthopaedie 1995.

2. 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.

- The correct extent of hyperextension of the head is LESS with an infant compared to an adult. true/false/?

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

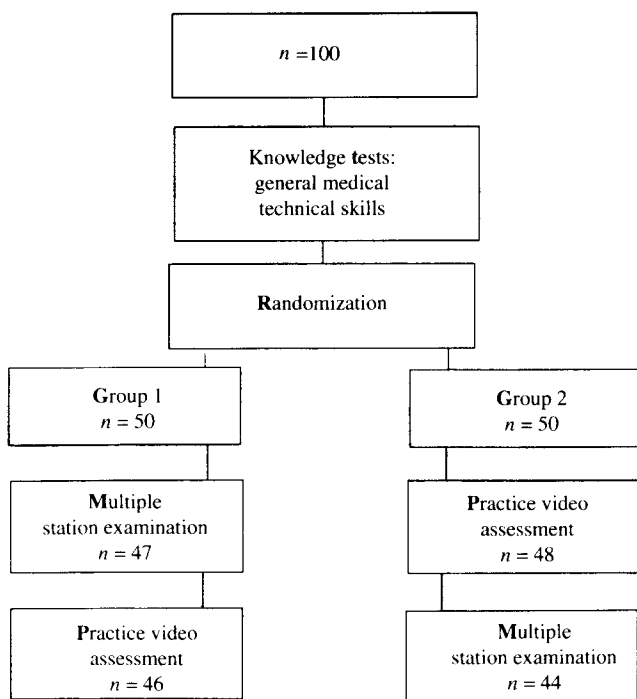
- This is a correct procedure for insufflation of an infant. true/false/?

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

- This is a correct rate for infants. true/false/?
- During resuscitation of an adult the adequate rate of compressions is closer to 80 per minute than to 60 per minute. true/false/?
- During resuscitation with one resuscitator the recommended schedule is:15 compressions followed by 1 insufflation. true/false/?

Literature: Leerboek elementaire reanimatie. Den Haag: Nederlandse Hartstichting 1996

---



**Figure 1** Study design.

videotaped. A few months later these GPs went through the practice video assessment of regular surgeries in daily practice. Group 2 started with the practice video assessment and this was followed by the station examination some months later. In the practice video assessment, medical performance including physical ex-

amination (and communication with patients) was recorded during 1 week with the patient's consent and using two cameras. By means of 10 pre-set criteria a representative sample of 16 consultations was selected from a logbook completed by the GPs.<sup>25</sup> In both tests medical performance was assessed by trained peer observers using the MAAS-Global scoring list<sup>21</sup>. Validity of both tests proved to be adequate in previous research.<sup>26</sup> Reliability, i.e. generalizability coefficients<sup>27</sup>, of the scores in the station examination was 0.59 and of the scores in the practice video assessment 0.83.

**Analysis**

The scores on the knowledge tests were calculated as the sum of correct minus incorrect scores (correction for guessing). Case scores in the multiple-station examination and practice video assessment were calculated by summing the number of item responses for each case. All scores were expressed as percentages of the maximum score. The test score was the mean score across cases. The predictive value of both written knowledge tests and the multiple-station examination for actual performance in the practice video assessment was assessed by calculating the bivariate correlations (Pearson, observed and corrected for attenuation) between these methods. Confounding of the predictive values by GPs' characteristics and by a time-order effect was assessed by using a multiple regression analysis with the practice video assessment scores as dependent

variables and scores on the knowledge tests, scores on the multiple station examination, professional characteristics and possible time-order effects as independent variables. In this way the percentages of explained variance in the practice scores were calculated ( $R$  and  $R^2$ ). Time-order effects were assessed by calculating the significance level of differences between scores of the multiple-station examination and the scores of the practice video assessment, between and within both groups (paired and unpaired  $t$ -tests).

## Results

### Drop out

In total 10 participants (10%) dropped out arguing that 'video observation is too threatening' (1%), 'on holiday' (4%) and 'too busy to participate' (5%). The study population ( $n = 90$ ) was representative of Dutch GPs for professional and personal variables, except for age (mean 43.1, SD 5.5; Dutch GPs mean 44.3, SD 7.4.  $t$ -test:  $t = 2.03$ ; degrees of freedom (d.f.) = 89;  $P < 0.05$ ).

### Scores and differences in scores, within and between both groups

Table 2 shows descriptive statistics for scores of the knowledge tests and multiple-station examination on medical competence and for scores of practice video assessment on medical performance, for groups 1 and 2. Scores on the general medical knowledge test were significantly higher than the scores on the knowledge test on technical skills ( $t = 21.5$ ; d.f. = 89;  $P < 0.01$ ; paired  $t$ -test). Time-order effects are revealed within

group 2 where scores between the multiple-station examination and practice video assessment differed significantly ( $t = 8.2$ ; d.f. = 43;  $P < 0.01$ ; paired  $t$ -test). In the multiple-station examination, group 2 scored significantly higher than group 1 ( $t = 3.7$ ; d.f. = 89;  $P < 0.05$ ; unpaired  $t$ -test). No significant difference was found between groups for scores of the practice video assessment ( $P < 0.05$ ). Apparently the scores of the multiple-station examination increased significantly when a practice video assessment was done beforehand (group 2), whereas practice scores appeared to be unaffected by previous assessment experiences. This time-order effect was therefore taken into account as an independent variable in regression analysis.

### Predictive values

Table 3 shows the correlations (observed and disattenuated with their significance levels) of scores of knowledge tests and scores of the multiple-station examination with scores of the practice video assessment. In both groups the two medical knowledge tests correlate significantly with medical performance in practice. These correlations are comparable with the correlation between scores of the multiple station examination and scores of the practice video assessment.

Table 4 shows the results of multiple regression analysis with practice video assessment scores as dependent variables and the scores of both knowledge tests and multiple-station examination, time-order effect (group 2) and professional characteristics as independent variables. The multiple  $R$  and  $R^2$  are given, representing percentages of variance explained by the variables in equation. Overall, the explained variance is

**Table 2** Scores knowledge tests, multiple-station examination and practice video assessment in percentages and differences of scores between and within both groups

	Number of items	Group 1 ( $n = 46$ )		Group 2 ( $n = 44$ )	
		Mean (SD)	Range	Mean (SD)	Range
Knowledge tests					
general medical	131	52.7 (11.5)	24.4–80.9	50.8 (11.0)	31.3–72.5
technical skills	124	31.1 (10.9)	6.5–54.0	31.4 (12.5)	6.5–54.8
Multiple-station examination	8 cases	64.4 (8.6) <sup>1</sup>	49.1–83.0	70.1 (5.9) <sup>1,2</sup>	55.7–84.4
Practice video assessment	16 cases	62.8 (7.6)	47.1–84.1	61.3 (6.2) <sup>2</sup>	47.6–75.8

<sup>1</sup> significant difference ( $P < 0.001$ ) between groups on same measure (multiple station examination).

<sup>2</sup> significant difference ( $P < 0.001$ ) within group 2 between different measures (station – practice).

**Table 3** Observed and disattenuated Pearson correlations between knowledge tests and practice video assessment (PVA), and between multiple-station examination and practice video assessment

	PVA group 1 (n = 46)		PVA group 2 (n = 44)	
	Observed	Disattenuated	Observed	Disattenuated
Knowledge tests				
General medical	0.41 <sup>1</sup>	0.57 <sup>1</sup>	0.33 <sup>2</sup>	0.46 <sup>1</sup>
Technical skills	0.41 <sup>1</sup>	0.59 <sup>1</sup>	0.32 <sup>2</sup>	0.46 <sup>1</sup>
Multiple-station examination	0.41 <sup>1</sup>	0.59 <sup>1</sup>	0.23	0.33 <sup>2</sup>

<sup>1</sup>  $P < 0.01$ .<sup>2</sup>  $P < 0.05$ .**Table 4** Stepwise regression analyses: practice video assessment (PVA) scores as dependent variables, respective test scores and scores on professional characteristics as independent variables

	Beta	t	R	R <sup>2</sup> (adjusted)
General medical knowledge test	0.26	2.61 <sup>1</sup>	0.37	0.13 (0.12)
Station examination	0.33	3.10 <sup>1</sup>	0.43	0.19 (0.17)
Time-order effect	-0.23	-2.19 <sup>2</sup>	0.48	0.23 (0.20)

<sup>1</sup>  $P < 0.01$ <sup>2</sup>  $P < 0.05$ 

moderate. The general medical knowledge test contributes to the explained variance more than the multiple-station examination does. Personal and professional characteristics of the GPs did not contribute to explained variance.

## Discussion

The aim of this study was to compare the predictive value of knowledge tests with the predictive value of a multiple-station examination for actual medical performance, assessed by video observation in GP's day-to-day practice. We realize that our study population is not representative, since the GPs participated voluntarily. However, since our study is a methodological study this non-representativeness is not a major problem. In addition, the predictive values of the different instruments for actual performance were studied within the same study population.

Both the general medical knowledge test and the knowledge test on skills proved to predict actual medical performance to the same extent as the multiple-station examination. These findings contrast with the hypothesis that competence-based tests using direct

observation, such as multiple station examinations and OSCEs, will have a stronger relationship with actual performance than knowledge tests. These findings may be explained by the reported 'audience-effect', i.e. the influence of observation on performance, in the multiple-station examination.<sup>26</sup> In a questionnaire, taken after the study, a majority of GPs reported they felt inhibited by the observation throughout the entire station examination, judged as an artificial and unfamiliar setting, whereas a minority said they were influenced in the practice video assessment. As a consequence, a majority of GPs judged the videotaped consultations of daily surgery as 'natural'. In the videotaped consultations of daily surgeries they recognized their normal 'working style' better than in the standardized station consultations. This audience effect may lead to a special variation in the GP's performance in each case of the multiple-station examination, which may disturb the correlation with actual performance in daily practice. Compared with these effects in the multiple-station examinations, written tests and observation in daily practice may be less intrusive.

The knowledge tests were sent to the participants, which could provide scope for cheating. However, the

assessment was purely educative and looking up things would minimize the educational value, as the participants was told. Moreover, cheating would have had a decreasing effect on the relationship between knowledge and performance. Cheating GPs probably would have a high knowledge test score and lower performance scores, having less ready knowledge available.

At variance with knowledge tests, GPs' (professional) characteristics did not contribute to the explanation of variation in performance scores. This suggests that knowledge tests are much better predictors for strengths and weaknesses in actual performance for groups of GPs than GPs' characteristics such as age, gender, being single-handed or being a GP-trainer. Since well-developed knowledge tests are available for screening purposes, the current system of postgraduate medical education, allowing GPs to choose on subjective needs, is highly questionable and should become a serious matter of debate.

Finally, we conclude that medical knowledge tests should be developed for use in assessment of practising GPs. Compared with multiple station examinations and OSCEs, knowledge tests can be used on a broad scale with fewer logistical problems. In addition, these predict actual performance relatively well. Nevertheless, the explained variance in actual performance by knowledge tests and station examination reported here, is too low to bridge the gap between competence and actual performance assessment. In assessment of practising GPs a combination of different methods, including observation in daily practice, is probably the most valid and reliable approach.

## References

- 1 Norman G. Can an examination predict competence? The role of recertification in maintenance of competence. *Annals RCPSC* 1991;24:121-4.
- 2 Rethans JJ, van Leeuwen Y, Drop R, van der Vleuten C. Competence and performance: two different concepts in the assessment of quality of medical care. *Family Pract* 1990;7:168-74.
- 3 Day SC, Norcini JJ, Webster GD. The effect of changes in medical knowledge on examination performance at the time of recertification. In: *Proceedings of the 27th Annual Conference on Research in Medical Education*. Chicago: Association of American Medical Colleges; 1988;139-44.
- 4 Lockyer JM. Physician performance: the roles of knowledge, skill, and environment. *Teaching and Learning in Medicine* 1992;4:86-96.
- 5 van Leeuwen YD, Pollemans MC, Mol SS, Eekhof JAH, Grol R, Drop MJ. The Dutch knowledge test for general practice: issue of validity. *Eur J Gen Pract* 1995;1:113-7.
- 6 Caulford PG, Lamb SB, Kaigas TB, Hanna E, Norman GR, Davis DA. Physician incompetence: specific problems and predictors. *Acad Med* 1994;69 (Suppl.):16-20.
- 7 Sibley JC, Sackett DL, Neufeld V, Gerrard B, Rudnick KV, Fraser W A randomized trial of continuing medical education. *New Engl J Med* 1982;306:511-5.
- 8 Grol R. Research and development in quality of care: establishing the research agenda. *Qual Health Care* 1996;5:1-8.
- 9 Newble DI, Baxter A, Elmslie G. A comparison of multiple choice and free response tests in examinations of clinical competence. *Med Educ* 1979;13:263-8.
- 10 McGuire C. Perspectives in assessment at >. *Acad Med* 1993;68 (Suppl.):S3-8.
- 11 Rethans JJ, Sturmans F, Drop MJ, van der Vleuten CPM, Hobus P. Does competence of general practitioners predict their performance? *BMJ* 1991;303:1377-80.
- 12 Norman GR, Davis DA, Painvin A, Rath D, Ragbeer M. Comprehensive assessment of clinical competence of family-general physicians using multiple measures. In: Bender W, Hiemstra R, Scherpbier A, Zwiestra (eds) *Teaching and assessing clinical competence*. Groningen: Boekwerk Publ. 1990; 357-364.
- 13 Page GG, Fielding DW. Performance on PMP's and performance in practice: are they related? *J Med Educ* 1980;55:529-37.
- 14 Ramsey PG, Carline JD, Inui YS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Int Med* 1989;110:719-26.
- 15 Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual performance? A study using simulated patients. *Brit J Gen Pract* 1994;44:153-6.
- 16 Kane MT. The assessment of professional competence. *Evaluation and the Health Professions* 1992;15:163-82.
- 17 Harden R, Gleeson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13:41-54.
- 18 Anderson MB, Kassebaum DG. Proceedings of the AAMC's consensus conference of the use of standardized patients in the teaching and evaluation of clinical skills. *Acad Med* 1993;68:437-83.
- 19 Reznick RK, Smee S, Baumber JS. Guidelines for estimating real cost of an objective structured clinical examination. *Acad Med* 1993;68:513-7.
- 20 Pollemans M. Kennistoetsing bij huisartsen. [Thesis]. [Assessment of knowledge of general practitioners]. Rijksuniversiteit Limburg; 1994
- 21 van Thiel J, van der Vleuten CPM, Kraan H. Assessment of medical interviewing skills: generalizability of scores using successive MAAS-versions. In: Harden RM, Hart IR & Mulholland H, editors. *Approaches to the Assessment of Clinical Competence* 536-541. Centre for Medical Education 1992.
- 22 Lamberts H, Wood M. (eds) *International Classification of Primary Care*. Oxford University Press; 1987.

- 23 van Leeuwen YD. Growth in knowledge of trainees in general practice, Figures on facts. [Thesis] Universitaire Pers; Maastricht. 1995.
- 24 Jansen JJM, Tan LHC, van der Vleuten CPM, van Luijk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. *Med Educ* 1995;**29**:247-53.
- 25 Ram P, Grol R, Rethans JJ, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ* in press.
- 26 Ram P, van der Vleuten C, Rethans JJ, Aretz K, Grol R. Assessment of practising general practitioners: comparison of video observation in a station exam using standardized patients with video observation of real surgery in daily practice. *Acad Med* 1999; in press.
- 27 Brennan RL. *Elements of Generalizability Theory*. Iowa City: American College Testing Publications; 1983.

*Received 18 December 1997; editorial comments to authors 30 March 1998; accepted for publication 1 April 1998*