

# Assessment of Practicing Family Physicians: Comparison of Observation in a Multiple-station Examination Using Standardized Patients with Observation of Consultations in Daily Practice

Paul Ram, MD, Cees van der Vleuten, MD, Jan-Joost Rethans, MD, PhD, Richard Grol, PhD,  
and Karin Aretz

## ABSTRACT

**Purpose.** Looking for a valid, reliable, and feasible method to collect data on the performances of practicing family physicians, the authors compare the measurement characteristics of a multiple-station examination (MSE) using standardized patients with those of a video assessment of regular consultations in daily practice (practice video assessment, PVA).

**Method.** In a cross-sectional study, consultations of 90 family physicians were videotaped both in an MSE and in their daily practices. Peer-observers used a validated instrument (MAAS-Global) to assess the physicians' communication with patients and their medical performances. The physicians were randomly divided into two groups, comparable for demographic characteristics, and half underwent the assessments in reverse order to test for time-order effects. Content validity, criterion validity, reliability, and feasibility of the two methods were compared.

**Results.** Content validity of the PVA was superior to that of the MSE, since the domain of general family prac-

tice care was better covered. Observed participants judged the videotaped practice consultations to be "natural," whereas hardly any family physician, after reviewing the videotaped consultations of the MSE, recognized his or her usual working style. Specific criteria made it possible to standardize real practice. Concerning criterion validity, only the medical-performance components of the two methods correlated. No correlation was found for the communication components. Real-practice performance proved to be less influenced by observation than was performance during the MSE. The reliabilities of the two methods, expected to be better in the controlled MSE, were comparable. The administration of the PVA was more flexible, less costly, and better accepted by the family physicians than was that of the MSE.

**Conclusion.** Assessment for quality improvement of family physicians' practices by video observation in daily practice is superior to video assessment in a simulated setting using standardized patients.

*Acad. Med.* 1999;74:62-69.

*Dr. Ram is a family physician and staff member, Department of Family Medicine, Maastricht University, Maastricht, The Netherlands. Dr. van der Vleuten is professor in educational research, Department of Educational Development and Research, Maastricht University. Dr. Rethans is a family physician and educational researcher, Dr. Grol is professor in quality improvement, and Ms. Aretz is research assistant, all at the Center for Quality of Care Research, University of Maastricht and University of Nijmegen, Nijmegen, The Netherlands.*

*Correspondence and requests for reprints should be addressed to Dr. Ram, Department of General Practice, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands; e-mail: <Paul.Ram@HAG.Unimaas.NL>.*

Improving the quality of care that medical professionals, such as family physicians, provide their patients demands the regular collection and evaluation of data on physicians' levels of competence (what they are capable of doing) and performances (what they do in their day-to-day practices). By assessing whether competence and performance meet quality criteria, educators can tailor activities to correct the physicians' real needs and deficiencies.<sup>1,2</sup>

Competence and performance should be assessed by valid, reliable, and feasible methods.<sup>3</sup> Objective assessment is preferred to self-assessment; studies have shown that physicians who rate their own performances usually overestimate the quality of their service delivery, and their self-selected postgraduate education changes their practice behaviors very little.<sup>4,5</sup> In addition, to be valid and reliable, assessment should both approximate real professional practice as closely as possible and main-

tain standardized test-taking conditions.<sup>6</sup>

Assessment of doctors' performances by direct observation in their offices is assumed to have high validity.<sup>7,8</sup> However, this "in vivo" assessment makes it difficult to achieve even modest levels of standardization, thus affecting reliability.<sup>9</sup> As regards validity, several factors may decrease the representativeness of the sample and hamper generalizations: logistics may limit the selection of doctor-patient consultations for assessment, patients may not consent to being observed, and observers may influence the family physicians' performances.<sup>10</sup>

An alternative is "in vitro" assessment. Well-developed methods of assessment using direct observation in simulated situations have been proposed in recent decades, including the objective structured clinical examination (OSCE) and the standardized patient (SP)-based test.<sup>11,12</sup> These methods simulate clinical situations in so-called "stations." The OSCE usually focuses on parts of an encounter, whereas the SP-based test typically uses standardized patients to simulate an entire encounter, each station focusing on an integrated approach of different reasons for the encounter.<sup>13</sup> Unlike non-standardized "in vivo" assessments, these "multiple-station examinations" offer the advantage of reliability by using uniform cases that can be directly observed and rated with objective checklists and rating scales. Studies into whether such a competence-based method can predict actual performance have resulted in conflicting evidence.<sup>14,15</sup>

To further investigate the strengths and weaknesses of competence and performance assessment, we compared "in vitro" video observation of family physicians in a station examination using standardized patients with "in vivo" video observation of the same physicians' consultations with their patients in daily practice. This experimental

study focused on validity, reliability, and feasibility to select the more efficient method for assessing the quality of family physicians.

## METHOD

### Subjects

We invited, by letter, all family physicians in the south of The Netherlands to participate in this study; 220 (25%) accepted our invitation. For budgetary reasons, we randomly selected 135 family physicians: 100 of them to be observed, 35 to serve as peer-observers. We compared the personal and professional characteristics of the participants with those of the national population of family physicians (using the *t* test and chi-square test) to determine their representativeness.

### Procedure and Instruments

To make the multiple-station examination (MSE) and practice video assessment (PVA) comparable, we designed the MSE to be as similar to real practice as possible and we standardized the PVA.

To create the MSE, we first developed a blueprint that reflected complaints common in general practice.<sup>16,17</sup> Using this blueprint, a panel of seven family physicians then wrote eight standardized cases that represented both acute and chronic diseases. Three of the cases contained multiple problems. The cases (with their International Classification of Primary Care (ICPC) chapters) were: non-insulin-dependent diabetes mellitus (T); aspecific stomach pain (D); asthma induced by exercise (R); a sore throat and partial deafness (eustachian salpingitis) (R + H); shoulder complaints (painful arc) (L); sleep disturbances and fatigue (P + B); angina pectoris and peripheral vascular disease (K); and headache caused by refractive disease (presbyopia) (F).

Checklists for scoring the participating physicians were derived from national guidelines for appropriate clinical performance.<sup>18</sup> Rooms were equipped as real consultation rooms and the participants were allowed to bring in their doctor's bags. The doctors remained in their rooms while eight well-trained standardized patients rotated from one doctor to another. The study's "time-pressure design" allowed the doctors to spend just 12 minutes with each patient (96 minutes in total). We instructed the participants to perform as naturally as possible and videotaped all encounters.

For the PVA, during one week of the study, we videotaped the participants' regular practice-site consultations. A video observation system, consisting of two cameras with built-in microphones, a monitor, and a recorder, was installed by trained electricians in each practice. All consultations were recorded both in the consulting room and in the examination room. The family physician was responsible for switching from one camera to the other when necessary and for logging, immediately after the consultation, patient and consultation data (such as patient's age and sex, number and nature of the complaints, and duration of consultation). Before the consultation, the receptionist informed the patient about the video recording and gave the patient two colored cards (green = consent, red = no consent). The patient then handed the appropriate card to the family physician at the start of the consultation. Video registrations were erased if patients revoked their permissions.

Since a practice video assessment is less standardized than is a multiple-station examination, we felt that more cases were necessary for analysis of the PVA. Following a blueprint similar to the one we used to develop the MSE (based on prevalences of diseases in general practice and on the national family physician task description<sup>16,17,19</sup>), we formulated ten criteria, which we

then used to select 16 consultations from each physician's logbook. (This process is described in greater detail elsewhere.<sup>20</sup>) These criteria were:

- The consultations had to represent at least eight different ICPC chapters.
- The consultations had to include representation of the five ICPC chapters with the highest prevalences (D = digestive system, K = cardiovascular system, L = locomotor system, R = respiratory system, and S = skin)
- The first five videotaped consultations would not be chosen.
- At least eight of the consultations had to involve complaints that have accepted treatment guidelines.
- At least 14 of the 16 consultations had to be between five and 15 minutes long.
- The patients in the consultations had to have an age range of under 18 to over 65.
- There had to be between six and ten women among the patients in the consultations.
- At least eight of the consultations had to be initial consultations; at least four had to be follow-ups.
- No more than two of the consultations could involve patients with more than two reasons for seeing the doctor.
- No more than two of the consultations could involve patients with only psychosocial problems.

To check for any time-order effect, we randomly divided the 100 participants into two groups of 50 (groups 1 and 2), with comparable personal and professional characteristics. The two groups went through the two test settings—MSE and PVA—in reverse order.

The 35 peer-observers, trained in four sessions, assessed the video recordings of the MSE and PVA consultations for quality of communication and medical performance. To avoid a "halo effect,"<sup>18</sup> each observer scored only four

MSE and six PVA consultations of a single doctor. To study generalizability for reliability analysis, 25% of all cases were observed independently by two observers. The observers assessed the consultations using the MAAS-Global (Maastricht Anamnestic and Advice Scoring list, Global),<sup>21</sup> a validated rating scale that contains 12 items on communication and four items on medical performance. The items are case-independent, scored on a seven-point Likert scale, and anchored on an extensive list of detailed criteria.<sup>22</sup> With regard to communication, seven items refer to successive phases in a consultation: entry, follow-up, exploration of patients' needs, communication about the physical examination, evaluation and diagnosis, communication on the management plan, and evaluation of the consultation. The five remaining items refer to general communication skills: providing information, exploring emotions, summarizing, ordering, and empathizing. The items concerning medical performance refer to history taking, physical

examination, evaluation and diagnosis, and management plan.

Based on the blueprints described above, we assessed content validity on the following aspects (see Table 1): the time-order effects for both MSE and PVA, the representativeness of the selected PVA cases, and the influences that differences in sample characteristics had on the PVA scores (such as number and content of complaints, age and sex of patients, and duration and type of consultations). Furthermore, we distributed a questionnaire (13 items, five-point Likert scale) to the observed participants to assess whether they found the MSE cases authentic, whether they felt that being observed influenced their performances (the "audience effect"), and whether they recognized their own usual working styles. Criterion validity was defined as the ability of the physician's performance on the MSE to predict his or her performance on the PVA.

Reliability was defined and analyzed as generalizability. We assessed the par-

**Table 1**

Methods Used to Compare the Psychometric Characteristics of a Multiple-station Examination (MSE) and a Practice Video Assessment (PVA), The Netherlands, 1996.		
Characteristic	MSE	PVA
<b>Validity</b>		
Content validity		
Content of cases selected	Selection criteria: prevalence/guidelines	Selection criteria: prevalence/guidelines
Time-order effect	<i>t</i> test	<i>t</i> test
Representativeness of cases	By definition	Statistics: % sample criteria fully met
Influence sample differences on scores	Not applicable (equal samples)	Regression analysis
Authenticity of cases	Questionnaire	Not applicable (implicit)
Audience effect	Questionnaire	Questionnaire
Recognition of working style	Questionnaire	Questionnaire
Criterion validity	Pearson correlation	Pearson correlation
Reliability	Generalizability analysis	Generalizability analysis
<b>Feasibility</b>		
Acceptance	Questionnaire	Questionnaire
Cost	Keeping accounts	Keeping accounts

Table 2

	Physician Group 1*		Physician Group 2*	
	MSE Score Mean (SD)	PVA Score Mean (SD)	MSE Score Mean (SD)	PVA Score Mean (SD)
Communication score	44.4 (10.3)	41.3 (7.7)	52.4 (8.6)	42.1 (7.1)
Medical performance score	64.9 (8.5)	62.8 (7.6)	70.7 (6.4)	61.3 (6.2)
Overall score	54.7 (8.4)	52.1 (7.1)	62.0 (6.6)	51.7 (6.1)

\*Physicians in Group 1 first took the MSE and then the PVA; physicians in Group 2 took the examinations in the opposite order.

participating physicians' acceptance of both procedures with a "feasibility and preference" questionnaire. Using a five-point Likert scale, the physicians answered 15 questions concerning the effects of the assessment procedures on their practice organization and one question about which of the two methods—MSE or PVA—they would prefer to see used in the future to assess both communication and medical performance.

#### Analysis

By summing the scores on each item for each case, we calculated the physicians' overall case scores, expressed as percentages (maximum 100%). The test score was the mean score across cases. These scores were also calculated separately for communication and medical performance (Table 2).

To study validity (Table 1), we analyzed the time-order effects of the design by calculating the significance of differences between the MSE scores and the PVA scores, within and between both groups (paired and unpaired *t* tests). To assess the representativeness of selected cases in the PVA, we calculated the percentages of samples that met the predetermined criteria. To measure the influence of

differences in PVA sample characteristics on performance scores, we performed a multiple regression analysis with the practice scores as dependent and the sample characteristics as independent variables. The MSE cases' authenticity and both the MSE and PVA's audience effects and recognition of usual working style were assessed by calculating the percentages of family physicians agreeing. Finally, we analyzed the criterion validity by calculating the observed correlations (Pearson) between the MSE and PVA. Those correlations were corrected for attenuation.

Generalizability theory was used to estimate reliability, calculated on the data set, which had multiple ratings per consultation.<sup>27,28</sup> In the MSE, different raters (peer-observers) assessed each case, and each case was similar for all doctors; therefore we used a "raters-nested-within-cases-crossed-with-persons" ANOVA. In the PVA, different raters assessed each case and, since each case, being a regular consultation, was unique to each doctor, we used a "raters-nested-within-cases-within-persons" ANOVA. These designs allowed both concurrent estimation of reliability of raters and cases and estimated projections of reliability from the actual number of cases and raters used in the

study to other hypothetical test conditions with different sample sizes.

To judge feasibility, we calculated descriptive statistics of the "feasibility and preference" questionnaire. Costs of both tests were registered by keeping accounts.

## RESULTS

In total ten participants (10%) dropped out because participation was "too threatening" (1%), they "went on holiday" (4%), or they were "too busy to participate" (5%). On demographic variables, the 90 remaining physicians were representative of Dutch family physicians, except with respect to age. Dutch family physicians had a mean age of 44.3 years  $\pm$  5.5; the physicians participating in the study had a mean age of 43.1  $\pm$  7.4 (unpaired *t* test:  $t = 2.03$ ;  $df = 89$ ;  $p < .05$ ).

#### Validity

Table 2 reports the communication, medical performance, and overall scores on the MSE and PVA. As regards time—order effects, within group 1, no significant difference was found between MSE and PVA scores. Within group 2, the scores on the MSE and the PVA differed significantly, for both communication and medical performance ( $p < .001$ ). Comparing the MSE scores of the two groups, we found that group 2 scored significantly higher than did group 1 on both communication and medical performance ( $p < .001$ ). No significant difference was found between the groups for the PVA scores. Apparently, scores on the MSE increased significantly when the PVA had been done previously (group 2), whereas PVA scores appeared to be consistent, irrespective of whether the MSE had been done before or not. In both groups, the scores for communication were significantly lower than were the scores for medical performance, in

**Table 3**

<b>Percentages of 85 Observed Physicians Who Recognized Their Own Working Style in Communication and Medical Performance and Who Thought That Observation Influenced Their Performances on a Multiple-station Examination (MSE) and a Practice Video Assessment (PVA), The Netherlands, 1996</b>				
	Physician Group 1*		Physician Group 2*	
	MSE	PVA	MSE	PVA
Recognition of communication style	6.3	93.7	16.7	83.3
Recognition of medical performance	6.3	93.7	12.2	87.8
Observation influenced performance	75.0	35.5	50.0	22.0
Observation did not influence performance	25.0	64.5	50.0	78.0

\*Physicians in Group 1 first took the MSE and then the PVA; physicians in Group 2 took the examinations in the opposite order.

both the MSE and PVA (all comparisons  $p < .001$ ).

When judging the authenticity of the eight MSE cases, 94% of the observed physicians found them authentic, 3% not authentic, and 3% expressed neutrality. In PVA the domain of general practice care was covered extensively, since at least eight different ICPC chapters (criterion 1) were represented in every sample of all participants and other criteria such as heterogeneity in age and sex of patients were met in the majority (> 70%) of all samples.<sup>24</sup> The sample characteristic, "total duration of 16 consultations," had a mean of  $156.62 \pm 19.62$  minutes and showed a significant relationship with PVA scores: the longer the duration, the higher the performances for both communication ( $\beta = 0.351$ ;  $p < .001$ ) and medical performance ( $\beta = 0.356$ ;  $p < .001$ ).

Table 3 reports data regarding the influences of observation on perceived performance and on recognition of usual working style, judged by the physicians after viewing the videotapes of their own performances. Comparing the two tests, we found that significantly more family physicians in both

groups 1 and 2 judged that they had been influenced by the observation in the MSE. This difference was significantly higher in group 1 (MSE first, PVA later) than in group 2 (40% and 28%, respectively;  $p < .001$ ).

Table 4 gives the observed and disattenuated Pearson correlations between the MSE and PVA for the communication, medical performance, and overall scores (separately for groups 1 and 2). For medical performance, we found a significant correlation between the scores on the MSE and PVA in both

**Table 4**

	Physician Group 1*		Physician Group 2*	
	Observed Pearson (p)	Disattenuated Pearson (p)†	Observed Pearson (p)	Disattenuated Pearson (p)†
Communication score	.136 (> .1)	.172 (> .05)	.200 (> .1)	.253 (> .05)
Medical performance score	.405 (< .01)	.586 (< .01)	.226 (> .1)	.327 (< .05)

\*Physicians in Group 1 first took the MSE and then the PVA; physicians in Group 2 took the examinations in the opposite order.  
 †"Disattenuated" means corrected for unreliability of the instruments.

groups. For communication, we found no significant correlation between the scores on the two tests, either in group 1 or in group 2.

**Reliability**

In Table 5, generalizability coefficients are reported as a function of the number of cases and the number of raters for both the MSE and the PVA. Each entry represents a reliability estimate for the given sample size of cases and raters. For example, using two raters, a value of 0.81 (PVA) and 0.84 (MSE) on the total score was found for eight cases, i.e., 1 hour 36 minutes of testing time. An benchmark of 0.80 is usually considered as an acceptable value, which is reached for all scores within two and a half hours of testing time, using two raters.

**Feasibility**

Fixed and variable costs were estimated for two different sample sizes of consultations of both methods. For a 12-case sample, the cost for the MSE was \$530 and for the PVA, \$375; for 16 cases, \$680 and \$440 respectively (Table 6). Stable costs refer to fixed costs, including the purchase and installation of video equipment in practice (PVA) and the costs of videotapes and rater train-

Table 5

Reliability: Generalizability Coefficients, as a Function of the Number of Cases and Raters, Measuring the Reliability of Scores for Physicians' Taking a Multiple-station Examination (MSE) and a Practice Video Assessment (PVA), The Netherlands, 1996*												
No. of Cases	Overall Score				Communication Score				Medical Performance Score			
	MSE		PVA		MSE		PVA		MSE		PVA	
	1 Rater	2 Raters	1 Rater	2 Raters	1 Rater	2 Raters	1 Rater	2 Raters	1 Rater	2 Raters	1 Rater	2 Raters
4	0.57	0.72	0.58	0.68	0.60	0.75	0.55	0.65	0.42	0.59	0.52	0.63
<b>8</b>	<b>0.72</b>	<b>0.84</b>	0.74	0.81	<b>0.75</b>	<b>0.86</b>	0.71	0.79	<b>0.59</b>	<b>0.74</b>	0.68	0.78
12	0.80	0.89	0.81	0.87	0.82	0.90	0.79	0.85	0.68	0.81	0.76	0.84
<b>16</b>	0.84	0.91	<b>0.85</b>	<b>0.90</b>	0.86	0.92	<b>0.83</b>	<b>0.88</b>	0.74	0.85	<b>0.81</b>	<b>0.87</b>
20	0.87	0.93	0.87	0.91	0.88	0.94	0.86	0.90	0.78	0.88	0.84	0.90

\* **Bold italics** = real sample sizes.

ing (PVA and MSE). The participants preferred the PVA over the MSE to assess both communication (73% vs 27%) and medical performance (66% vs 34%).

#### DISCUSSION

In this study, we compared a multiple-station examination (MSE) at the medical school, using standardized patients, with a practice video assessment (PVA) in the family physicians' offices. We studied the validity, reliability, and feasibility of using each of the two methods for assessing the physicians' performances in consultations with patients. Our results showed that the PVA was superior in all aspects. We realize that our study population may not be representative, since the family physicians participated voluntarily, showing their readiness to be assessed. Because the MSE scores of the physicians in group 2 (who first took the PVA) were significantly higher than the MSE scores of the physicians in group 1 and because there was no significant difference between the groups' PVA scores, we conclude that a time-order effect (resulting in better scores for competence) occurred only in one direction: from "in

vivo" assessment to "in vitro" assessment. Apparently, only assessment in simulated environments is influenced by previous assessment experiences. This stability in PVA scores may be ex-

plained by the common interaction between family physicians and their patients in daily practice.

Both settings achieved satisfactory content validity: the family physicians

Table 6

Costs (in U.S.\$) for Assessing One Physician Using One Peer-Observer Rater in a Multiple-Station Examination (MSE) and a Practice Video Assessment (PVA), The Netherlands, 1996*				
Costs	Assessing 12 Cases		Assessing 16 Cases	
	MSE	PVA	MSE	PVA
Administration	60	54	75	67
Equipment/rooms	52	<b>22</b>	70	<b>22</b>
Videotapes	7	<b>22</b>	7	<b>22</b>
Rater training	<b>45</b>	<b>45</b>	<b>45</b>	<b>45</b>
Observation	112	112	150	150
Feedback	15	15	22	22
Standardized patients	75	—	100	—
Script cases	7	—	10	—
Travel	45	—	51	—
Locum tenens†	112	—	150	—
Installation of video equipment	—	<b>75</b>	—	<b>75</b>
Review of physicians' logbooks	—	30	—	37
TOTAL	530	375	680	440

\* **Bold italics** = stable costs.  
† On the days that physicians were participating in the MSE, they had to hire substitute physicians to see their patients.

confirmed the authenticity of the MSE's standardized cases, and, as was evident from comparison with the blueprints, the PVA encounters covered the domain of general practice care. However, the PVA consultations were more authentic than were the MSE cases, since initial and follow-up consultations with children and older patients occurred in the natural daily context.<sup>24</sup> Moreover, almost all the family physicians judged the videotaped practice consultations to be "natural," whereas hardly any of the family physicians recognized their own working styles in the standardized station cases, since they met unknown standardized patients in an artificial setting. The influence of variation in total duration of the 16 PVA consultations on performance scores reflects differences between family physicians in their perceptions of tasks, work loads, and working styles. Therefore, our selection criterion (consultations lasting between five and 15 minutes) contributes to the representativeness of the samples. Finally, concurrent validity between MSE and PVA is ambiguous, since only a moderate correlation was found for medical performance. Overall, the validity of the PVA was superior to that of the MSE.

The reliability results for both methods are encouraging, compared with other studies.<sup>29-31</sup> We expected the results to be superior for the controlled MSE, but after 16 cases, the benchmark of 0.80 was met for all scores (except for medical performance on the MSE). This finding favors the PVA. Family physicians may perform more consistently in their authentic daily practices, whereas they show more variability from case to case in an unfamiliar competence test. Adding another rater to each case substantially improved the reliability in both test formats, suggesting considerable rater bias. Therefore, efforts to improve reliability should be focused on rater training. Overall, the expected loss of reliability of the PVA as

compared with the MSE was not confirmed.

Finally, the multiple-station examination was more expensive than the practice video assessment, through the use of standardized patients, laboratory facilities, and travel costs for participants and standardized patients. Furthermore, the MSE was more burdensome for the participants because of travel time, disruption of patient care by being away from their practices, and the fixed period of organization. In contrast, with the PVA patient care was continued normally and the planning of the period of video recordings was more flexible without increasing the cost. Again, the practice video assessment appeared to be superior to the multiple-station examination.

We conclude that the results for validity, reliability, and feasibility favor the practice video assessment over the multiple-station examination. In quality assessment of the communication and medical performances of practicing family physicians, video observation in daily practice is therefore to be preferred to assessment by observation in a controlled simulated situation.

Efforts to improve "video assessment on the job" for any specialty should be focused on "standardization with validation" of observation of real physician-patient interactions, on training raters to score complex daily consultations, and on developing nationally accepted guidelines for both actual communication and medical performance in order to enable setting standards. Different specialties in different countries may benefit from such an assessment, provided that procedures developed in our study are followed meticulously, assessment instruments with good psychometric characteristics are used, and a good classification system for diseases is available. Further research of the educational effect of video assessment in daily practice is needed.

## REFERENCES

1. Grol R, Wensing M, Jacobs A, Baker R (eds). *Quality Assurance in General Practice. The State of the Art in Europe*. Utrecht, The Netherlands: NHG and EQuiP, 1993.
2. New perspectives and opportunities in quality assurance: our role. In: Federation of State Medical Boards of the United States Annual Meeting. San Diego, CA, 1988.
3. 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.
4. McPhee SJ, Bird JA, Fordham D, Rodnick JE, Osborn EH. Promoting cancer preventing activities by primary care physicians. *JAMA*. 1991; 266: 393-401.
5. Sibley JC, Spitzer WO, Rudnick KV, et al. Quality of care appraisal in primary care: a quantitative method. *Ann Intern Med*. 1975; 83: 46-52.
6. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. In: Schmidt H, Norman G (eds). *Advances in Health Sciences Education: Theory and Practice: Volume 1*. Dordrecht, The Netherlands/Boston/London: Kluwer Academic Publishers, 1996.
7. Rethans JJ, Westin S, Hays R. Methods for quality assessment in general practice. *Fam Pract*. 1996; 13: 468-76.
8. Wakefield J. Direct observation. In: Neufeld VR, Norman GR (eds). *Assessing Clinical Competence*. New York: Springer Publishing Company, 1985: 51-71.
9. Kane MT. The assessment of professional competence. *Eval Health Prof*. 1992; 15: 163-82.
10. Elstein A, Shulman LS, Sprafka SA. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press, 1978.
11. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *BMJ*. 1975; 1: 447-51.
12. Stillman P, Swanson D. Ensuring the clinical competence of medical school graduates through standardized patients. *Arch Intern Med*. 1987; 147: 1049-52.
13. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med*. 1990; 2: 58-76.
14. Rethans JJ, van Leeuwen Y, Drop R, Sturmans F, van der Vleuten C. Performance and competence: two different constructs in the assessment

- of quality of medical care. *Fam Pract.* 1990; 7: 168-74.
15. Campbell LM, Murray TS. Assessment of competence. *Br J Gen Pract.* 1996; 46: 619-22.
  16. Metsemakers JFM, Höppener P, Knottnerus JA, et al. The Registration Network Family Practices: a computerized health information system in The Netherlands. *Br J Gen Pract.* 1992; 42: 102-6.
  17. Lamberts H, Wood M (eds). *ICPC: International Classification of Primary Care.* Oxford, U.K.: Oxford University Press, 1987.
  18. Rutten GEH, Thomas S (eds). *NHG-standaarden voor de huisarts [Guidelines for the general practitioner].* Utrecht, The Netherlands: Bunge/NHG, 1993.
  19. Springer MP (ed). Basic job description for the general practitioner. Utrecht, The Netherlands: Dutch National Association of General Practitioners, 1983.
  20. Ram PM, Grol RPTM, Rethans JJ, Schouten B, van der Vleuten CPM, Kester A. Assessment of general practitioners by video observation of medical and communicative performance in daily practice: issues of validity, reliability, and feasibility. *Med Educ.* In press.
  21. van Thiel J, van der Vleuten CPM. Reliability and feasibility of measuring interviewing skills using the revised Maastricht History Taking and Advice Checklist. *Med Educ.* 1991; 25: 224-9.
  22. van Thiel J, van Dalen J, Ram P. MAAS-globaal criterialijst. Maastricht University, Maastricht, The Netherlands 1995 (internal publication).
  23. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley, 1972.
  24. Brennan RL. *Elements of Generalizability Theory.* Iowa City, IA: American College Testing Publications, 1983.
  25. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: testing the reliability of the Leicester assessment package. *Br J Gen Pract.* 1994; 44: 293-6.
  26. Newble D, Swanson D. Psychometric characteristics of the objective structured clinical examination. *Med Educ.* 1988; 22: 325-34.
  27. Norcini J, Swanson D. Factors influencing testing time, requirements for simulation-based measurements: do simulations ever yield reliable scores? *Teach Learn Med.* 1989; 1: 85-91.